RELEASING DIFFERENTIALLY PRIVATE SYNTHETIC MICRO-DATA WITH BAYESIAN GANS

WORKING PAPER

Christian Arnold
Cardiff University
CF10 3AT, UK
arnoldc6@cardiff.ac.uk

Marcel Neunhoeffer
University of Mannheim
A5 6, D-68159 Mannheim, Germany
mneunhoe@mail.uni-mannheim.de

Sebastian Sternberg
University of Mannheim
A5 6, D-68159 Mannheim, Germany
ssternbe@mail.uni-mannheim.de

August 30, 2018

ABSTRACT

This paper shows how to generate differentially private synthetic data using generative adversarial nets (GANs). We bring together insights from three literatures. First, generating artificial copies of original data is considered the gold standard in differential privacy, since any further analysis of this kind of data does not spend any extra amount of the privacy budget. While this literature has used machine learning to generate synthetic data on relatively trivial data sets, we show how to handle even complex data structures. Second, GANs became prominent in learning and generating the representation of visual and audio data. However, unlike in the context of synthetic visual and audio data, synthetic micro-data requires to take account not only of the point estimate, but also has to capture the diversity of the original data. We therefore apply, third, Bayesian GAN. We show how BayesGAN can generate differentially private data when injecting the right amount of noise during training with a Stochastic Gradient Langevin Dynamics sampler. In our paper, we are the first to generate differentially private data using BayesGAN. So far, our experiments show that we generate differentially private micro-data that are at least as useful for analysis and prediction as synthetic data generated with other, so far considered methods. In addition, we also incorporate the privacy loss parameters ϵ and δ into our framework which allows users to control the desired privacy loss of the synthetic data.

Keywords Bayesian GAN · GAN · Machine Learning · Synthetic Data · Differential Privacy

1 Introduction

In the era of Big Data, concerns about privacy and data protection are omni-present: Scientific studies often use confidential data for their studies. Government agencies hold data about citizens that are equally sensitive. And companies – think Facebook, Google or Twitter – are collecting data about consumer (online) behaviour at an unprecedented level. But often, this data needs to be shared. To make studies replicable, scientists have to disclose code and data. Companies and governments want to share data with those who have the technology and/or knowledge to analyse it–be it other profit or non-profit organisations. Simple anonymisation techniques such as removing personally identifiable information have long been shown to not being enough (see Sweeney, 1997). Original data simply cannot be shared if high levels of data protection need to be guaranteed. In contrast, while any highly anonymised data might be safe to share, it is questionable whether this data still contains enough information for analysis.

In statistics, the treatment of privacy-sensitive data has a long history under the name of "statistical disclosure control" or "statistical disclosure limitation". In this context, synthetic data is a is statistical disclosure limitation technique which aims at releasing data to the public while protecting privacy-sensitive information. The main idea of synthetic micro-data is to release micro-data, that is, data on individual records, synthesized based on the information in the original data. The usage of synthetic data for statistical disclosure has a long history and was first proposed by Rubin (1993) and Little (1993) in the spirit of multiple imputation. Nowadays, there exist a wide range of methods to produce synthetic

micro-data (Reiter and Raghunathan, 2007; Reiter, 2005; Drechsler and Reiter, 2010; Kinney et al., 2011; Drechsler and Reiter, 2011; Drechsler, 2011; Manrique-Vallier and Hu, 2018). However, existing synthetic data approaches often rely on heuristic arguments about data safety – e.g. assumptions about the resources of an intruder (computing power, time), or the availability of auxiliary information (see Dwork et al., 2017).

Over the last decade, a more rigorous disclosure protection standard – *differential privacy* (DP) – became increasingly popular in academia, business and government alike. DP originated in cryptography and relies on mathematical terms to provide strong privacy guarantees (Dwork et al., 2006). In short, the general idea of DP is the requirement of an outcome of a randomized data analysis procedure (a simple statistic such as a mean, or the output of an algorithm such as a data synthesizer) not to change much when this outcome is calculated from two neighbouring data sets that differ by only one record (e.g. one individual). In other words, DP guarantees that the difference between any two adjacent datasets does not disclose information about any individual observation.

In this paper, we show how to generate DP micro-data. We use a generative adversarial network (GAN) (Goodfellow et al., 2014) in a Bayesian framework (BayesGAN) (Saatchi and Wilson, 2017). Sampling with Stochastic Gradient Langevin Dynamics, we add noise to the gradients in line with the bounds formulated in Wang et al. (2015), thus adding differential privacy to the BayesGAN (DP-BayesGAN).

Our solution offers a number of advantages over existing approaches to generate micro-data. First, it is easy to use as it does not require complex hardware architectures, assumptions about distributions and functional forms, specification of variables, feature matching or label smoothing. Second, DP-BayesGAN is capable to cover highly complex and multi-modal data structures, and is less prone to suffer from mode collapse like a standard GAN would be. This is an important aspect, given that most real-world data is highly complex and multi-modal. Third, a standard GAN can still implicitly disclose privacy-sensitive information about the original data. Hitaj et al. (2017) show for instance that it is possible to reconstruct original training samples from generated samples. Our proposed framework does not suffer from this problem due to the incorporation of DP during the noise added to the gradient during optimization.

We therfore make three clear contributions. First, we show that BayesGAN can easily be made differentially private. While there already exist work on differentially private GANs (Beaulieu-Jones et al., 2017; Triastcyn and Faltings, 2018; Xie et al., 2018), we are the first, to the best of our knowledge, who show that DP can be incorporated into the Bayesian GAN framework. Second, we suggest a way to generate synthetic data at any user defined level of the privacy parameter ϵ and δ. Third, we show the usefulness of Bayesian GANs in an application on the generation of differentially private synthetic micro-data.

