Privacy-Preserving Adversarial Network (PPAN) for Continuous non-Gaussian Attributes

Mohammadhadi Shateri
Department of Systems Engineering
École de technologie supérieure, Montreal, Canada
Email: mohammadhadi.shateri@etsmtl.ca

Fabrice Labeau
Department of Electrical and Computer Engineering
McGill University, Montreal, Canada
Email: fabrice.labeau@mcgill.ca

Abstract—A privacy-preserving adversarial network (PPAN) was recently proposed as an information-theoretical framework to address the issue of privacy in data sharing. The main idea of this model was to use mutual information as the privacy measure and adversarial training of two deep neural networks, one as the mechanism and another as the adversary. The performance of the PPAN model for the discrete synthetic data, MNIST handwritten digits, and continuous Gaussian data was evaluated and compared to the analytically optimal trade-off. In this study, we evaluate the PPAN model for continuous non-Gaussian data, i.e. smart meters data, where lower and upper bounds of the privacy-preserving problem are used. These bounds include the Kraskov (KSG) estimation of entropy and mutual information that is based on the k-th nearest neighbor. In addition to the synthetic data sets, a practical case for hiding the actual electricity consumption from smart meter readings is examined. The results show that for continuous non-Gaussian data, the PPAN model performs within the determined optimal ranges and close to the lower bound.

Index Terms—Privacy-preserving adversarial network, Adversarial training, Mutual information, Continuous non-Gaussian attributes, Smart meters.

I. INTRODUCTION

In the era of data mining, the explosion of data collection is enabling researchers to increasingly learn and develop sophisticated models on real-life data sets. There are thousands of research works in many areas including medicine, education, electrical power, or business in which data mining tools are applied to large published data sets. Unfortunately, these data sets may include private information about individuals. Therefore, data collection and curation organizations should sanitize the data before releasing/sharing it. The privacy-preserving data sharing models aim at generating a released data Z from an observation W which is a sanitized version of useful/public attribute Y and at the same time minimizes the possibility of inferring about the sensitive/private attribute X. There are several studies that use the well-known differential privacy to ensure privacy protection [1], [2]. It is shown in many studies that although effective in preserving privacy, differential privacy degrades the utility of the data extensively [3]. However, it is out of the scope of our study.

An information-theoretical approach based on distorting useful data was proposed in [4]–[13]. A real-time scenario using directed information was studied in [8], [9] and was applied to smart meter cases while the case where an attacker is equipped with additional knowledge (side information) and its impact on the privacy-distortion curve was proposed in [14]. In other studies, the non-causal scenarios were considered in which using the Kullback-Leibler(KL) divergence, the privacy risk was quantified as the mutual information between private attribute X and released/shared data Z. In [4] for the Gaussian attributes, they used a numerical procedure (a modification of the steepest descent algorithm) to solve this privacy-distortion trade-off. The results were compared with the quadratic-Gaussian lower and upper bounds of mutual information. However, it was not discussed how including the private attribute in observation samples —W = (X, Y)—can boost performance. In another study [7] using the same information-theoretical approach, a model including adversarially trained deep neural networks was presented and titled privacy-preserving adversarial networks (PPAN). The idea of adversarial training of neural networks was first introduced in [15]. The PPAN model was applied to both useful data as observation (W = Y) and full data as observation (W = (X, Y)). The results of the PPAN model for discrete synthetic data, MNIST handwritten digits, and continuous Gaussian data were examined and found to be very close to the optimal trade-off that was derived analytically. However, the performance of the PPAN model for continuous non-Gaussian data cannot be properly assessed using this framework. The reason could be presented as the fact that generally there is no closed-form solution for the privacy-distortion trade-off. In practical applications, there are several cases, such as those related to smart meters [8], where attributes come from a continuous range but with a non-Gaussian distribution. Therefore, the natural challenge raised here is how the PPAN model can be applied and evaluated in these cases?

In our study, a method for evaluating the PPAN model for continuous non-Gaussian attributes is provided. In this method, first a lower and upper bound on the performance of the PPAN model are determined; and then the estimated privacy-distortion trade-off of the network is compared with these bounds. These bounds include the Kraskov–Stögbauer–Grassberger (KSG) estimation of entropy and mutual information [16]. This study is important in the field of privacy in (big) data sharing since unlike most...
available studies which are based on a data-driven approach, it provides theoretical insights (although by considering simplified assumptions such as Gaussian approximation). In addition to synthetic data cases, a practical scenario for privacy in smart meter data sharing is considered. This can motivate the applicability of the proposed setting for other practical cases such as those that include images.

