# Using GAN's to generate private synthetic data

Anthony Della Pella

**Abstract**

We utilize a novel mechanism commonly found in the differential privacy literature to generate synthetic data sets that guarantee differential privacy. This work differs from other similar literature in that the theoretical analysis is very simple and streamlined as opposed to more powerful results which rely on privacy analysis of the solver. The work was implemented and some practical questions were explored with the implementation, however the results do not match the state of the art.

## Contents

# 1 Introduction

It is widely accepted that researchers using machine learning models, such as GANs, may accidentally reveal sensitive information about training samples. One prominent example of this is the "Netflix Prize" challenge which two reserachers from UT Austin were able to identify Netflix customers by matching the data sets presented in the contest to other resources available online.

To safeguard the privacy of these samples, various approaches have been proposed. One particularly promising approach utilizes the theoretically sound notion of *differential privacy*. In practice, there are several methodologies to incorporate differential privacy in GAN training. Additionally, differentially private GANs can be used to generate synthetic data for the benefit of scientific endeavors such as data analysis, experimental trials, and more general investigative research requiring the use of ordinarily hard to access data. Both of these notions are discussed below and some theoretical results are discussed alongside implementations demonstrating the efficacy and downsides to such approaches.

# 2 Problem statement

Generically speaking, the problem that we seek to solve is multifaceted. On one hand, sharing of data (in particular, data that may be difficult to obtain) is crucial to speeding up the rate of scientific advancement. One key example of this is in medical studies. It would be beneficial to have the ability to generate synthetic data that still satisfies statistical properties of the true data which can be shared with other researchers.

On the other hand, sensitive data (such as that found in many medical studies) is covered both in legal language and general ethical standards such as HIPAA [1]. Of particular relevance in recent times has been the standards behind releasing data in a way that doesn't allow for identification of individuals. One example of this is the now infamous among privacy researchers "Netflix Challenge" [2]. One has to be careful when releasing synthetic data as well because without some privacy guarantees it may be possible to recover sensitive information from the generated data set.

Unfortunately, studies such as [3] and [4] both fail to provide proofs of demonstrated efficient differentially private data synthesis. On the other hand, [5] provides a theoretically sound approach to GANs with differential privacy guarantees, however the methodology is expensive and involves utilizing a differentially private SGD.

With this "theoretical" emphasis, we can define a (mathematical) problem in the following way:

Given a database $X$ distributed according to some distribution $\mathcal{D}$ we seek to generate a synthetic database $A = \mathcal{M}(X)$ (where $\mathcal{M}$ is some privacy preserving mechanism) such that

1. $A$ follows the distribution $\mathcal{D}$.
2. Privacy is guaranteed. In our case, we seek to guarantee the notion of "differential privacy".

# 3 Methods / Contribution

As we've seen in class, GANs provide a mechanism for generating and augmenting data based on some ground data set [6, 7]. Thus, it seems reasonable to solve the problems posed in Section 2 using this machinery. We rely heavily on the notion of differential privacy, so it is stated here. For a more robust introduction see [8].

**Definition 1** (Differential Privacy)**.** *Given two databases $X$ and $X'$ which differ on exactly one element, a randomized algorithm $\mathcal{M}$ is said to be $(\epsilon, \delta)$-differentially private if for all events $S$*

$$\mathbb{P}(\mathcal{M}(X) \in S) \leq \exp(\epsilon)\mathbb{P}(\mathcal{M}(X') \in S) + \delta. \tag{1}$$

A common way to release databases in a differentially private manner is to noise up the responses to a query. One common distribution we draw this noise from is the Laplace distribution. The Laplace mechanism is defined below:

**Definition 2** (Laplace Mechanism)**.** *The Laplace mechanism preserves $(\epsilon, 0)$-differential privacy according to the following rule: Given a query $f$, return the true result to $f$ plus a noise term $\epsilon_0$ where $\epsilon_0 \sim \text{Lap}(|b|/\epsilon)$.*

