Cybersecurity & AI

Freie Universität Berlin

How to Provably Generate Private Synthetic Data

Gerhard Wunder & Benedikt Groß
Cybersecurity and AI Research Group,
Freie Universität Berlin

18.08.2022

Outline

Introduction

Privacy metrics

Latent variable models

Differentially private generative models

Conclusion

Outline

Introduction

Privacy metrics

Latent variable models

Differentially private generative models

Conclusion

The challenge of data privacy

- Who collects data and why?
  - Government: Population statistics, decision making, law enforcement (taxes), ...
  - Research: Develop new methods, find structural imbalances (social sciences), ...
  - Companies: Analyse customer behaviour, market research, improve business, ...
- Privacy risks
  - Membership inference attack: Being identified part of a data set can make you vulnerable
  - Feature inference attack: Sensitive features in collected data
  - ...
  - GDPR: Any feature can compromise someone's privacy
- Remedy(?): Anonymization (or De-Identification)
  - However: Re-identification with additional data possible

Example: Netflix challenge

- In 2007 Netflix released a data set with ~ 500000 records of user data for their Netflix challenge[1]
- The data set was de-identified by removing all personally identifiable information
- Correlating the data set with data from IMDb enabled to sucessfully re-identify many records
- Re-identification was able by matching ratings and/or times when users watched movies on Netflix/wrote reviews on IMDb

Arvind Narayanan and Vitaly Shmatikov. "Robust de-anonymization of large sparse datasets". In: *2008 IEEE Symposium on Security and Privacy (sp 2008)*. IEEE. 2008, pp. 111–125

[1] `https://www.kaggle.com/datasets/netflix-inc/netflix-prize-data`

Another recent example

- Recently, the state police in Berlin stopped detailed reporting on crimes against LGBT community. Before, they had reported anonymized data on a regular base. Why is that?
- It turned out that 'state officials' called for a stop because the released data could potentially be aligned with other data, e.g. from LGBT help agencies

Preußische Allgemeine

29.07.2022

Zeitung für Deutschland · Das Ostpreußenblatt · Pommersche Zeitung

Kriminalität

Fragwürdiger Datenschutz in der Kritik

Berlin: Jeden Tag zwei Attacken auf Homosexuelle – Doch die Täter bleiben meist im Dunkeln

Die Opferberatungsstelle geht davon aus, dass die meisten Taten überhaupt nicht in der Statistik auftauchen. „Das Dunkelfeld liegt unserer Einschätzung nach bei 80 bis 90 Prozent." In diesem Jahr ist eine realistische Einschätzung der Lage sogar noch schwieriger geworden als in der Vergangenheit. Maneo hat nämlich erstmals von der Polizei keine anonymisierten Eckdaten zu Tatzeiten, Tatorten und Ereignissen von Übergriffen mehr erhalten. Das Zurückhalten der anonymisierten Polizeidaten geht laut Berichten lokaler Medien auf den Datenschutzbeauftragten der Berliner Generalstaatsanwaltschaft zurück.

Dieser soll rechtliche Bedenken angemeldet haben, dass durch den Abgleich mit den von Hilfsvereinen gespeicherten Datensätzen die Daten nicht anonymisiert, sondern personenbezogen seien. Schon vor Maneo hatten andere Opferberatungsstellen berichtet, sie hätten auch zu antisemitischen Übergriffen keine Daten mehr erhalten.

- Problem: Now only average (first-order) statistics is provided but no information on Where ? When ? How?

A better privacy primitive: Synthetic data

- So, as just seen: Classical anonymization techniques **fail** to protect privacy. This calls for **private synthetic data** using some clever ML tools

Benefit of synthetic data

- Provides anonymization
- Can be released for third parties to analyze (unlike privacy-preserving data analysis methods!)
- Can be generated in huge quantities

Objectives of this talk

- Revisit standard methods:
  - Even on a synthetic data set, membership inference can be conducted![2]
  - DP-SGD is a standard method but accuracy can be heavily hampered
- Outline a new method:
  - Generative models have an inherent random mechanism that can (should) be exploited

[2] Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. "Synthetic data–anonymisation groundhog day". In: *arXiv preprint arXiv:2011.07018* (2021).