The rest of the paper is organized as follows. In Section II we present a background on the privacy-distortion problem formulation and we tackle this problem using deep neural networks in an adversarial fashion. Then, in Section III, the theoretical results are proposed and discussed in detail. Experimental results are presented and discussed in Section IV. Finally, concluding remarks are given in Section V.

II. BACKGROUND: PRIVACY-DISTORTION TRADE-OFF FORMULATION

Consider private data $X \in \mathbb{R}^{m_X}$, useful data $Y \in \mathbb{R}^{m_Y}$, and observed data $W \in \mathbb{R}^{m_W}$ which are modeled as jointly distributed random variables by data model $P_{W|X,Y}$ over space $\mathcal{W} \times \mathcal{X} \times \mathcal{Y}$. The observed data $W$ could be just an observation of the useful data $Y$, i.e. $W = Y$ or a combination of the useful and private data as $W = (Y,X)$. The goal of the privacy-aware data sharing model is designing a mechanism $P_{W|Z}$ to generate released data $Z \in \mathbb{R}^{m_Y}$ (by observing $W$) that shares minimum information with the private data $X$ while provides the maximum utility of the useful data $Y$. In this study, the sensitive information leakage is quantified by mutual information $I(X;Z)$ between released data $Z$ and private attribute $X$ while the utility is inversely quantified as expected distortion $d(Y,Z)$ between released data $Z$ and useful attribute $Y$. Therefore, we are interested in the following optimization problem:

**Problem A.** Given the useful data $Y$, private data $X$, and observed data $W$, the privacy-aware data sharing mechanism $P_{Z|W}$ can be formulated as the solution to the following optimization problem:

$$
\begin{align*}
\mathbb{E}(\delta) &= \text{minimize} \quad I(X;Z) \\
\text{subject to} \quad \mathbb{E}[d(Y,Z)] &\leq \delta
\end{align*}
$$

(1)

where the parameter $\delta$ denotes the maximum allowed distortion that the model can tolerate on the shared data $Z$.

**Lemma 1.** Consider two random variables $X$ and $Z$ with conditional distribution $P_{X|Z}(x|z)$. For any conditional distribution $Q_{X|Z}$ the mutual information $I(X;Z)$ can be approximated as follows:

$$
I(X;Z) = H(X) + \max_{Q_{X|Z}} \mathbb{E}\left[ \log Q_{X|Z}(x|z) \right]
$$

(2)

where the expectation $\mathbb{E}[\cdot]$ is with respect to $P_{X|Z}(x|z)$

**Proof.** The Lemma 1 can be proved using variational lower bound of mutual information [17]. For the details of the proof, the reader is referred to [7].

**Proposition 1.** The privacy-distortion optimization problem $A$ can be written as the following minimax optimization problem:

$$
\min_{P_{Z|W}} \max_{Q_{X|Z}} \mathbb{E}\left[ \log Q_{X|Z}(x|z) \right] + \lambda \left( \max_{0 \leq d(Y,Z) \leq \delta} \mathbb{E}[d(Y,Z)] - \delta \right)^2
$$

(3)

where $\lambda > 0$ is Lagrange penalty coefficient.

**Proof.** By substituting equation (2) in (1) and dropping constant term $H(X)$ the privacy-distortion optimization problem (1) can be written as the minimax problem (3).

It should be noted that, depending on application, different distortion metrics $d(\cdot)$ including $Pr[Y \neq Z]$ and $\mathbb{E}[\|Y-Z\|^2]$ can be used. The minimax problem (3) can be interpreted in an adversarial training context in which the adversary network uses the released data $Z$ to estimate the posterior $Q_{X|Z}(x|z)$ by maximizing the log-likelihood $\mathbb{E}[\log Q_{X|Z}(x|z)]$ while the releaser(mechanism network) attempts to prevent that by minimizing this log-likelihood (See Fig. 1). Following equation (3), the mechanism and adversary networks are realized as a neural network (NN) with the following loss functions:

$$
\begin{align*}
Loss_A &= -\mathbb{E}\left[ \log Q_{X|Z}(x|z) \right] \\
Loss_M &= \lambda \left( \max_{0 \leq d(Y,Z) \leq \delta} \mathbb{E}[d(Y,Z)] - \delta \right)^2 - Loss_A
\end{align*}
$$

(4)

where $Loss_A$ is the loss function of the adversary network and $Loss_M$ is the loss function of the mechanism network.