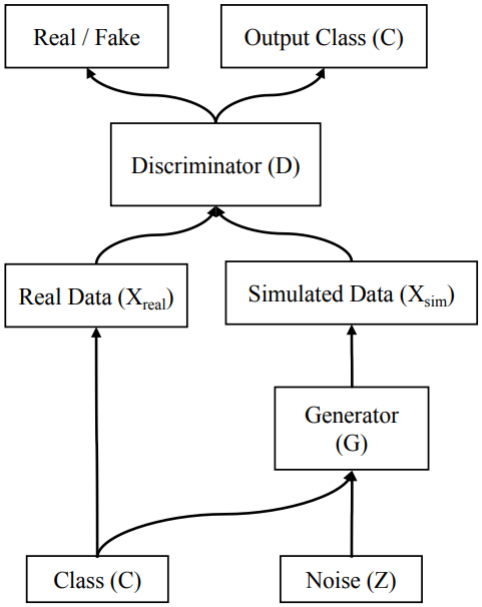
*A few notes about this definition.*

- *The Laplace distribution is defined by density function*

$$\text{Lap}(|b|/\epsilon) = \frac{\epsilon}{2|b|} \exp\left(-\frac{\epsilon|x|}{|b|}\right). \tag{2}$$

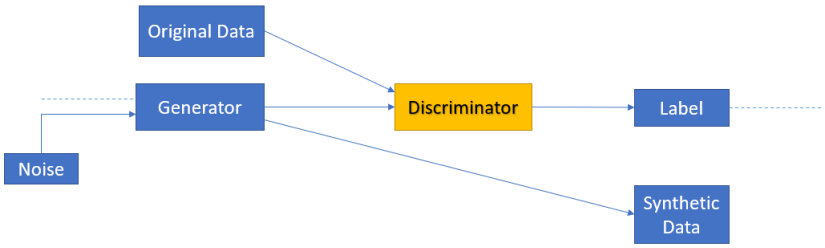
- *We will choose $|b| = \Delta f$ which is the $\ell_1$ distance of $f$ between databases that are $\ell_1$ distance apart.*

With a brief summary of what we will need from the theory of differential privacy out of the way we can discuss the propsed methodology of this project. We will modify a well known GAN structure used to generate synthetic data [4]. In particular, they suggest using the following structure:



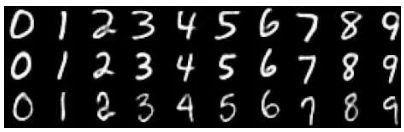

Under this structure, they train the discriminator using the differentially private mechanisms presented first in [5]. The primary limitation of this is that the theoretical guarantees are hard (if not impossible) to prove, and while this private training step performs well, there is no theoretical analysis presented.

To ease the difficulty of the theoretical analysis, and achieve formal privacy guarantees, we propose an alternate strategy. Namely, consider the model structure below:




Here, we will inject some noise into the discriminator using the Laplace mechanism.

(a) Real MNIST Digits



(b) Synthetic MNIST Digits with Privacy

**Fig. 1**: Samples of Real and Synthetic MNIST Digits

# 4 Results

We prove the following theorem that successfully answers the problem outlined in Section 2.

**Theorem 1.** *Given a database $X$ with some distribution $\mathcal{D}$, we generate a synthetic database $A = \mathcal{M}(X)$ such that:*

- $A \sim \mathcal{D}$.
- $\mathcal{M}$ *is a differentially private mechanism – in particular,*

$$\mathbb{P}(\mathcal{M}(X) \in S) \leq \exp(\epsilon)\mathbb{P}(\mathcal{M}(X') \in S) + \delta. \tag{3}$$

In addition to the theoretical result, we implemented the above model using PyTorch. One of the most convenient parts about using GANs to generate synthetic data is that through the course of training, you end up with both a way to generate synthetic data, and a model to classify relevant data. Doing so on the standard MNIST digits dataset (following closely along with [9, 10]) we can generate the images in Fig. 1 using our differentially private Laplace mechanism AND classify the fake data with around 90% accuracy.

# 5 Discussion and Conclusion

The practical results, namely generating a privacy preserving dataset using GANs are very much not novel. In this case, the primary novel contribution in terms of implementation is the mechanism (although more on that later). The real contribution of this work is the simplified theoretical analysis that allows us to prove something about the differential privacy guarantees of the synthetic data set without relying on the [5] framework. In particular, our analysis is much more streamlined.