Outline

Introduction

Privacy metrics

Latent variable models

Differentially private generative models

Conclusion

Differential privacy

Differential privacy is a mathematical formulation of privacy

Differential Privacy

A randomized algorithm $\mathcal{A}$ is called $\epsilon \geq 0$ Differentially private, if for all neighboring databases D_1, D_2 and for all subsets $S \subset \mathrm{Im}(\mathcal{A})$ it holds

$$\mathbb{P}\left[\mathcal{A}(D_1) \in S\right] \leq \exp(\epsilon) \cdot \mathbb{P}\left[\mathcal{A}(D_2) \in S\right], \quad (1)$$

where D_0, D_1 are neighboring data sets, i.e. only differ by 1 record.

- Offers strict guarantees, irrespective of attacker knowledge
- Property of the analysis, not the data or the output of an algorithm
- Analysis outcome is essentially the same (for small ϵ), regardless of which data set was used
- Hence it can not reveal any thing about a specific record

Differential Privacy

Cybersecurity & AI   Freie Universität Berlin

(ϵ, δ)-DP

A randomized algorithm $\mathcal{A}$ is called $\epsilon \geq 0$, $\delta \geq 0$ differentially private, if for all neighboring databases D_1, D_2 and for all subsets $S \subset \mathrm{Im}(\mathcal{A})$ it holds

$$\mathbb{P}\left[\mathcal{A}(D_1) \in S\right] \leq \exp(\epsilon) \cdot \mathbb{P}\left[\mathcal{A}(D_2) \in S\right] + \delta \tag{2}$$

Robustness to post-processing

If a randomized mechanism $\mathcal{A}$ is (ϵ, δ)-DP, then so is $F \circ \mathcal{A}$ for any function F.

Composition

The composition of n randomized mechanisms $\mathcal{A}_i$, each (ϵ_i, δ_i)-DP, is $(\sum_i \epsilon_i, \sum_i \delta_i)$-DP.

- Many other variants exist, e.g. Réyni-DP,MI-DP, f-DP, Gaussian-DP[3], ...
- Advanced composition theorems allow to tighter bound the privacy for specific compositions.

[3] Jinshuo Dong, Aaron Roth, and Weijie Su. “Gaussian Differential Privacy”. In: *Journal of the Royal Statistical Society* (2021).

DP as hypothesis test

Membership inference attack

The adversary conducts the following hypothesis test:
H_0 : Training set is D_0.
H_1 : Training set is D_1.

Trade-off function

Hardness of hypothesis test problem is characterized by the trade-off between type I and type II error rates. Let P, Q denote the probability distributions $\mathcal{A}(D_0)$ resp. $\mathcal{A}(D_1)$. Let ϕ be any rejection rule for testing H_0 against H_1. Then the trade-off function $T(P, Q) : [0, 1] \to [0, 1]$ is defined as

$$\alpha \mapsto \inf_{\phi} \{1 - \mathbb{E}_Q[\phi] : \mathbb{E}_P[\phi] \leq \alpha\} \quad (3)$$

f-DP

A randomized algorithm $\mathcal{A}$ is f-differentially private, if

$$T(\mathcal{A}(D_0), \mathcal{A}(D_1)) \geq f,$$

for all neighboring data sets D_0, D_1.

Interpreting privacy guarantees

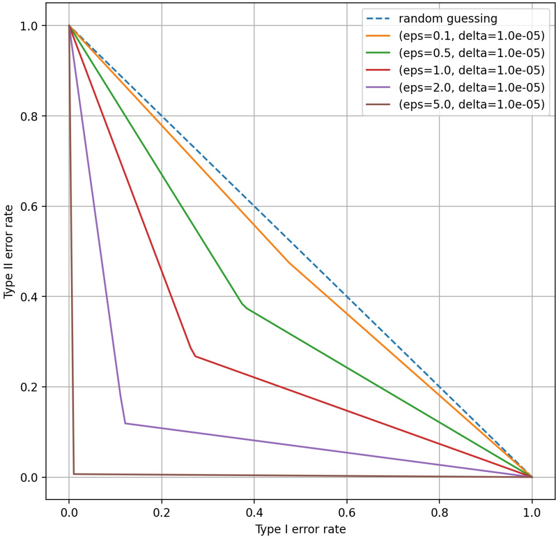


Figure: (ϵ, δ)-DP trade-off curves

Membership Inference Attack (formal)

