DP-SGD vs PATE: Which Has Less Disparate Impact on GANs?

Georgi Ganev[1 2]

[1]UCL [2]Hazy
georgi.ganev.16@ucl.ac.uk

Abstract

Generative Adversarial Networks (GANs) are among the most popular approaches to generate synthetic data, especially images, for data sharing purposes. Given the vital importance of preserving the privacy of the individual data points in the original data, GANs are trained utilizing frameworks with robust privacy guarantees such as Differential Privacy (DP). However, these approaches remain widely unstudied beyond single performance metrics when presented with imbalanced datasets. To this end, we systematically compare GANs trained with the two best-known DP frameworks for deep learning, DP-SGD, and PATE, in different data imbalance settings from two perspectives – the size of the classes in the generated synthetic data and their classification performance.

Our analyses show that applying PATE, similarly to DP-SGD, has a disparate effect on the under/over-represented classes but in a much milder magnitude making it more robust. Interestingly, our experiments consistently show that for PATE, unlike DP-SGD, the privacy-utility trade-off is not monotonically decreasing but is much smoother and inverted U-shaped, meaning that adding a small degree of privacy actually helps generalization. However, we have also identified some settings (e.g., large imbalance) where PATE-GAN completely fails to learn some subparts of the training data.

1 Introduction

Generative machine learning models, and in particular Generative Adversarial Networks (GANs) (Goodfellow et al. 2014), have received increasing attention from both researchers (Szpruch et al. 2019; Van Der Schaar and Maxfield 2020) and government organizations (Benedetto et al. 2018; NIST 2018b,a; NHS England 2021) as a promising solution to the individual-level data sharing problem. The underlying idea is to train generative models to learn the distribution of the (real) data, generate new high-quality (synthetic) samples from the trained model, and release synthetic, rather than real, data.

However, recent research has shown that generative models, including GANs, may leak sensitive information about the training samples through overfitting and memorization (Carlini et al. 2019; Webster et al. 2019; Meehan, Chaudhuri, and Dasgupta 2020) as well as susceptibility to privacy attacks such as membership inference attacks (Hayes et al. 2019; Chen et al. 2020; Stadler, Oprisanu, and Troncoso 2020). The state-of-the-art approach to protect against such vulnerabilities is training the models to satisfy Differential Privacy (DP) (Dwork et al. 2006; Dwork, Roth et al. 2014). DP mechanisms protect against attempts to infer the inclusion of any record in the training data by bounding their individual contribution, usually through perturbation. In this paper, we will focus on the two most widely used DP techniques for training deep learning models – DP-SGD (Abadi et al. 2016) and PATE (Papernot et al. 2016, 2018).

Even though DP mechanisms guarantee rigorous privacy protection, they degrade the performance of the model. Furthermore, this accuracy drop is likely to be disparate, affecting the underrepresented subpopulations of the data disproportionately more. For example, in the case of deep learning classifiers (Bagdasaryan, Poursaeed, and Shmatikov 2019; Farrand et al. 2020; Suriyakumar et al. 2021) empirically illustrate the disparate degradation caused by DP-SGD. However, comparisons between DP-SGD and PATE are still relatively unstudied in this light, with (Uniyal et al. 2021) doing so for classifiers, and more recently, (Ganev, Oprisanu, and De Cristofaro 2021) demonstrating the said effects in generative models trained on imbalanced tabular data. To fill this gap, we set out to examine and compare two GAN models trained with DP guarantees (DP-WGAN and PATE-GAN) on imbalanced image data (MNIST) in several imbalance settings.

Research Question. Does applying DP-SGD and PATE to GANs lead to similar disparate effects when trained on imbalanced data, or more specifically, on the minority and majority classes in terms of size and accuracy of the resulting synthetic data?

Main Findings. Our experiments could be summarized to:

- Overall, both models exhibit disparity in terms of size and accuracy, but the effects are much smaller for PATE-GAN. Furthermore, PATE-GAN offers a much better privacy-utility trade-off and performs better even for tight privacy budgets (0.5). We believe this is due to the teacher-discriminators setup.
- In terms of size, the two models behave in opposite directions with increased privacy – DP-WGAN "evens" the classes while PATE-GAN increases the imbalance.
- In the presence of a single highly imbalanced class ("8"

reduced to 10% its original size), PATE-GAN fails to learn the whole subpopulation. This is not the case for DP-WGAN.