In the following section, we will first consider related work. We then take a closer look at the challenges when generating differentially private data and proceed to explain how GANs can be a great means to generate synthetic copies from original dataset. We then make bayesian GANs differentially private and present our approach for modelling the privacy parameters ϵ and δ into the generator; thus allowing to *ex-ante* define the privacy budget that can be spent when generating synthetic data. We provide an overview over the so far implemented experiments and offer concluding remarks.

2 Related Work

Our paper joins insights from three related fields, that so far do not communicate to one another: The machine-learning community around Generative Adversarial Nets, scholars from computer science and statistics interested in Bayesian formulations of differential private data analysis and finally (applied) statisticians who care about the creation of differentially private micro-data. While there have been first attempts to build bridges across respectively two combinations of these sub-fields, we are the first to draw from all three of these literatures.

2.1 Synthetic Data as a Statistical Disclosure Limitation Technique

Synthetic data was introduced as a solution for statistical disclosure control. A statistical model based on the original data generates "artificial", or "fake" data. The intuition behind synthetic data is simple: some or all of the original values in the data set should be replaced by values sampled from some appropriate probability distribution so that the statistical properties of the original data set are preserved. Anybody who seeks to analyse the original data will have access to the synthetic and use this data instead.

Synthetic data as a statistical disclosure limitation technique has been introduced almost three decades ago.[1] Rubin (1993) and Little (1993) first advanced synthetic data methods in the spirit of multiple imputation. After being formulated as a proper framework (Raghunathan, 2003), a series of papers elaborated it further (Abowd and Lane, 2004;

[1] For overviews, see e.g. (Snoke et al., 2016; Drechsler, 2011).

Abowd and Woodcock, 2004; Reiter and Raghunathan, 2007; Reiter, 2002; Drechsler and Reiter, 2010; Kinney et al., 2010, 2011). Machine Learning techniques were used early on (Reiter, 2005) and increasingly also included more complex approaches (Caiola and Reiter, 2010; Drechsler and Reiter, 2011).

Synthetic data for disclosure control has several important (theoretical) advantages as compared to the traditional statistical disclosure limitation methods such as data swapping, aggregation or cell suppression. The results are similar no matter whether an analyses is performed on the synthetic or original data. Also, synthetic data can preserve confidentiality more easily, because a synthetic data record is not any respondent's actual data record. Therefore, the identification of individuals or privacy sensitive samples in general is difficult if not impossible. Finally, if the estimation method is appropriate, the approach can allow the data users to draw valid and correct inferences for a variety of estimands. An analyst does not need to know any particular assumptions about how the synthetic data were created.

2.2 Generative Adversarial Nets

Goodfellow et al. (2014) introduces Generative Adversarial Nets. Delivering promising results on prediction problems, GANs became particularly prominent for their capacity to generate artificial, yet realistic looking images (Salimans et al., 2016). Arjovsky et al. (2017) offer a theoretically informed framework that helps convergence and counter well known fallacies–like mode collapse.

Recently, a number of attempts have been made to formulate Bayesian versions of GANs. Saatchi and Wilson (2017) offer a complete Bayesian treatment of semi-supverised and un-supervised learning of GANs. Introducing hierarchical implicit models, Tran et al. (2017) also propose a Bayesian formulation of GANs by placing a prior on the network parameters θ. Wang et al. (2018) propose a destillation framework for GANs that makes storage of sampled MCMC parameters more efficient.

2.3 Differential Privacy

Differential Privacy originated in a subfield of Computer Science concerned with cryptography (Dwork et al., 2006). The concept has been gaining increasing attention beyond the borders of the subfield and has influenced applications in various other areas.

2.3.1 DP and Bayes

Modelling parameters as distributions, the Bayesian perspective lends itself quite naturally to differential privacy. However, in the light of modern machine learning models with abundant parametarisation and large data sets, marginalising posterior distributions with traditional MCMC techniques becomes increasingly challenging. The advent of Stochastic Gradient methods were an important step to apply Bayesian approaches to modern machine learning models (Chen et al., 2014; Welling and Teh, 2011). Building on these efforts, Wang et al. (2015) are the first to bring differential privacy to Bayesian samplers. Li et al. (2017) continue in this line; they re-formulate more narrow bounds for step sizes and thus achieve state-of-the-art training results while at the same time satisfying privacy concerns. Differentially private stochastic gradient samplers were possible, because scholars adopted insights from machine learning and cryptography. Applying these insights to Bayesian GANs and making them differentially private spills this knowledge back to the machine learning literature.

2.3.2 DP and GANs

Differential privacy has also been considered in deep learning. Abadi et al. (2016) lay an important foundation for later applications to GANs. They clip gradients and add gaussian noise at each step, and then calculate the overall suffered privacy loss ϵ using moments accounting.

GANs have been used to address privacy concerns more broadly, for example to protect privacy in images (e.g. Tripathy et al., 2017; Wu et al., 2018) or–in contrast–to attack privacy in the context of distributed learning setting (Hitaj et al., 2017).

More recently, differential privacy as a more narrow definition of privacy has been applied to GANs, too. A number of studies follows the framework of Abadi et al. (2016). (Beaulieu-Jones et al., 2017) generate differentially private data from a medical trial. (Xie et al., 2018) also produce differentially private data: They establish differential privacy by combining noise and weights clipping. The privacy loss is calculated through moments accounting. Triastcyn and Faltings (2018) use GANs to explicitly generate micro-data. To hide sensitive data, they enforce DP on the penultimate layer by clipping its $L2$ norm and adding Gaussian noise. They then evaluate the privacy parameter ϵ empirically on the basis of pairwise comparisons between every possible pair of adjacent data sets.