Membership Inference Attack

Given a target record r, and a synthetic data set $D_{syn} = \mathcal{A}(D)$, generated by a model $\mathcal{A}$ that was trained on private data D, conduct the following hypothesis test:

$$H_0 : r \in D,$$
$$H_1 : r \notin D.$$

Reconstruction Attack

1. Find n nearest points $\{r_1, \ldots, r_n\} \subseteq D_{syn}$

$$\text{For } i = 1, 2, \ldots, n : \quad r_i = \underset{r_{syn} \in D_{syn} \setminus \{r_1, \ldots, r_{i-1}\}}{\arg\min} \|r - r_{syn}\|$$

2. Compute $s = \frac{1}{n} \sum_{i=1}^{n} \|r - r_i\|$
3. Accept H_0, if $s \leq t$ for some predefined threshold $t > 0$.
4. Trade-off curve by varying decision threshold t

Outline

Variational AutoEncoder (VAE)

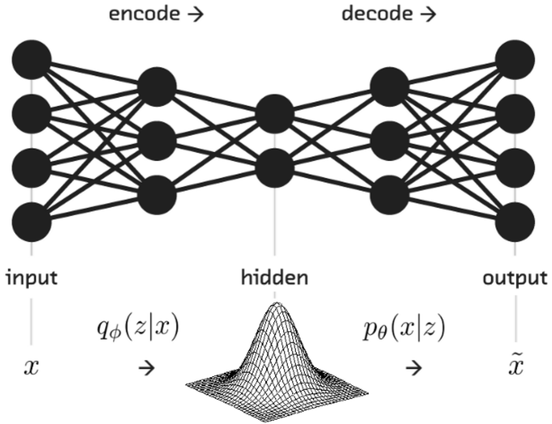


Figure: VAE schematic[4]

[4] image source: `https://towardsdatascience.com/what-the-heck-are-vae-gans-17b86023588a`

Cybersecurity & AI

Freie Universität Berlin

Latent variable models

Model data distribution by

$$p_\theta(x, z) = p(z)p_\theta(x|z), \tag{4}$$

where $p(z)$ is a (simple) prior and $p_\theta(x|z)$ is parameterized by a neural network. Synthetic data is generated as follows:

1. Sample latent variable $z \sim p(z)$
2. Sample data according to $x \sim p_\theta(x|z)$

The intractable posterior

$$p_\theta(z|x) = \frac{p(z)p_\theta(x|z)}{p(x)}$$

is approximated by a parameterized approximate posterior $q_\phi(z|x)$. This gives rise to two joint distributions

$$p_\theta(x, z) = p(z)p_\theta(x|z),$$
$$q_\phi(x, z) = p(x)q_\phi(z|x).$$

A unifying approach to latent variable models

Many generative models can be seen as a Lagrangian of a mutual information optimization objective subject to consistency constraints[5]:

$$\min_{\theta,\phi} \alpha_1 I_{q_\phi}(x;z) + \alpha_2 I_{p_\theta}(x;z) + \Lambda^T \mathcal{D},$$

where

- $I_{q_\phi}(x;z) = \mathbb{E}_{q_\phi(x,z)}[\log q_\phi(x,z) - \log q_\phi(z)p(x)]$ is the MI under $q_\phi(x,z)$ and $I_{p_\phi}(x;z)$ is the MI under $p_\theta(x,z)$,
- $\mathcal{D} = [D_1, \dots, D_m]$ are consistency constraints of the form $D_i = D(q||p)$ for some divergence $D(\cdot||\cdot)$, such that $D_i = 0 \Rightarrow p_\theta(x,z) = q_\phi(x,z)$.
- Here, (p,q) can be any pair of $(p_\theta(x,z), q_\phi(x,z))$, $(p_\theta(x|z), q_\phi(x|z))$, $(p_\theta(z|x), q_\phi(z|x))$, $(q(x), p_\theta(x))$, $(p(z), q_\phi(z))$.
- Λ is a vector of Lagrange multipliers.
- $\alpha_i > 0 \Rightarrow$ minimize MI, $\alpha_i < 0 \Rightarrow$ maximize MI, controls information flow. $\alpha_1 = \alpha_2 = 0$ corresponds to plain ELBO (no MI optimization).

[5] Shengjia Zhao, Jiaming Song, and Stefano Ermon. “The information autoencoding family: A lagrangian perspective on latent variable generative models”. In: *arXiv preprint arXiv:1806.06514* (2018).