- Applying some degree of privacy actually serves as regularization to PATE-GAN and helps the performance up to a point, unlike DP-WGAN, for which any privacy protection deteriorates the utility.

2 Preliminaries

In this section, we present some background on DP, GANs, and the two generative models used in our experiments.

2.1 Differential Privacy (DP)

A randomized algorithm $\mathcal{A}$ satisfies (ϵ, δ)-DP if and only if, for any two adjacent datasets D_1 and D_2 (differing in a single record), and all possible outputs S of $\mathcal{A}$, the following holds (Dwork et al. 2006; Dwork, Roth et al. 2014):

$$P[\mathcal{A}(D_1) \in S] \leq \exp(\epsilon) \cdot P[\mathcal{A}(D_2) \in S] + \delta$$

Put simply, one cannot distinguish whether any individual's data was part of the input dataset from observing the algorithm's output. The privacy budget ϵ denotes the level of indistinguishability while δ is a probability of privacy failure. We focus on the two most widely used DP techniques for deep learning – DP-SGD (Abadi et al. 2016) and PATE (Papernot et al. 2016, 2018).

2.2 Generative Adversarial Networks (GANs)

A GAN is a deep learning model consisting of two neural networks, a generator, and a discriminator. They "compete" against each other in a min-max "game," the former produces synthetic data while the latter distinguishes real from generated samples until they reach equilibrium.

DP-WGAN. DP-WGAN (Alzantot and Srivastava 2019) utilizes DP-SGD in place of the standard SGD during training. DP-SGD guarantees privacy by bounding the individual gradients (using clipping and perturbation) of the discriminator and relying on the moments accountant method to track the overall privacy budget.

Furthermore, the model uses the WGAN architecture (Arjovsky, Chintala, and Bottou 2017) to improve stability during training.

PATE-GAN. PATE-GAN (Jordon, Yoon, and Van Der Schaar 2018) modifies the PATE framework for training GANs. The model replaces the standard single discriminator with k teacher-discriminators, trained on disjoint partitions of the real data. In turn, the standard student model is replaced by a student-discriminator trained on noisy (real/synthetic) labels predicted by the teachers. The noisy aggregation is where DP guarantees the privacy protection. Also, the proposed model eliminates the need for publicly available data.

Even though both DP-WGAN and PATE-GAN were initially proposed for tabular data given the remarkable success of GANs on images, we decided not to adjust their architectures.

3 Experimental Evaluation

In this section, we explain our evaluation methodology and discuss the experimental findings.

3.1 Evaluation Methodology

We run all of our experiments on MNIST (LeCun, Cortes, and Burges 2010) and imbalance the class "8" to maintain consistency with (Papernot et al. 2016; Bagdasaryan, Poursaeed, and Shmatikov 2019; Uniyal et al. 2021). First, since the classes are slightly imbalanced, we get rid of all images per class exceeding 5,000. Then, we imbalance the dataset, making "8" the minority/majority class in one of the following three settings[1]:

1. *Minority* – undersample "8" to 10%/25% its original size while keeping the other classes balanced.
2. *Majority* – keep "8" untouched but undersample all other classes to 10%/25% their sizes.
3. *Mixed* – turn "8" into minority/majority class by making it 25% of the largest/smallest class and randomly imbalance all other classes in a uniformly decreasing manner.

We train 5 DP-WGAN and PATE-GAN models for each setting, generate 5 synthetic datasets with a size equal to the input data (per trained model) and report mean and standard deviations. We experiment with privacy budgets (ϵ) of 0.5, 5, 15, and infinity ("non-DP"). We measure the class distributions in the resulting synthetic datasets as well as class recall from classifiers (logistic regression similar to (Ganev, Oprisanu, and De Cristofaro 2021)) trained on the real/synthetic data and tested on put-aside test data. We also report RMSE for sizes and truncated[2] RMSE (TRMSE) for recall weighted by the real sizes in App. B. A summary of the privacy-utility trade-off for both models in all settings is plotted in Fig. 5.

Last, for PATE-GAN in all settings, we also experiment with a different number of teachers – 1, 10, 50, and 100.

We use the same hyperparameters as the original implementations and set $\delta = 10^{-5}$ for all experiments.