In contrast to the context free privacy solution differential privacy is offering, Huang et al. (2017) introduce GANs as means to generate context aware privatisation schemes and mask data of interest on the basis of predefined variables of interest. Building on this, Huang et al. (2018) reconceptualise the generator as "privatizer" and use the GAN framework to find an optimal privacy mechanism.

Our model clearly relates to existing efforts. We add a Bayesian version of a GAN to the existing model landscape, offering a natural way to differential privacy: Bayesian samplers have been shown to naturally follow requirements for differential privacy, while at the same time being robust implementations that are easy to estimate.

3 Preliminaries

To fully understand our DP innovation, we introduce key concepts more formally.

3.1 Differential Privacy

Differential Privacy (DP) is a mathematical concept and strong privacy guarantee with roots in cryptography (Dwork, 2006; Dwork and Roth, 2013; Nissim et al., 2017).

(ϵ, δ)-DP is defined by Dwork and Roth (2013) as:

A randomized algorithm $\mathcal{M}$ with domain $\mathbb{N}^{|\mathcal{X}|}$ is (ϵ, δ)-differentially private if for all $\mathcal{S} \subseteq \text{Range}(\mathcal{M})$ and for all $x, y \in \mathbb{N}^{|\mathcal{X}|}$ such that $||x - y||_1 \leq 1$[2]:

$$Pr[\mathcal{M}(x) \in \mathcal{S}] \leq e^{\epsilon} Pr[\mathcal{M}(y) \in \mathcal{S}] + \delta.$$

For $\delta = 0$ this is called pure DP or ϵ-DP.

Nissim et al. (2017) put the definition of ϵ-DP into words: "Differential privacy mathematically guarantees that anyone seeing the result of a differentially private analysis will essentially make the same inference about any individual's private information, whether or not that individual's private information is included in the input to the analysis."

ϵ is the so called privacy loss parameter. It quantifies how much information can be learnt about any individual included in x, in a worst case scenario of an attacker with arbitrary side knowledge and arbitrary computational power. For small values of ϵ the potential privacy loss for any observation in x is small. Larger values of ϵ mean that the potential privacy loss is higher, with $\epsilon = \infty$ meaning that all the information about any individual can be learnt (i.e. by publishing the original data set).

For our application we consider a popular relaxation of ϵ-DP, with $\delta > 0$, so called (ϵ, δ)-DP. In the words of Dwork and Roth (2013): "$(\epsilon, 0)$-differential privacy ensures that, for every run of the mechanism $\mathcal{M}(x)$, the output observed is (almost) equally likely to be observed on every neighboring database, simultaneously. In contrast (ϵ, δ)-differential privacy says that for every pair of neighboring databases x, y, it is extremely unlikely that, ex post facto the observed value $\mathcal{M}(x)$ will be much more or much less likely to be generated when the database is x than when the database is y" (18).

What makes DP so strong is that it does not depend on attacker capabilities–an attacker can have arbitrary side knowledge and even arbitrary computational power. This is an important advantage of DP over other existing risk measures in the Statistical Disclosure Limitation (SDL) literature, where usually assumptions about the attacker have to be made.

3.2 Synthetic Data

The intuitive idea behind synthetic data is that an attacker could only learn about synthetic people and not real people or any other type of observation for that matter. In general, there are two ways to generate synthetic data (Manrique-Vallier and Hu, 2018). In sequential modelling, each variable at a time is modelled on the basis of the rest of the data and the respective resulting models are used to generate synthetic data (Van Buuren et al., 2006). Reiter (2005) and Caiola and Reiter (2010) are the first to use Classification and Regression Trees (CARTs) for this purpose. A second approach is to model the data distribution jointly. Discrete data have been treated by Matthews et al. (2010) and Hu et al. (2014).

Yet this idea alone does not guarantee privacy in the framework of DP. Sophisticated attackers with lots of computational power and arbitrary side knowledge (e.g. about who is in the original data set) might still be able to reconstruct real

[2]This means that x and y are two adjacent data sets that are differing only by one row.

observations when being presented synthetic data. Only *differentially private* micro data can guarantee privacy against attacks that do not require assumptions about attackers' capabilities.

Once differentially private data is generated, it holds yet another important advantage–this time over differentially private algorithms to analyse original data: Even when using an ϵ differentially private algorithm on original data, any analysis will spend some part of the privacy budget. The more analyses are run, the more likely it is that original data can be identified by an attacker. In contrast, synthetic data spends this privacy budget only once: when generating the synthetic data. Analyses can be run an arbitrary number of times on synthetic data without increasing the disclosure risk (Dwork and Roth, 2013; Nissim et al., 2017).

Differentally private data synthesis is a very recent area of research. Only six years ago, McClure and Reiter (2012) note that: "There is a long way to go before differentially private synthetic data generation becomes feasible for highly complex datasets." For an excellent overview over the quick advances of differentially private data synthesis see Bowen and Liu (2016).

When generating synthetic data, there needs to be a measure for the utility. Snoke et al. (2018) suggest to use the propensity score mean-squared error pMSE to measure differences between original and synthetic data. For DP-synthetic data the choice of ϵ is not trivial and a "social question". The data owner has to decide on how much privacy loss is accaptable. Yet, a data release is only useful if some of the statistical utility is preserved. Complicating matters is that there is no straightforward interpretation of ϵ in the context of synthetic data.