Latent variable models

$$
\begin{tikzcd}[column sep=huge]
\mathcal{X} \arrow[r, "q_\phi(z|x)"'] & \mathcal{Z} \arrow[r, "p_\theta(z|x)"'] & \mathcal{X}
\end{tikzcd}
$$

$$
\begin{tikzcd}[column sep=huge]
\mathcal{X} \arrow[r, "q_\phi(z|x)"'] \arrow[r, dotted, bend left, "\min \alpha_1 I_{q_\phi}(x;z)"] & \mathcal{Z} \arrow[r, "p_\theta(z|x)"'] & \mathcal{X}
\end{tikzcd}
$$

$$
\begin{tikzcd}[column sep=huge]
\mathcal{X} \arrow[r, "q_\phi(z|x)"'] & \mathcal{Z} \arrow[r, "p_\theta(z|x)"'] \arrow[r, dotted, bend left, "\min \alpha_2 I_{p_\theta}(x;z)"] & \mathcal{X}
\end{tikzcd}
$$

$$
\begin{tikzcd}[column sep=huge]
\mathcal{X} \arrow[r, "q_\phi(z|x)"'] \arrow[r, dotted, bend left, "\min \alpha_1 I_{q_\phi}(x;z)"] & \mathcal{Z} \arrow[r, "p_\theta(z|x)"'] \arrow[r, dotted, bend left, "\min \alpha_2 I_{p_\theta}(x;z)"] & \mathcal{X}
\end{tikzcd}
$$

Evidence Lower Bound (ELBO)

Cybersecurity & AI

Freie Universität Berlin

- VAE optimizes the evidence lower bound[6]

$$
\begin{aligned}
\log p_\theta(x) &= \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x)\right] \\
&= \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p_\theta(x,z)}{p_\theta(z|x)}\right] \\
&= \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p_\theta(x,z)q_\phi(z|x)}{q_\phi(z|x)p_\theta(z|x)}\right] \\
&= \underbrace{\mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right]}_{\text{ELBO } \mathcal{L}_{\theta,\phi}(x)} + \underbrace{\mathbb{E}_{q_\phi(z|x)}\left[\log \frac{q_\phi(z|x)}{p_\theta(z|x)}\right]}_{\mathrm{KL}\left(q_\phi(z|x)||p_\theta(z|x)\right)\geq 0}
\end{aligned}
$$

- ELBO is a lower bound on the marginal likelihood of the data $\log p_\theta(x) \geq \mathcal{L}_{\theta,\phi}(x)$
- Maximizing the ELBO does two desirable things:
  1. The marginal likelihood is maximized, i.e. the generative model gets better.
  2. The KL distance between approximate and true posterior gets smaller, so $q_\phi(z|x)$ gets better.

[6] Diederik P Kingma and Max Welling. “Auto-encoding variational bayes”. In: *arXiv preprint arXiv:1312.6114* (2013).

ELBO II

$$
\begin{aligned}
\mathcal{L}_{\theta,\phi}(x) &= \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right] \\
&= \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p_\theta(x|z)p(z)}{q_\phi(z|x)}\right] \\
&= \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] - \underbrace{\mathbb{E}_{q_\phi(z|x)}\left[\log \frac{q_\phi(z|x)}{p(z)}\right]}_{\mathrm{KL}\left(q_\phi(z|x)||p(z)\right)}
\end{aligned}
$$

- The first term measures the reconstruction error of data point x.
- The second term draws the approximate posterior $q_\phi(z|x)$ towards the prior $p(z)$[7].

$$
\begin{aligned}
ELBO(\theta,\phi) = \frac{1}{N}\sum_{i=1}^{N} &\mathbb{E}_{q_\phi(z_i|x_i)}\left[\log p_\theta(x_i|z_i)\right] \\
&- \left(\log N - \mathbb{E}_{q(z)}\left[H(q(n|z))\right]\right) \\
&- \mathrm{KL}(q(z)||p(z))
\end{aligned}
$$

[7] Matthew D Hoffman and Matthew J Johnson. “Elbo surgery: yet another way to carve up the variational evidence lower bound”. In: *Workshop in Advances in Approximate Bayesian Inference, NIPS*. vol. 1. 2. 2016.