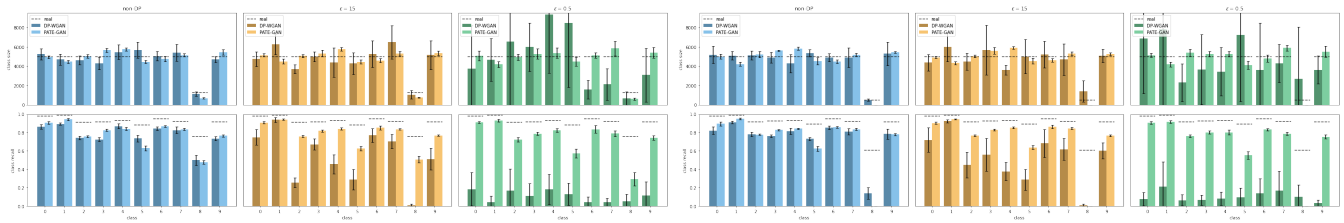
3.2 Minority Class Results

We observe the results in Fig. 1 and Tab. 3. Overall, PATE-GAN exhibits far better performance for both imbalances – it preserves the counts even for lower ϵ budgets, the recall drop is not so acute, and its standard deviation is much lower. DP-WGAN recall looks random for $\epsilon = 0.5$, which means that the classifiers failed to learn anything, most likely due to bad quality of the synthetic data.

Looking at the minority class "8," however, PATE-GAN fails to generate any digits for imbalance 10%. Surprisingly, this phenomenon occurs in the "non-DP" case as well. This could be because the teachers fail to pass samples "8" classified as real to the student even though when applied to classification, PATE is more robust under similar imbalance levels (Uniyal et al. 2021).

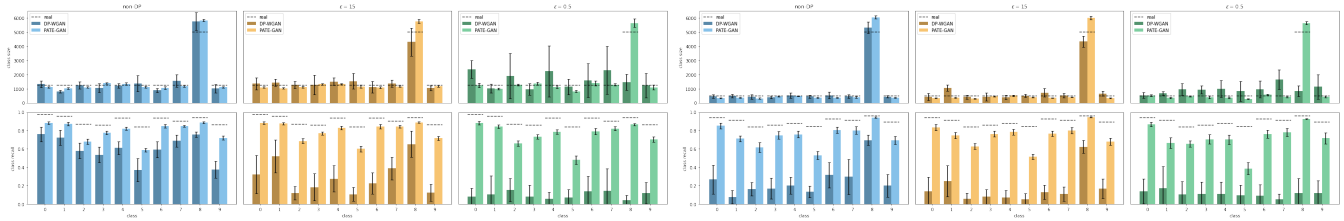
[1] we also experiment with balanced settings in App. A

[2] we do not want to penalize the score if the recall achieved on the synthetic data is better than on the real



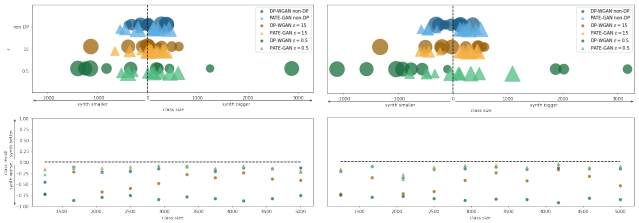
(a) Minority (25%) (b) Minority (10%)

Figure 1: Class size (top) and recall (bottom) of synthetic data generated by DP-WGAN and PATE-GAN trained with different privacy budgets ϵ on imbalanced MNIST, where class "8" is decreased to 25% and 10% of its original size.



(a) Majority (25%) (b) Majority (10%)

Figure 2: Class size (top) and recall (bottom) of synthetic data generated by DP-WGAN and PATE-GAN trained with different privacy budgets ϵ on imbalanced MNIST, where all classes apart from "8" are decreased to 25% and 10% of their original size.



(a) Mixed, Minority (25%) (b) Mixed, Majority (25%)

Figure 3: Class size (top) and recall (bottom) of synthetic data, relative to real, generated by DP-WGAN and PATE-GAN trained with different privacy budgets ϵ on imbalanced MNIST, where "8" is minority/majority class accounting to 25% of the largest/smallest class while all other classes are randomly subsampled in a uniformly decreasing manner.

While expectedly DP-WGAN's performance monotonically drops both in terms of size and recall with decreasing ϵ, PATE-GAN's performance actually increases when DP is applied, e.g., $\epsilon = 15$ and 5 yield better results than "non-DP" in terms of size for both imbalances and in terms of recall for imbalance 10%. This is most likely due to the fact that the teacher-discriminators are exposed to different subsets of the real data, and as result, do not learn exactly the same distributions as well as the noise added to their votes, which further enables generalization.