![Fig. 1. Privacy preserving model in an adversarial fashion. The seed noise $U$ is concatenated to the observed data $W$ and is used to provide randomization for the mechanism network.](image.png)

In [7] it is assumed that the posterior $Q_{X|Z}(x|z)$ is Gaussian with mean $\mu(Z)$ and covariance matrix $\Sigma_{\mu}(Z)$. Applying this assumption to equation (4) and approximating expectation $\mathbb{E}(\cdot)$ empirically by averaging over $n$ samples (considering the law of large number), equation (4) can be rewritten as follows:

$$
\begin{align*}
Loss_A &\approx -\frac{1}{n} \sum_{i=1}^{n} \log N(x_i; \mu(z_i), \Sigma(z_i)) \\
Loss_M &\approx \frac{1}{n} \sum_{i=1}^{n} \lambda \left( \max_{0 \leq d(y_i,z_i) \leq \delta} \mathbb{E}[d(Y,Z)] - \delta \right)^2 - Loss_A
\end{align*}
$$

(5)

where $N(x,\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^\epsilon det(\Sigma)}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$ and $\Sigma$ is covariance matrix. The model presented in Fig. 1 and includes two neural networks with loss functions (5) is called Privacy-Preserving
Adversarial Networks (PPAN) model [7]. During the training, the output of the Mechanism (releaser) and the output of the Adversary are used to update the parameters of each network. But after training the model, just the releaser is used.

III. Theoretical Results

To assess the PPAN model for cases where attributes are continuous, generally, two cases including Gaussian and non-Gaussian can be considered. The continuous Gaussian case is discussed in detail and with many examples in [7].

Proposition 2. Assume useful data Y and private attribute X are two continuous Gaussian random variables.

(a) For the "useful data only" case where observed data \( W = Y \), and X and Y have zero mean value, the problem A has a closed-form solution as follows:

\[
I_{useful}(X;Z) = \max \left\{ 0, \frac{1}{2} \log \frac{1}{1 - \rho^2 + \frac{\delta}{\sigma^2_Y}} \right\}
\]  
(6)

(b) For the "full data" case, i.e. \( W = (X,Y) \), and unit variance X and Y, the problem A has a closed-form solution as follows:

\[
I_{full}(X;Z) = \begin{cases} 
0, & \delta \geq \rho^2 \\
\frac{1}{2} \log \frac{1}{1 - \rho^2} - \frac{\delta}{\sigma^2_Y}, & \delta < \rho^2
\end{cases}
\]  
(7)

where \( \rho \) is the correlation coefficient between X and Y and \( \sigma^2_Y \) is the variance of Y.

For the proof of the Proposition 2, the reader is referred to [7]. It should be mentioned that these theoretical solutions proposed in Proposition 2 are used as the baseline for the case of the Gaussian attributes. In this case, after generating the released data Z using the PPAN model, the actual distortion and the leakage of the network are calculated using (8).

\[
\text{Distortion} = \frac{1}{n} \sum_i |y_i - z_i|^2; \quad \text{Leakage} = -\frac{1}{2} \log(1 - \rho^2_{xz})
\]  
(8)

where \( \rho_{xz} \) is the correlation coefficient between private attribute X and released data Z while the leakage in equation (8) is the mutual information of bivariate Gaussian [18].

A. Lower and upper bounds on Leakage-Distortion Trade-off

For the cases where the private and useful attributes X and Y are continuous but not Gaussian, the equations (6) and (7) for theoretical leakage calculations (baseline) and equation (8) for network leakage calculation do not hold. Therefore, in this study for the non-Gaussian cases, the amount of network leakage is calculated using a non-parametric estimation of the mutual information. Moreover, a lower and upper bound for the leakage-distortion tradeoff \( \mathbb{R}(\delta) \) in problem A is theoretically defined using a non-parametric estimation of entropy and mutual information as a baseline. To this end, the Kraskov–Stögbauer–Grassberger (KSG) estimation of entropy and mutual information is used.

**Definition 1.** Consider a random sample of size N from d-dimensional random vector \( X = (X_1, ..., X_d) \). The Kraskov–Stögbauer–Grassberger (KSG) estimation of differential entropy \( \hat{H}_{knn}(X) \) based on the k-nearest-neighbor distances can be stated as follows [16]:

\[
\hat{H}_{knn}(X) = \frac{d}{N} \sum_{i=1}^{N} \log(e_i) + \log(c_d + \psi(N) - \psi(k))
\]  
(9)

where \( e_i \) is twice the distance from \( x_i \) to its k-th neighbor, \( \psi(m) = \frac{d}{2} \ln(T(m)) \) is digamma function, and \( c_d \) is the volume of a unit d-dimensional ball which for the Euclidean norm \( c_d = \frac{\pi^{d/2}}{\Gamma(d/2 + 1)} \).