3.3 Generative Adversarial Nets as a Means to Generate Complex Synthetic Data

The basic idea of a GAN is surprisingly intuitive. At its core, a GAN is a minimax game with two competing actors–a discriminator (D) trying to tell real from synthetic samples and a generator (G) to produce realistic synthetic samples from random noise.

We use the same illustrative example as Goodfellow et al. (2014) to make GANs (and the adjustments later on) more accessible: "The generative model can be thought of as analogous to a team of counterfeiters, trying to produce fake currency and use it without detection, while the discriminative model is analogous to the police, trying to detect the counterfeit currency. Competition in this game drives both teams to improve their methods until the counterfeits are indistiguishable from the genuine articles."

In GANs the team of counterfeiters, the generator, is a neural network which is trained to produce realistic synthetic data examples from random noise. And the police, the discriminator, is a neural network with the goal to classify fake and real data. The generator network is trained to be able to fool the discriminator network, and uses the feedback of the discriminator to generate increasingly realistic "fake" data that should eventually be indistinguishable from the original ones. At the same time, the discriminator is constantly adapting to the more and more improving generating abilities of the generator. Thus, the "threshold" where the discriminator is fooled increases along with the faking capabilities of the generator. This goes on until equilibrium is reached[3].

Formally, this two-player minimax game can be written as:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}(x)}\Big[\log D(x)\Big] + \mathbb{E}_{z\sim p_z(z)}\Big[\log(1-D(G(z)))\Big] \tag{1}$$

where $p_{data}(x)$ is the distribution of the real data, X is a sample from $p_{data}(x)$. The generator network $G(z)$ takes as input z from $p(z)$, where z is a random sample from a probability distribution $p(z)$[4]. Passing the noise z through G then generates a sample of synthetic data which is then fed into the discriminator $D(x)$. The discriminator takes as input a set of labeled data, either real (x) from $p_{data}(x)$ or generated $(G(z))$, and is trained to distinguish between real data and synthetic data[5]. D is trained to maximize the probability of assigning the correct label to training examples and samples from $G(z)$. G is trained to minimize $log(1 - D(G(z)))$. Thus, the goal of the discriminator is to maximize function V, whereas the goal of the generator is to minimize it.

The equilibrium point for the GANs is that the G should model the real data and D should output the probability of 0.5 as the generated data is same as the real data – that is, it is not sure if the new data coming from the generator is real or fake with equal probability.[6]

[3]Interestingly, a GAN is therefore a dynamic system where the optimisation process is seeking not a minimum, but an equilibrium. This is in stark contrast to standard deep learning systems, where the entire loss landscape is static.

[4]Usually GANs are set up to either sample from uniform or Gaussian distributions.

[5]This is a standard binary classification problem, and thus the standard binary cross-entropy loss with a sigmoid function at the end can be used.

[6]Note the connection to the measure of general utility presented by Snoke et al. (2018). The explicit goal of a GAN is to maximize general utility, and therefore a natural way to generate fully synthetic data.

GANs have been shown to be capable of generating sophisticated synthetic copies from image and audio data (complicated high-dimensional distributions from which it was difficult to sample from). In this paper we suggest to make use of these powerful algorithms to produce synthetic micro-data. After all, any synthetic data approach to differential privacy requires that the data generating algorithm is capable of generating high quality artificial data. Those who analyse synthetic data can only detect relationships that a generating algorithm is actually capable of replicating (Manrique-Vallier and Hu, 2018).

3.4 Bayesian GANs for Diverse Synthetic Data (Instead of Some Very Good Data Points)

One common problem of GANs, as described above, is so called mode collapse. Loosely speaking, this means that the generator G focuses on few examples it knows will trick the discriminator. In the context of the illustrative example, mode collapse means that the team of counterfeiters learns how to fake 20 Euro notes very well. Enthusiastic about the initial success, they focus entirely on producing 20 Euro notes while not even attempting to fake other notes as well.

For fully synthetic data, mode collapse is particularly detrimental. It means that the collapsed generator G' would make "copies" of the most likely real data points in $p_{data}(x)$ and that diversity in a sample would be lost. But even without mode collapse, all inference about the original data distribution $p_{data}(x)$ focusses on the most probable mode of the Generator Likelihood only. GANs generate parameters θ_G and θ_D that represent one mode of a very likely multi-modal distribution–but one mode only. This nuisance in the context of standard GANs becomes a real problem when attempting to generate diverse synthetic data sets. To capture and copy from the full distribution $p_{data}(x)$, the GAN has to explore the complete posterior distribution of the generator G and the discriminator D. Saatchi and Wilson (2017) recently proposed a Bayesian formulation of a GAN, a powerful framework for the *full* exploration of the parameter posteriors.

The Bayesian GAN (BayesGAN) is a straightforward Bayesian formulation of the traditional GAN (or Maximum Likelihood GAN). Instead of finding the most likely parameter vectors $\theta_{\mathbf{G}}$ for the Generator and $\theta_{\mathbf{D}}$ for the Discriminator, the goal is to fully "represent the posterior distribution over the parameters" of both. The posteriors given in Saatchi and Wilson (2017) are:

$$p(\theta_G|\mathbf{z}, \theta_D) \propto \left(\prod_{i=1}^{n_G} D(G(\mathbf{z_i}; \theta_G); \theta_D\right) p(\theta_G|\alpha_G) \tag{2}$$

$$p(\theta_D|\mathbf{z}, \mathbf{X}, \theta_G) \propto \prod_{i=1}^{n_D} D(\mathbf{x_i}; \theta_D) \times \prod_{i=1}^{n_G} (1 - D(G(\mathbf{z_i}; \theta_G); \theta_D)) \times p(\theta_D|\alpha_D) \tag{3}$$

To sample from the posteriors Saatchi and Wilson (2017) propose to use stochastic gradient hamiltonian monte carlo (SGHMC) an sample from $p(\theta_G|\mathbf{z}, \theta_D)$ and $p(\theta_D|\mathbf{z}, \mathbf{X}, \theta_G)$ at each training step.