3.3 Majority Class Results

The results are in Fig. 2 and Tab. 4. For DP-WGAN, there is a significant drop in recall for all undersampled digits, even for the "non-DP" case, unlike PATE-GAN, which again has a very stable behavior.

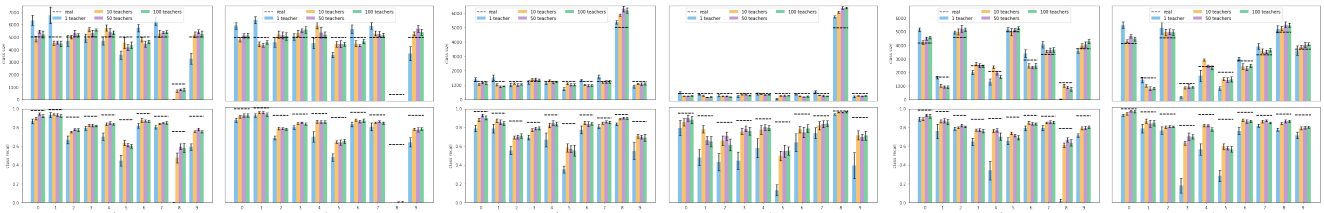
In terms of size, PATE-GAN generates more imbalanced datasets by producing more "8s" and uniformly fewer other digits for all ϵ budgets but again achieves better results for $\epsilon = 15$ than "non-DP." Unlike the minority class setting, no classes "disappear" when the imbalance is increased from 25% to 10%. For DP-WGAN, increasing the imbalance leads to much worse performance, which allows us to speculate that PATE-GAN needs smaller training data to capture the underlying distribution and is more robust to imbalance.

Interestingly, some digits suffer a lot more in terms of recall than others (e.g., "2," "5," "9"), which could be explained because they are visually close to "8" but their sizes are much smaller.

3.4 Mixed Class Results

The results are displayed in Fig. 3 and Tab. 5. First, we note that with increased privacy PATE-GAN, again, has much lower variability/spread in terms of size and a smaller drop in terms of recall. We also clearly observe the opposing size effects the two generative models exhibit, similarly to (Ganev, Oprisanu, and De Cristofaro 2021) – DP-WGAN makes the classes more uniform, i.e., large classes are reduced, and small classes are increased, while PATE-GAN further enforces the imbalance, large classes become even bigger.

In terms of recall, the performance of PATE-GAN for all privacy budgets ϵ drops slightly in a uniform manner across all classes and actually exceeds the "non-DP" DP-WGAN. This observation hints that underrepresented groups do not suffer unequally under the PATE framework when the imbalance is not severe. As for DP-WGAN, the size seems to be an important factor as for $\epsilon = 15$ smaller classes bear more considerable drop, which is in agreement with previous work (Bagdasaryan, Poursaeed, and Shmatikov 2019; Farrand et al. 2020). Recall for $\epsilon = 0.5$ looks random.



(a) Min (25%) (b) Min (10%) (c) Maj (25%) (d) Maj (10%) (e) Mix, Min (25%) (f) Mix, Maj (25%)

Figure 4: Class size (top) and recall (bottom) of synthetic data generated by PATE-GAN for $\epsilon = 5$ in all settings.

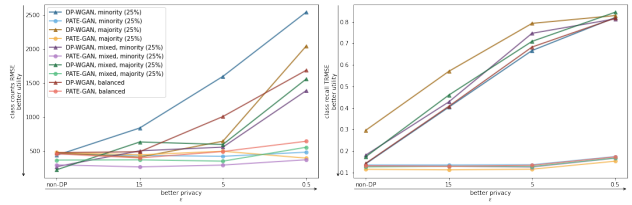


Figure 5: Class size RMSE (left) and recall TRMSE (right) vs ϵ trade-off of synthetic data generated by DP-WGAN and PATE-GAN in all settings.

3.5 Number of Teacher-Discriminators Results

Here, we repeat all the experiments for PATE-GAN with $\epsilon = 5$ in the three settings (Minority, Majority, Mixed) but vary the number of teacher-discriminators. The results can be seen in Fig. 4 and Tab. 6. Overall, we observe that the best performance is achieved when we set the number of teachers to 10 or 50, with some small exceptions. This behavior is fully expected and in line with previous work (Jordon, Yoon, and Van Der Schaar 2018; Uniyal et al. 2021) as having too few/many teachers could fail to generalize or learn at all. For instance, we see that when we have a single teacher, the standard deviation is much higher across the board. One interesting case is the Minority setting with imbalance 10% in Fig. 4b, where increasing the number of teachers improves performance, and we get the best results for 100. This still fails to generate any "8s," unfortunately.