Given more time, the "moments accountant" method of [5] to measure privacy losses in GANs could've been used to measure the loss in the implementation of the mechanism. Moreover, there wasn't enough time to solve why the model only achieved 90% classification accuracy on the synthetic data. This was the primary exploration done before time ran out on the project.

On the theoretical side, one could attempt to prove tighter privacy guarantees using the moments accountant to keep track of privacy losses at a finer scale. In particular, a similar approach was used in [11] whereby it was found that while the theoretical results hold in general, the privacy guarantees become increasingly bad for

more complex data sets. It is worth noting here that this work was formulated using the Gaussian mechanism (which isn't as easy to analyze for differential privacy), but the results of their model were more accurate than those presented here.

In another direction, one could use this implementation for several purposes. One could be a cryptographic application where users train a GAN based on their profile, and then synthetic data is used as the users key. Similar approaches to cryptographic security have been used (at least theoretically), but none relying on the notion of differential privacy for GANs.

As a last application, it would be interesting to (and the author may pursue) implement this technique on handwritten digits in ones own handwriting style. This would effectively generate a font, and then using GANs one could generate synthetic handwritten text from an individual. The applications of this combined with privacy guarantees on the synthetic data could have applications in things like forgery detection of signed memorabilia.

# Appendix A Proof of Main Theorem

The proofs of the theorem and propositions below are trivial calculations relying only on basic properties of probability, the definition of differential privacy, and basic set theory. Their proofs are contained within the references.

**Theorem 2** (Post Processing [8]). *Let $\mathcal{M}$ be a randomized algorithm that is $(\epsilon, \delta)$-differentially private. Let $g$ be an arbitrary randomized mapping. Then $g \circ \mathcal{M}$ is $(\epsilon, \delta)$-differentially private.*

**Proposition 3** (Noise Layer [11]). *The outputs of a NN with an $(\epsilon, \delta)$-differentially private noise layer $L$ also satisfy $(\epsilon, \delta)$-differential privacy.*

**Proposition 4** (Weights [11]). *The weights $\vec{\omega}_i$ of a NN with $(\epsilon, \delta)$- differentially private labels $y$ are also $(\epsilon, \delta)$-differentially private at each update step $i$.*

*Proof of Theorem 1.* The synthetic data will satisfy the desired privacy guarantee if we can show that the gradient updates themselves satisfy the privacy guarantee [5]. More formally, we need to show that any GAN with generator $G$ and discriminator $\Delta$ with an $(\epsilon, \delta)$-differentially private noise layer $L$ has gradient updates that preserve the same $(\epsilon, \delta)$-privacy guarantee.

This follows by applying Theorem 2 first to guarantee privacy of our discriminator $\Delta$ via Proposition 3 and then apply this result to guarantee the privacy of our generator $G$ by Proposition 4. □

# References


[1] Assistance, H.C.: Summary of the hipaa privacy rule. Office for Civil Rights (2003)

[2] Narayanan, A., Shmatikov, V.: How to break anonymity of the netflix prize dataset. arXiv preprint cs/0610105 (2006)

[3] Torfi, A., Fox, E.A., Reddy, C.K.: Differentially private synthetic medical data generation using convolutional gans. Information Sciences **586**, 485–500 (2022)

[4] Beaulieu-Jones, B.K., Wu, Z.S., Williams, C., Lee, R., Bhavnani, S.P., Byrd, J.B., Greene, C.S.: Privacy-preserving generative deep neural networks support clinical data sharing. Circulation: Cardiovascular Quality and Outcomes **12**(7), 005122 (2019)

[5] Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L.: Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pp. 308–318 (2016)

[6] Figueira, A., Vaz, B.: Survey on synthetic data generation, evaluation methods and gans. Mathematics **10**(15), 2733 (2022)

[7] Tanaka, F.H.K.d.S., Aranha, C.: Data augmentation using gans. arXiv preprint arXiv:1904.09135 (2019)

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

[9] Shetty, R.D.: Gan on mnist with pytorch (2022)

[10] Lin, L.: Generating mnist digit images with generative adversarial network

[11] Triastcyn, A., Faltings, B.: Generating differentially private datasets using gans (2018)