Illustrative example

- Binarized MNIST data + outlier
- Train different VAE variants on 10000 samples
- 2-dimensional latent space
- Create 10000 synthetic data points by sampling from standard normal prior $p(z)$
- Reconstruction attack as privacy evaluation method (1000 training set members, 1000 non-members)

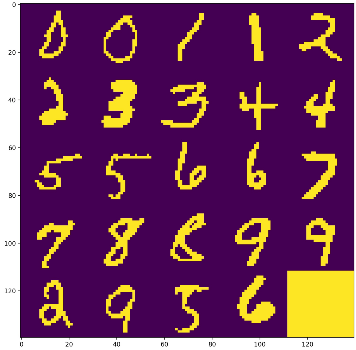


Figure: Examples of regular data points and outlier

Reconstruction attack - VAE

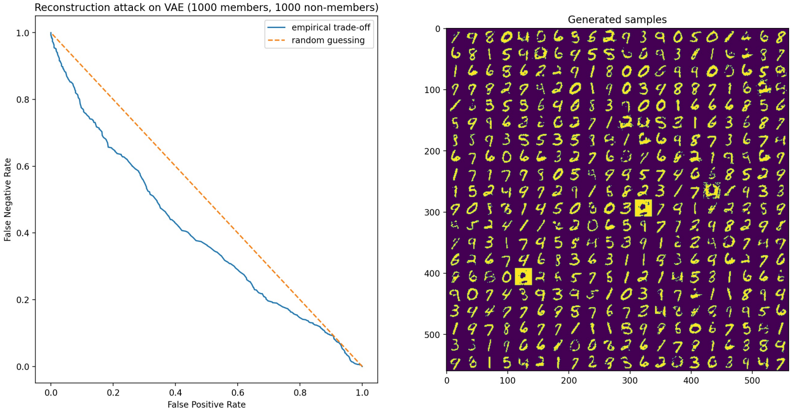
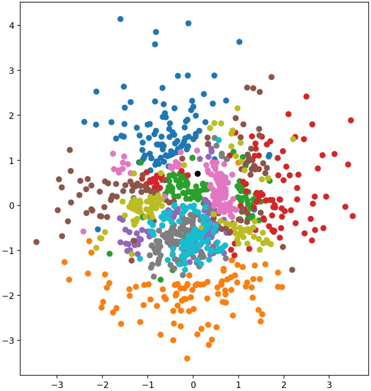


Figure: VAE: reconstruction attack trade-off (left), generated samples (right)

Latent space representation - VAE



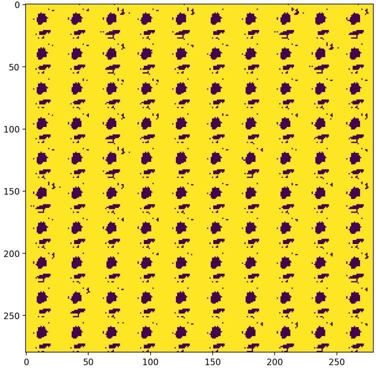


Figure: Latent embeddings of training data (left), reconstruction of outlier (right)

Intermediate conclusion I

- Synthetic data alone is not sufficient to protect privacy
- While normal data is somehow protected, outliers are very susceptible to privacy breaching attacks[8]
- Mixed/categorical data is easier to attack (MNIST does not contain a lot of sample-specific information)

[8] Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. "Synthetic data–anonymisation groundhog day". In: *arXiv preprint arXiv:2011.07018* (2021).

Outline

Differentially private generative models

- Explicit privacy protection measures are necessary also for generative models
- Privacy protection degrades data utility, there is a privacy/utility trade-off
- Variational autoencoder trained with DP-SGD.

Differentially private stochastic gradient descent[9]

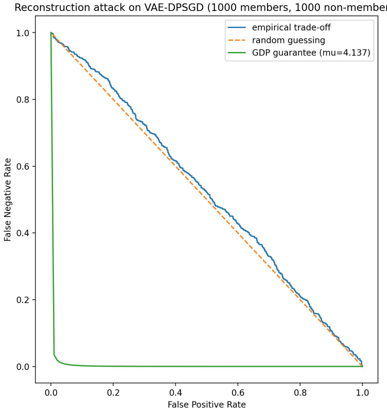
1. Clip per sample gradients
2. Add suitably scaled Gaussian noise to clipped gradients
3. Average noisy gradients
4. Gradient descent step
5. Track privacy budget

Careful!

Privacy protection degrades utility!