4 Related Work

There is a rich body of research proposing GANs trained with DP guarantees in various domains. They are predominantly GANs variants utilizing modifications of DP-SGD, e.g., DPGAN (Xie et al. 2018) for images and Electronic Health Records (EHR), dp-GAN (Zhang, Ji, and Wang 2018) and DP-CGAN (Torkzadehmahani, Kairouz, and Paten 2019) for images, dp-GAN-TSCD (Frigerio et al. 2019) for tabular and time series data, etc. In comparison, there are only a couple of models incorporation PATE into GANs – PATE-GAN (Jordon, Yoon, and Van Der Schaar 2018) and G-PATE (Long et al. 2021) which ensures DP for the generator by connecting a student generator with an ensemble of teacher discriminators. (Fan 2020) provides a more in-depth survey of various DP GANs.

Researchers have studied the disparate effects of DP mechanisms in different contexts. (Kuppam et al. 2019; Tran et al. 2021b) demonstrate that if fund allocations are based on DP statistics, smaller districts could get more resources at the expense of larger ones, compared to what they would receive without DP.

Prior work has also analyzed the disparate effects of DP-SGD applied to deep neural network classifiers (Bagdasaryan, Poursaeed, and Shmatikov 2019; Farrand et al. 2020; Suriyakumar et al. 2021) trained on imbalanced datasets. They empirically demonstrate that the less represented subgroups in the dataset that suffer lower accuracy to start with lose even more utility when DP is applied. For example, (Farrand et al. 2020) show that even small imbalances and loose privacy guarantees could lead to disparate impacts. (Uniyal et al. 2021) find that CNNs trained with PATE suffer from disparate accuracy drops as well, but less severely than with DP-SGD. There are also papers focusing on learning DP classifiers with fairness constraints (Jagielski et al. 2019; Tran, Fioretto, and Van Hentenryck 2021) and on analyzing the PATE framework from a fairness point of view (Tran et al. 2021a). While these efforts consider discriminative models, we examine generative ones.

Finally, analyzing the DP disparity on generative models, (Cheng et al. 2021) show that training classifiers on balanced DP synthetic images could result in increased majority subgroup influence and utility degradation. Focusing on tabular data, (Pereira et al. 2021) look at single-attribute subgroup fairness and overall classification while (Ganev, Oprisanu, and De Cristofaro 2021) analyze class as well as single/multi-attribute subgroup classification parity over a variety of imbalances and privacy budgets. They find that the disparate effects of DP could be opposing depending on the specific generative model and DP mechanism.

Our work is perhaps closest in spirit to (Uniyal et al. 2021; Ganev, Oprisanu, and De Cristofaro 2021) but we focus on generative models unlike the former and use image data and have a more disciplined approach to constructing the class imbalances, unlike the latter.

5 Conclusion

Through our extensive experiments and analysis, we demonstrate that applying DP methods to GANs to preserve the privacy of the data records could lead to disparate effects on class distributions in the generated synthetic data and the performance of downstream tasks. Overall, PATE exhibits a more desirable behavior than DP-SGD, better privacy-utility trade-off, and while, unfortunately, it still disproportionately affects the minority subparts of the data, it does so to a less severe extend.

References

Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H. B.; Mironov, I.; Talwar, K.; and Zhang, L. 2016. Deep learning with differential privacy. In *ACM CCS*.

Alzantot, M.; and Srivastava, M. 2019. Differential Privacy Synthetic Data Generation using WGANs. https://github.com/nesl/nist_differential_privacy_synthetic_data_challenge/.

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein generative adversarial networks. In *ICML*.

Bagdasaryan, E.; Poursaeed, O.; and Shmatikov, V. 2019. Differential privacy has disparate impact on model accuracy. In *NeurIPS*.

Benedetto, G.; Stanley, J. C.; Totty, E.; et al. 2018. The creation and use of the SIPP synthetic Beta v7. 0. *US Census Bureau*.

Carlini, N.; Liu, C.; Erlingsson, Ú.; Kos, J.; and Song, D. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *USENIX Security*.