In their application Saatchi and Wilson (2017) show that BayesGAN is able to recover highly complex multi modal distributions on which the standard GAN approaches failed. This makes BayesGAN a promising framework to produce diverse synthetic data. Another advantage of the Bayesian formulation of the GAN is that it is rather simple to make it differentially private.

4 Differentially Private BayesGAN

To make BayesGAN DP we rely on a DP Stochastic Gradient MCMC sampler. Since only the Discriminator has access to the training data it is sufficient to make the Discriminator DP (see also Xie et al., 2018; Triastcyn and Faltings, 2018). We achieve this by injecting noise into the gradients during training of BayesGAN.

In the context of the illustrative example this means that the police is disturbed (eg. only has blurred vision) while trying to tell real notes from fake ones. This will also reduce the capabilities of the team of counterfeiters as they cannot learn more about the difference between real notes and their fake notes.

For the first application of a DP-BayesGAN we rely on the DP Stochastic Gradient Langevin Dynamics (SGLD) sampler. Wang et al. (2015) prove that SGLD is "differentially private for free if the parameters are chosen appropriately". In particular, the amount of noise injected into the gradients during training is given by:

$$n \sim \mathcal{N}\left(0, \frac{128NTL^2}{\tau\epsilon^2} \log(\frac{2.5NT}{\tau\delta}) \log(\frac{2}{\delta})\eta_t^2\right) \tag{4}$$

as long as, $T \geq \frac{\epsilon^2 N}{32\tau \log(2/\delta)}$.

Where ϵ and δ are the privacy loss parameters for (ϵ, δ)-DP. N is the number of observations in the training data set. T is the number of data passes during training[7]. L is the Lipschitz constant. τ is the number of training examples in a minibatch and η_t is the step size at training iteration t. [8]

This implies the following algorithm for DP-BayesGAN based on the BayesGAN algorithm by Saatchi and Wilson (2017).

η is the learning rate, N the number of observations in the training data $\mathbf{x}$, τ is the size of the minibatch, T is the number of data passes. Like Saatchi and Wilson (2017) we take J_g and J_d simple MC samples for the generator and discriminator respectively, and M SGLD/DP-SGLD samples for each simple MC sample;

for $t = 1 : \lfloor NT/\tau \rfloor$ **do**

  for *number of MC iterations* J_g **do**

    Sample J_g noise samples $\{\mathbf{z}^{(1)}, ..., \mathbf{z}^{(J_g)}\}$ from noise prior $p(\mathbf{z})$. Each $\mathbf{z}^{(i)}$ has τ samples;

    Update sample set representing $p(\theta_g|\theta_d)$ by running SGLD updates for M iterations:

$$\theta_g^{j,m} \leftarrow \theta_g^{j,m} + \eta_t \left(\sum_{i=1}^{J_g} \sum_{k=1}^{J_d} \frac{\partial \log p(\theta_g|\mathbf{z}^{(i)}, \theta_d^{k,m})}{\partial \theta_g} \right) + \mathbf{n}; \mathbf{n} \sim \mathcal{N}(0, 2\eta_t I)$$

    ;

    Append $\theta_g^{j,m}$ to sample set ;

  end

  for *number of MC iterations* J_d **do**

    Sample minibatch of J_d noise samples $\{\mathbf{z}^{(1)}, ..., \mathbf{z}^{(J_g)}\}$ from noise prior $p(\mathbf{z})$. Each $\mathbf{z}^{(i)}$ has τ samples;

    Sample minibatch of τ data samples x;

    Update sample set representing $p(\theta_d|\mathbf{z}, \theta_g)$ by running DP-SGLD updates for M iterations:

$$\theta_d^{j,m} \leftarrow \theta_d^{j,m} + \eta_t \left(\sum_{i=1}^{J_d} \sum_{k=1}^{J_g} \frac{\partial \log p(\theta_d|\mathbf{z}^{(i)}, \mathbf{x}, \theta_g^{k,m})}{\partial \theta_d} \right) + \mathbf{n}; \mathbf{n} \sim \mathcal{N}\left(0, \frac{128NTL^2}{\tau\epsilon^2} \log(\frac{2.5NT}{\tau\delta}) \log(\frac{2}{\delta}) \eta_t^2 I\right)$$

    ;

    Append $\theta_d^{j,m}$ to sample set ;

  end

end

Algorithm 1: DP-BayesGAN

This means that before training the BayesGAN the researcher can set the desired amount of ϵ and δ. In the following section we explore how different levels of ϵ affect the general utility of the synthetic micro-data generated with DP-BayesGAN.

5 Experiments

With the following experiments we show three features of DP-BayesGAN. First, we show that the non-private BayesGAN can compete with current state of the art models to produce synthetic data of highly complex data. Second, we show that with introducing DP-BayesGAN the general utility of the data decreases as expected. And third, that the disclosure risk indeed decreases for smaller values of ϵ in the DP-BayesGAN.