[9] Martin Abadi et al. “Deep learning with differential privacy”. In: *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 2016, pp. 308–318.

Reconstruction attack - VAE-DPSGD



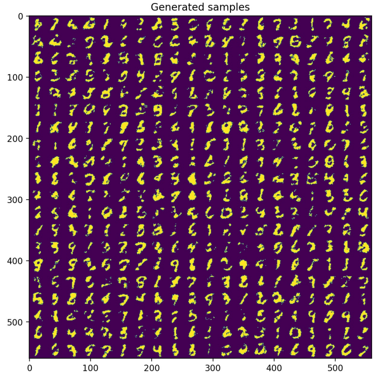


Figure: VAE-DPSGD: reconstruction attack trade-off (left), generated samples (right)

Latent space representation - VAE-DPSGD

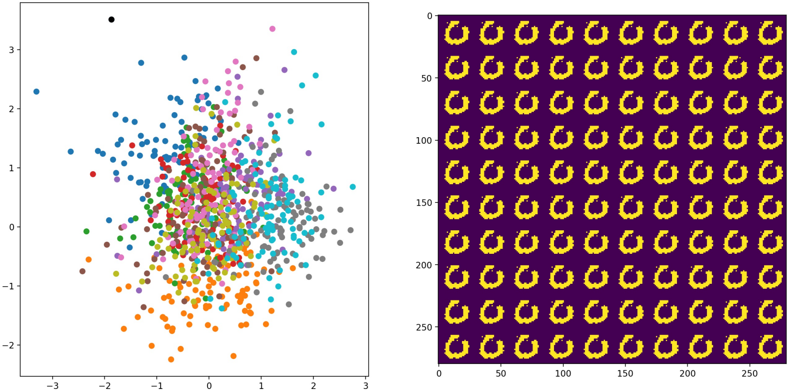


Figure: Latent embeddings of training data (left), reconstruction of outlier (right)

Intermediate conclusion II

- Artificially adding noise to the training is cumbersome and detrimental to the convergence of VAE
- Tight privacy budgets don't allow for many training epochs $\rightarrow$ sample quality can not be controlled properly
- Variational autoencoders already incorporate stochastic mechanisms per default
- Can we leverage on the stochastic sampling procedure inherent in latent variable models for privacy?

Idea

- Constrain encoder/decoder mechanism wrt. continuity modulus
- While typical data is already well protected, outliers are in danger of being identified, even in synthetic data
- Implicitly distinguish between 'data manifold' and outliers

Lipschitz-constrained VAE

Lipschitz continuity

A mapping f between normed spaces is called Lipschitz continuous, if

$$\forall x, y \in \mathrm{dom}(f) : \|f(x) - f(y)\| \leq L\|x - y\|,$$

with the appropriate norms for domain and image space. For a differentiable function f this becomes

$$\|f(x) - f(y)\| \leq \sup_z \|\nabla f(z)\| \|x - y\|.$$

- Constraining the Lipschitz constant of a function encourages simpler mappings, since inputs that are close can not be mapped far away from each other.
- VAE-GP: Add gradient penalty on the decoder to (β-)VAE objective:

$$\min_{\theta,\phi} \frac{1}{N} \sum_{i=1}^{N} \|x - d_\theta(e_\phi(x))\|^2 + \beta \, \mathrm{KL}(e_\phi(x)||p(z))$$
$$+\gamma \frac{1}{M} \sum_{j=1}^{M} \max(\|\nabla_z d_\theta(z_j)\|^2 - L), 0)^2, \tag{5}$$

for random points $z_1 \ldots, z_M$ in latent space.

Lipschitz-constrained VAE

- Our main contribution is to appropriately regularize the VAE objective (GP-VAE)

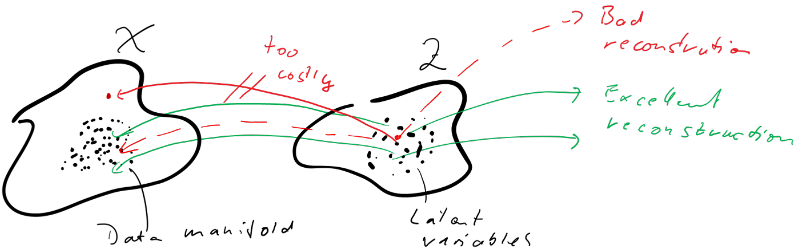
Theorem

Suppose both neural networks have infinite model power. Suppose that the decoder variance is bounded below and that the deterministic part of the decoder is L/2-Lipschitz uniformly in the model parameters and for a subset $\mathcal{D} \in \mathcal{Z}$ with $\mathbb{P}\left[Z \in \mathcal{D}\right]] \geq 1 - \delta$. Then, the encoder is $(L, \delta$-DP and the sum of Type I and Type II errors is bounded below below by $1 - L$.