Chen, D.; Yu, N.; Zhang, Y.; and Fritz, M. 2020. Gan-leaks: A taxonomy of membership inference attacks against generative models. In *ACM CCS*.

Cheng, V.; Suriyakumar, V. M.; Dullerud, N.; Joshi, S.; and Ghassemi, M. 2021. Can You Fake It Until You Make It? Impacts of Differentially Private Synthetic Data on Downstream Classification Fairness. In *ACM FAccT*.

Dwork, C.; McSherry, F.; Nissim, K.; and Smith, A. 2006. Calibrating noise to sensitivity in private data analysis. In *TCC*.

Dwork, C.; Roth, A.; et al. 2014. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*

Fan, L. 2020. A survey of differentially private generative adversarial networks. In *The AAAI Workshop on Privacy-Preserving Artificial Intelligence*.

Farrand, T.; Mireshghallah, F.; Singh, S.; and Trask, A. 2020. Neither private nor fair: Impact of data imbalance on utility and fairness in differential privacy. In *Workshop on Privacy-Preserving Machine Learning in Practice*.

Frigerio, L.; de Oliveira, A. S.; Gomez, L.; and Duverger, P. 2019. Differentially private generative adversarial networks for time series, continuous, and discrete open data. In *IFIP SEC*.

Ganev, G.; Oprisanu, B.; and De Cristofaro, E. 2021. Robin Hood and Matthew Effects–Differential Privacy Has Disparate Impact on Synthetic Data. *arXiv:2109.11429*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *NeurIPS*.

Hayes, J.; Melis, L.; Danezis, G.; and De Cristofaro, E. 2019. Logan: Membership inference attacks against generative models. In *Proceedings on Privacy Enhancing Technologies*, 1.

Jagielski, M.; Kearns, M.; Mao, J.; Oprea, A.; Roth, A.; Sharifi-Malvajerdi, S.; and Ullman, J. 2019. Differentially private fair learning. In *ICML*.

Jordon, J.; Yoon, J.; and Van Der Schaar, M. 2018. PATE-GAN: Generating synthetic data with differential privacy guarantees. In *ICLR*.

Kuppam, S.; McKenna, R.; Pujol, D.; Hay, M.; Machanavajjhala, A.; and Miklau, G. 2019. Fair decision making using privacy-protected data. *arXiv:1905.12744*.

LeCun, Y.; Cortes, C.; and Burges, C. 2010. MNIST handwritten digit database. *ATT Labs*.

Long, Y.; Wang, B.; Yang, Z.; Kailkhura, B.; Zhang, A.; Gunter, C. A.; and Li, B. 2021. G-PATE: Scalable Differentially Private Data Generator via Private Aggregation of Teacher Discriminators. *NeurIPS*.

Meehan, C.; Chaudhuri, K.; and Dasgupta, S. 2020. A non-parametric test to detect data-copying in generative models. In *AISTATS*.

NHS England. 2021. A&E Synthetic Data. https://data.england.nhs.uk/dataset/a-e-synthetic-data.

NIST. 2018a. 2018 Differential Privacy Synthetic Data Challenge. https://www.nist.gov/ctl/pscr/open-innovation-prize-challenges/past-prize-challenges/2018-differential-privacy-synthetic.

NIST. 2018b. 2018 The Unlinkable Data Challenge. https://www.nist.gov/ctl/pscr/open-innovation-prize-challenges/past-prize-challenges/2018-unlinkable-data-challenge.

Papernot, N.; Abadi, M.; Erlingsson, U.; Goodfellow, I.; and Talwar, K. 2016. Semi-supervised knowledge transfer for deep learning from private training data. *arXiv:1610.05755*.

Papernot, N.; Song, S.; Mironov, I.; Raghunathan, A.; Talwar, K.; and Erlingsson, Ú. 2018. Scalable private learning with pate. *arXiv:1802.08908*.

Pereira, M.; Kshirsagar, M.; Mukherjee, S.; Dodhia, R.; and Ferres, J. L. 2021. An Analysis of the Deployment of Models Trained on Private Tabular Synthetic Data: Unexpected Surprises. *arXiv:2106.10241*.

Stadler, T.; Oprisanu, B.; and Troncoso, C. 2020. Synthetic Data – Anonymization Groundhog Day. *arXiv:2011.07018*.

Suriyakumar, V. M.; Papernot, N.; Goldenberg, A.; and Ghassemi, M. 2021. Chasing Your Long Tails: Differentially Private Prediction in Health Care Settings. In *ACM FAccT*.