To assess the utility of the synthetic copies of the data set we use the general utility measure for synthetic data as proposed by Snoke et al. (2018).

[7] Note that T is not the number of training iterations. The number of total training iterations is given by $\lfloor NT/\tau \rfloor$

[8] We note that narrower bounds for DP-SGLD are possible () but haven't been explored for this first version of the paper. It is also possible to derive a DP version of a stochastic gradient hybrid monte carlo (SGHMC) sampler which still has to be implemented as well.

5.1 A Complex Data Generating Process and Synthetic Copies

To assess the different synthetic copies we set up a complicated true data generating process (DGP). In particular, we rely on the DGP proposed by Montgomery and Olivella (2018). They generate 500 observations of 40 explanatory variables and one dependent variable. The explanatory variables include "symmetric and asymmetric variables, continuous and categorical variables, and correlated and independent variables" and are generated as follows.

$$x_{1i} \sim Gamma(8, 2);$$
$$x_{2i} \sim Gamma(10, 1);$$
$$[x_3 \quad x_4 \quad x_5]_i' \sim MVN([2 \quad 3 \quad 6], [1.5 \quad 0.5 \quad 3.3]' I_3);$$
$$[x_6 \quad x_7 \quad x_8]_i' \sim Multinom(\left[\tfrac{1}{3} \quad \tfrac{1}{3} \quad \tfrac{1}{3}\right], n = 1)$$
$$[x_9 \quad x_{10}]_i' \sim MVN\left([-0.3 \quad 2], \begin{bmatrix} 1.5 & 0.685 \\ 0.685 & 5.5 \end{bmatrix}\right)$$
$$[x_{11} \quad \dots \quad x_{40}]_i' \sim MVN(\mu, I_{30})$$

With μ being a sample of 30 integers sampling with replacement from the integers 2 to 10.

For the complicated DGP the outcome variable is generated as:

$$y = \begin{cases} x_1 - x_1^2 - x_2^2 - 15x_1x_2x_{10} + P_3(x_{10}) \times \begin{bmatrix} 10 & -5 & 0.9 \end{bmatrix}) & \text{if } x_{10} < 2.5 \\ 1750 + 350x_{10} & \text{if } x_{10} \geq 2.5 \end{cases}$$

Where P_n is the polynomial-generating function.

We use this DGP to illustrate, that the BayesGAN is capable of capturing complex high dimensional data sets.

For the experiment we set up the generator network similar to the experiment in Saatchi and Wilson (2017) as a two-layer neural network: 100-1000-41, fully connected, with ReLU activations. Consistently the discriminator network is a two-layer neural network: 41-1000-1, fully connected, with ReLU activations. We place a $\mathcal{N}(\mathbf{0}, I)$ prior on the weights of the BayesGAN.

To compare the BayesGAN to a state of the art synthetic data generator we rely on the conditional CART synthesizer as detailed in Nowok et al. (2016) and implemented in the `synthpop` package for `R`. We use BayesGAN and `synthpop` and generate ten synthetic copies of the original data set with each method. We then calculate the general utility of each of the copies and report the average across the ten synthetic data sets in Figure 1.

Furthermore, Figure 2 shows that BayesGAN as well as `synthpop` are able to capture the complicated interactive structure of the true DGP[9].

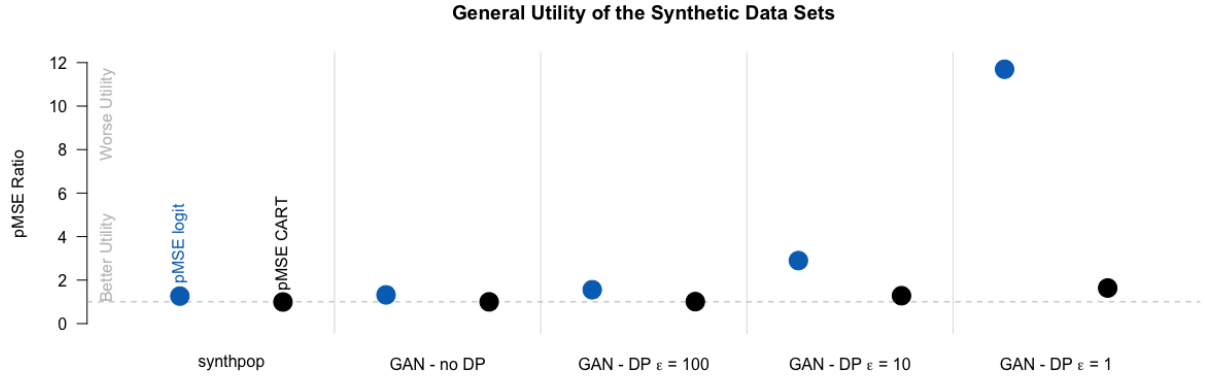

Figure 1: General Utility of the synthetic data sets.