- Step 1: Prove 'a posteriori sampling' style theorem[10]
- Step 2: Show (surprising) equivalence of Lipschitz and a posteriori condition. Show encoder DP (not enough though!)
- Step 3 (necessary): Formulate the 'right' hypothesis test and use equivalence of MI-DP and DP, Pinsker inequality and Neyman-Pearson theory

[10] Yu-Xiang Wang, Stephen Fienberg, and Alex Smola. "Privacy for free: Posterior sampling and stochastic gradient monte carlo". In: *International Conference on Machine Learning*. PMLR. 2015, pp. 2493–2502.

Reconstruction attack - VAE-GP



Reconstruction attack - VAE-GP

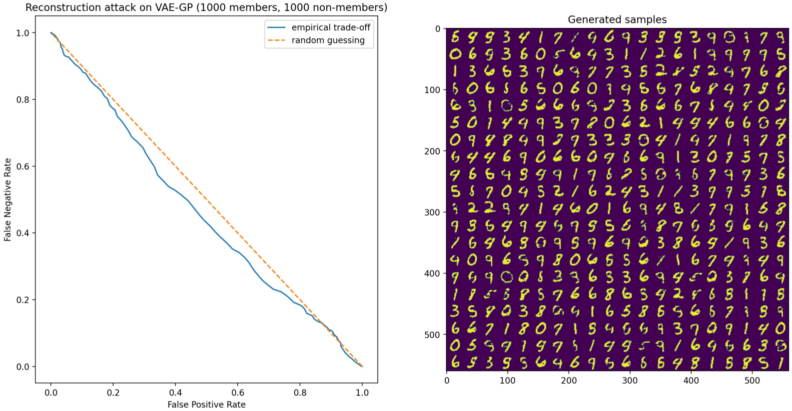
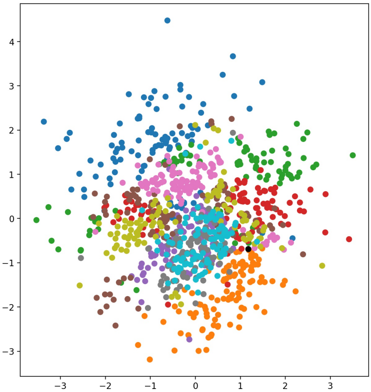


Figure: VAE-GP: reconstruction attack trade-off (left), generated samples (right)



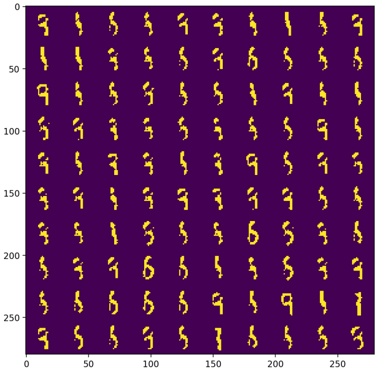


Figure: Latent embeddings of training data (left), reconstruction of outlier (right)

Outline

Final conclusion

Cybersecurity & AI
Freie Universität Berlin

- Ubiquitous data collection calls for privacy preserving analysis methods
- Classical anonymization techniques are not sufficient!
- Also VAE and DP-VAE fail in many cases:
  - Privacy-Utility trade-off not satisfying
  - Analysis has to be conducted by the data holder
- New method for generating synthetic data discussed which exploits the inherent VAE randomized mechanisms (not on top of it)
- Many open problems:
  - We ran algorithms on tabular data (adult set): Tradeoffs even worse
  - Privacy parameters difficult to adjust possibly, data curation needed
  - Lipschitz contraints notoriously difficult to guarantee, relation to robustness theory (see Barret et. al "Certiably Robust Variational Autoencoders", arxiv 2022)

Thank You for Your Attention! [11]

[11] g.wunder@fu-berlin.de or benedikt.gross@fu-berlin.de