Szpruch, L.; Snow, D.; Cohen, S.; Cucuringu, M.; and Horvath, B. 2019. Synthetic data generation for finance and economics. https://www.turing.ac.uk/research/research-projects/synthetic-data-generation-finance-and-economics.

Torkzadehmahani, R.; Kairouz, P.; and Paten, B. 2019. Dp-cgan: Differentially private synthetic data and label generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.

Tran, C.; Dinh, M. H.; Beiter, K.; and Fioretto, F. 2021a. A Fairness Analysis on Private Aggregation of Teacher Ensembles. *arXiv:2109.08630*.

Tran, C.; Fioretto, F.; and Van Hentenryck, P. 2021. Differentially Private and Fair Deep Learning: A Lagrangian Dual Approach. *AAAI*.

Tran, C.; Fioretto, F.; Van Hentenryck, P.; and Yao, Z. 2021b. Decision making with differential privacy under the fairness lens. In *IJCAI*.

Uniyal, A.; Naidu, R.; Kotti, S.; Singh, S.; Kenfack, P. J.; Mireshghallah, F.; and Trask, A. 2021. DP-SGD vs PATE: Which Has Less Disparate Impact on Model Accuracy? *arXiv:2106.12576*.

Van Der Schaar, M.; and Maxfield, N. 2020. Synthetic data: breaking the data logjam in machine learning for healthcare. https://www.vanderschaar-lab.com/synthetic-data-breaking-the-data-logjam-in-machine-learning-for-healthcare/.

Webster, R.; Rabin, J.; Simon, L.; and Jurie, F. 2019. Detecting overfitting of deep generative networks via latent recovery. In *IEEE CVPR*.

Xie, L.; Lin, K.; Wang, S.; Wang, F.; and Zhou, J. 2018. Differentially private generative adversarial network. *arXiv:1802.06739*.

Zhang, X.; Ji, S.; and Wang, T. 2018. Differentially private releasing via deep generative model (technical report). *arXiv:1801.01594*.

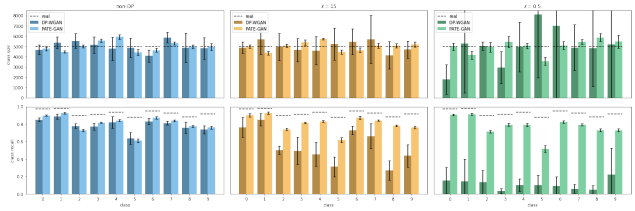


Figure 6: Class size (top) and recall (bottom) of synthetic data generated by DP-WGAN and PATE-GAN trained with different privacy budgets ϵ on balanced MNIST.

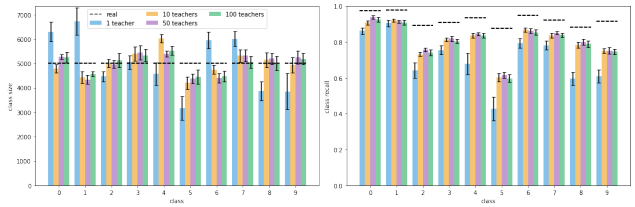


Figure 7: Class size (left) and recall (right) of synthetic data generated by PATE-GAN for $\epsilon = 5$ in balanced settings.

A Balanced Class Results

In this section, we experiment with balanced settings (i.e., we do not artificially imbalance the dataset) to serve as a baseline for the imbalanced settings. The results are displayed in Fig. 6 and 7 as well as Tab. 1 and 2. In short, yet again, PATE-GAN performs better than DP-WGAN – it manages to keep the class size balance and utility with increased privacy and has a much lower deviation. Unlike all other settings, even for the "non-DP" case, the size RMSE on synthetic data produced by PATE-GAN is smaller than DP-WGAN. Interestingly, even with no imbalance, PATE-GAN still benefits from applying a small privacy budget

Imbalance	Balanced			
ϵ	no-DP	15	5	0.5
Size RMSE				
DP-WGAN	**473.38**	487.00	1004.19	1683.49
PATE-GAN	455.78	**396.14**	490.72	640.70
Recall TRMSE				
DP-WGAN	**0.1428**	0.408	0.6830	0.8173
PATE-GAN	0.1316	**0.129**	0.1338	0.1729

Table 1: Class size RMSE and recall TRMSE summary corresponding to App. A, Fig. 6 and 5.