[9]Note that Var1 is the standardized y from the DGP above and Var11 is the standardized x_{10}

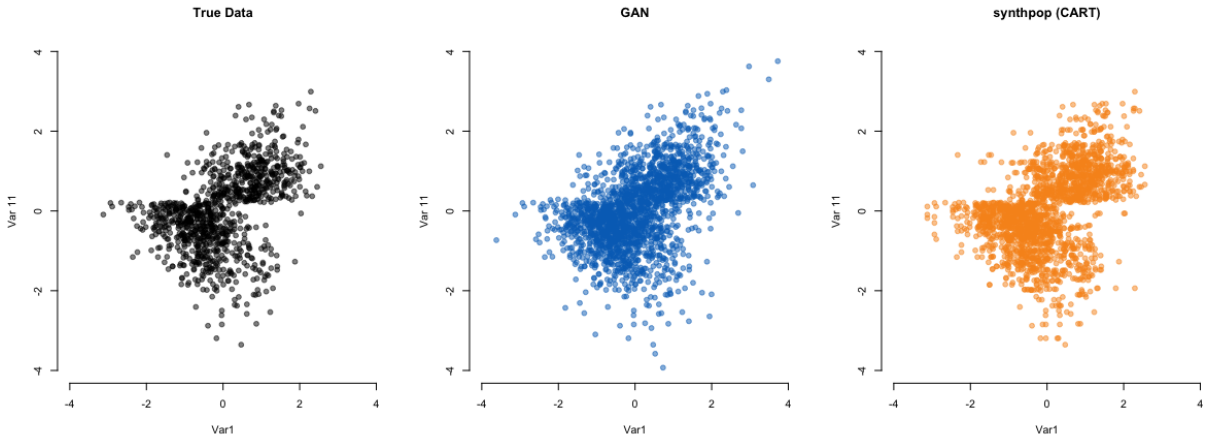

Figure 2: A closer look at the data.

5.2 DP Synthetic Data Sets

To generate the DP synthetic data sets we use the same BayesGAN setup as described above and additionally set ϵ and δ such that additional noise is injected into the gradients of the discriminator during training.

As panels 3, 4 and 5 in Figure 1 show, the general utility of the data decreases with added noise. This means that the logit model and the more flexible cart model can more easily distinguish real from synthetic data. This is what we expect, the more privacy preserving we want our generated data to be.

5.3 Privacy Analysis

tbd

6 Conclusion

In this paper, we investigate the problem of differentially private synthetic microdata. While there exists a large number of methodological approaches to produce synthetic microdata, these methods often rely on heuristic arguments instead of rigorously quantifying the disclosure protection that is offered.

We offer a solution that allows for producing high quality synthetic microdata while at the same time fulfilling the strong criteria of differential privacy. In particular, we employ Bayesian Generative Adversarial Networks to produce high quality synthetic microdata that is also differentially private. Using a Stochastic Gradient Langevin Dynamics sampler, we add noise to the gradients in line with the bounds formulated in Wang et al. (2015), thus generating DP synthetic micro-data with DP-BayesGAN. We demonstrate the usefulness of our approach in an experiment on simulated data with a complex data generating process. Our findings show that BayesGAN can compete with current state of the art models to produce synthetic microdata. We also demonstrate that as expected, the general utility of the synthetic data decreases with introducing the DP-BayesGAN.

Our paper improves existing work in several ways. First, we show that BayesGAN can easily be made differentially private. While there already exist work on differentially private GANs (Beaulieu-Jones et al., 2017; Triastcyn and Faltings, 2018; Xie et al., 2018), we are the first, to the best of our knowledge, who show that DP can be incorporated into the BayesGAN framework. Second, we suggest a way to generate synthetic data at any user defined level of the privacy loss parameters ϵ and δ. Third, we show the usefulness of BayesGAN in an application on the generation of differentially private synthetic micro-data.

In upcoming work, we like to address the following points. First, we are aware that our approach so far does not directly take into account categorical data (Manrique-Vallier and Hu, 2018). Second, we want to provide a deeper investigation of the relationship between privacy level and the data utility. One way to do this would be to follow related work and analyse this relationship using the output of images. Third, our approach is so far not capable to cover structural zeros. These are entries in categorical data that are logically impossible (Manrique-Vallier and Hu, 2018). A similar issue to be addressed are skip patterns (a battery of questions only asked if certain conditions are met).

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016). Deep Learning with Differential Privacy. In *Proceedings of the 23rd ACM Conference on Computer and Communications Security*, Vienna, Austria.

Abowd, J. M. and Lane, J. (2004). New Approaches to Confidentiality Protection: Synthetic Data, Remote Access and Research Data Centers. In Domingo-Ferrer, J. and Torra, V., editors, *Privacy in Statistical Databases*, volume 3050, pages 282–289.

Abowd, J. M. and Woodcock, S. D. (2004). Multiply-Imputing Confidential Characteristics and File Links in Longitudinal Linked Data. In *Privacy in Statistical Databases*, volume 3050, pages 290–297.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein GAN.

Beaulieu-Jones, B. K., Wu, Z. S., Williams, C., and Greene, C. S. (2017). Privacy-preserving generative deep neural networks support clinical data sharing. *bioRxiv*, page 159756.

Bowen, C. M. and Liu, F. (2016). Comparative Study of Differentially Private Data Synthesis Methods.

Caiola, G. and Reiter, J. P. (2010). Random Forests for Generating Partially Synthetic, Categorical Data. *Transactions on Data Privacy*, 3(1):27–42.

Chen, T., Fox, E. B., and Guestrin, C. (2014). Stochastic Gradient Hamiltonian Monte Carlo.

Drechsler, J. (2011). *Synthetic Datasets for Statistical Disclosure Control*, volume 53. Springer.

Drechsler, J. and Reiter, J. P. (2010). Sampling with synthesis: A new approach for releasing public use census microdata. *Journal of the American Statistical Association*, 105(492):1347–1357.

Drechsler, J. and Reiter, J. P. (2011). An Empirical Evaluation of Easily Implemented, Nonparametric Methods for Generating Synthetic Datasets.