Imbalance	Balanced
ϵ	5
Size RMSE	
1 Teacher	1140.23
10 Teachers	490.72
50 Teachers	425.23
100 Teachers	**354.31**
Recall TRMSE	
1 Teacher	0.2438
10 Teachers	0.1338
50 Teachers	**0.1246**
100 Teachers	0.1350

Table 2: Class size RMSE and recall TRMSE summary corresponding to App. A and Fig. 7.

(e.g., $\epsilon = 15$). Similarly to 3.3, for PATE-GAN, the classes that suffer from the biggest accuracy drop with increased privacy are the ones that are visually similar to each other, "5," "8," "9," and "2." These observations allow us to speculate that the issues and effects observed in Sec. 3 are not only due to the class imbalance but are magnified by it.

B Tables

In this section, we present summary tables of all experiments with imbalanced settings in Sec. 3.

Imbalance	Minority (25%)				Minority (10%)			
ϵ	no-DP	15	5	0.5	no-DP	15	5	0.5
Size RMSE								
DP-WGAN	**433.95**	835.85	1594.41	2537.48	**303.24**	681.29	921.00	1902.04
PATE-GAN	456.18	433.06	**419.82**	480.61	519.16	489.00	**477.81**	558.09
Recall TRMSE								
DP-WGAN	**0.1406**	0.4038	0.6667	0.8201	**0.1396**	0.3913	0.7205	0.8305
PATE-GAN	**0.1353**	**0.1353**	0.1363	0.1743	0.1435	0.1412	**0.1407**	0.1746

Table 3: Class size RMSE and recall TRMSE summary corresponding to Sec. 3.2, Fig. 1 and 5.

Imbalance	Majority (25%)				Majority (10%)			
ϵ	no-DP	15	5	0.5	no-DP	15	5	0.5
Size RMSE								
DP-WGAN	469.23	**407.40**	641.74	2038.68	**242.38**	499.73	935.04	3048.58
PATE-GAN	476.98	441.95	494.13	**393.46**	771.72	749.01	777.01	**477.82**
Recall TRMSE								
DP-WGAN	**0.2961**	0.5702	0.7928	0.8291	**0.5114**	0.5853	0.7326	0.8129
PATE-GAN	0.1152	**0.1128**	0.1157	0.1529	0.1242	0.1248	**0.1237**	0.1537

Table 4: Class size RMSE and recall TRMSE summary corresponding to Sec. 3.3, Fig. 2 and 5.

Imbalance	Mixed, Minority (25%)				Mixed, Majority (25%)			
ϵ	no-DP	15	5	0.5	no-DP	15	5	0.5
Size RMSE								
DP-WGAN	**267.49**	500.89	554.88	1383.69	**221.31**	630.52	594.04	1554.73
PATE-GAN	294.14	**264.34**	291.20	369.03	364.75	367.94	**348.80**	552.61
Recall TRMSE								
DP-WGAN	**0.1809**	0.4283	0.7468	0.8138	**0.1731**	0.4600	0.7093	0.8447
PATE-GAN	0.1304	0.1297	**0.1254**	0.1670	**0.1264**	0.1294	0.1281	0.1674

Table 5: Class size RMSE and recall TRMSE summary corresponding to Sec. 3.4, Fig. 3 and 5.

Imbalance	Min (25%)	Min (10%)	Maj (25%)	Maj (10%)	Mix, Min (25%)	Mix, Maj (25%)
ϵ	5	5	5	5	5	5
Size RMSE						
1 Teacher	1165.98	929.45	**307.98**	**556.47**	589.25	728.41
10 Teachers	**419.82**	477.81	494.13	777.01	**291.20**	**348.80**
50 Teachers	475.89	484.62	733.51	994.75	397.74	460.87
100 Teachers	423.28	**364.32**	677.30	988.51	440.97	410.11
Recall TRMSE						
1 Teacher	0.2576	0.2283	0.2268	0.2862	0.2737	0.2673
10 Teachers	0.1363	0.1407	0.1157	0.1237	0.1254	0.1281
50 Teachers	**0.1269**	0.1411	**0.1135**	**0.1227**	**0.1216**	**0.1241**
100 Teachers	0.1359	**0.1405**	0.1187	0.1281	0.1370	0.1300

Table 6: Class size RMSE and recall TRMSE summary corresponding to Sec. 3.5 and Fig. 4.