Dwork, C. (2006). Differential Privacy. In *33rd International Colloquium on Automata, Languages and Programming, part II (ICALP 2006)*, volume 4052, pages 1–12, Venice, Italy. Springer Verlag.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating Noise to Sensitivity in Private Data Analysis. In *Proceedings of the Third Theory of Cryptography Conference*, pages 265–284.

Dwork, C. and Roth, A. (2013). The Algorithmic Foundations of Differential Privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407.

Dwork, C., Smith, A., Steinke, T., Ullman, J., and Paulson, J. A. (2017). Exposed! A Survey of Attacks on Private Data. *Annual Review of Statistics and Its Application*, pages 61–84.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Nets. *Advances in Neural Information Processing Systems 27*, pages 2672–2680.

Hitaj, B., Ateniese, G., and Perez-Cruz, F. (2017). Deep Models Under the GAN: Information Leakage from Collaborative Deep Learning.

Hu, J., Reiter, J. P., and Wang, Q. (2014). Disclosure Risk Evaluation for Fully Synthetic Categorical Data. In Domingo-Ferrer, J., editor, *Privacy in Statistical Databases*, pages 185–199, Cham. Springer International Publishing.

Huang, C., Kairouz, P., Chen, X., Sankar, L., and Rajagopal, R. (2017). Context-aware generative adversarial privacy. *Entropy*, 19(12):1–32.

Huang, C., Kairouz, P., Chen, X., Sankar, L., and Rajagopal, R. (2018). Generative Adversarial Privacy. In *Privacy in Machine Learning and Artificial Intelligence*.

Kinney, S. K., Reiter, J. P., and Berger, J. O. (2010). Model selection when multiple imputation is used to protect confidentiality in public use data. *Journal of Privacy and Confidentiality*, 2(2):3–19.

Kinney, S. K., Reiter, J. P., Reznek, A. P., Miranda, J., Jarmin, R. S., and Abowd, J. M. (2011). Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database. *International Statistical Review*, 79(3):362–384.

Li, B., Chen, C., Liu, H., and Carin, L. (2017). On Connecting Stochastic Gradient MCMC and Differential Privacy.

Little, R. J. (1993). Statistical analysis of masked data. *JOURNAL OF OFFICIAL STATISTICS*, 9(2):407–426.

Manrique-Vallier, D. and Hu, J. (2018). Bayesian non-parametric generation of fully synthetic multivariate categorical data in the presence of structural zeros. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 181(3):635–647.

Matthews, G. J., Harel, O., and Aseltine, R. H. (2010). Examining the robustness of fully synthetic data techniques for data with binary variables. *Journal of Statistical Computation and Simulation*, 80(6):609–624.

McClure, D. and Reiter, J. P. (2012). Differential privacy and statistical disclosure risk measures: An investigation with binary synthetic data. *Transactions on Data Privacy*, 5(3):535–552.

Montgomery, J. M. and Olivella, S. (2018). Tree-Based Models for Political Science Data. *American Journal of Political Science*, 62(3):729–744.

Nissim, K., Steinke, T., Wood, A., Bun, M., Gaboardi, M., O 'brien, D. R., and Vadhan, S. (2017). Differential Privacy: A Primer for a Non-technical Audience * (Preliminary version). (1237235).

Nowok, B., Raab, G. M., and Dibben, C. (2016). Synthpop: Bespoke Creation of Synthetic Data in R. *Journal of Statistical Software*, 74(11).

Raghunathan, T. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19(1):1–16.

Reiter, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics*, 18(4):531.

Reiter, J. P. (2005). Using CART to generate partially synthetic, public use microdata. *Journal of Official Statistics*, 21(3):441–462.

Reiter, J. P. and Raghunathan, T. E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, 102(480):1462–1471.

Rubin, D. B. (1993). Discussion: Statistical Disclosure Limitation.

Saatchi, Y. and Wilson, A. G. (2017). Bayesian GAN. (Nips):1–16.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016). Improved Techniques for Training GANs. pages 1–10.

Snoke, J., Raab, G., Nowok, B., Dibben, C., and Slavkovic, A. (2016). General and specific utility measures for synthetic data.

Snoke, J., Raab, G., and Slavkovic, A. (2018). General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*.

Sweeney, L. (1997). Weaving Technology and Policy Together to Maintain Coddentiality. *Journal of Law, Medicine and Ethics1*, 25:98–110.

Tran, Ranganath, and Blei, M. (2017). Deep and Hierarchical Implicit Models. *arXiv:1702.08896 [cs, stat]*.

Triastcyn, A. and Faltings, B. (2018). Generating Artificial Data for Private Deep Learning.

Tripathy, A., Wang, Y., and Ishwar, P. (2017). Privacy-Preserving Adversarial Networks.

Van Buuren, S., Brand, J. P., Groothuis-Oudshoorn, C. G., and Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76(12):1049–1064.

Wang, K.-C., Vicol, P., Lucas, J., Gu, L., Grosse, R., and Zemel, R. (2018). Adversarial Distillation of Bayesian Neural Network Posteriors. In *Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden*.

Wang, Y.-X., Fienberg, S. E., and Smola, A. (2015). Privacy for Free: Posterior Sampling and Stochastic Gradient Monte Carlo. pages 1–30.

Welling, M. and Teh, Y.-W. (2011). Bayesian learning via stochastic gradient Langevin dynamics. *Icml'2011*, pages 681–688.

Wu, Y., Yang, F., and Ling, H. (2018). Privacy-Protective-GAN for Face De-identification.

Xie, L., Lin, K., Wang, S., Wang, F., and Zhou, J. (2018). Differentially Private Generative Adversarial Network.