**Definition 2.** Considering the KSG estimation of the differential entropy in Definition 1 and the expansion of the mutual information as \( I(X;Z) = H(X) + H(Z) - H(X,Z) \), the KSG estimation of \( I(X;Z) \) is calculated as follows [16]:

\[
\hat{I}_{knn}(X;Z) = \psi(N) + \psi(k) - \psi(n_i + 1) + \psi(n_i + 1) > 0
\]  
(10)

where \( <.,.> \) denotes to averages both over \( i \in [1, ..., N] \) and over all realization of random samples, \( n_i(i) \) is the number of points \( x_i \) whose distance from \( x_i \) is less than \( e(i)/2 \), \( n_i(\approx) \) is the number of points \( z_i \) whose distance from \( z_i \) is less than \( e(i)/2 \), and \( e(i) \) is a maximum \( e_i(i), e_i(i) \).

For the rest of this study, two general cases for the observed data W including "useful data only" \( W = Y \) and "full data" \( W = (X,Y) \) are considered. Since for the continuous non-Gaussian attributes, the optimization problem A does not have a closed-form solution, a lower and upper bound of the achievable mutual privacy guarantee will be derived using the KSG estimators. Without loss of generality, it is assumed that all the attributes are zero mean.

**Theorem 1.** For the optimization problem A, given random variable X and Y with standard deviation \( \sigma_X \) and \( \sigma_Y \) and the correlation coefficient \( \rho \) and assuming observed data W \( \mathcal{Y} \), we can say:

(a) Lower bound:

\[
\mathbb{R}(\delta) \geq \max \left\{ 0, \hat{I}_{knn}(X;Z) - 0.5 \log \left( 2 \pi e \frac{\delta^2}{\sigma^2_Y} \left( \frac{\rho^2}{\sigma^2_Y} + (1 - \rho^2)^2 \right) \right) \right\}
\]  
(11)

(b) Upper bound:

\[
\mathbb{R}(\delta) \leq \hat{I}(X;Y) \times (1 - \frac{\delta}{\sigma^2_Y})
\]  
(12)

**Proof.** (a) Considering the mutual information \( I(X;Z) \) we can say:

\[
I(X;Z) = H(X) - H(Z|X) = H(X) - H(X - \rho \sigma_X Z/\sigma_Y |Z)
\]

\[
\geq H(X) - H(X - \rho \sigma_X Z/\sigma_Y )
\]

\[
\geq H(X) - H(N(0, \mathcal{E}[X - \rho \sigma_X Z/\sigma_Y ]^2))
\]

\[
= H(X) - 0.5 \log \left( 2 \pi e \mathcal{E}[X - \rho \sigma_X Z/\sigma_Y ]^2 \right)
\]  
(13)
where (i) holds because conditioning reduces entropy and (ii) is true since, for a given value of the second moment, the zero-mean normal distribution has maximum entropy. The first term in inequality (13) is replaced by KSG entropy estimation (9). According to [7] by considering distortion condition \( E[(Y - Z)^2] \leq \delta \), it can be said that: 
\[
E[(X - pX) X/\sigma_X^2]^2 \leq \sigma_X^2(1 - \rho^2) + \rho^2 \delta \sigma_X^2/\sigma_Z^2.
\]
Therefore, by considering the fact that the mutual information is non-negative, the lower bound (11) is derived. 

(b) Considering the convexity of the mutual information and the KSG estimation of the mutual information (10), the upper bound (12) can be derived. For the details of the proof, the readers are referred to [4].

Therefore, equations (11) and (12) together form the lower and upper bounds on the performance of the PPAN model for the non-Gaussian attributes with useful data as the observation. It should be noted that for unit variance attributes, at \( \delta = 1 \) the lower and upper bounds intersect. For the "full data" as observation, a similar upper bound as the one in equation (12) can be assumed. Therefore, for the rest of this study, the focus will be on the lower bound.

**Theorem 2. For the optimization problem A, given random variable X and Y both with unit standard deviation and the correlation coefficient \( \rho \), and assuming observed data \( W = (Y, X) \), we can say:**

\[
R(\delta) \geq \max \{0, R_{UVLB}(\delta)\}
\]  
(14)

where \( R_{UVLB}(\delta) \) (unit-variance lower bound) is as follows:

\[
R_{UVLB}(\delta) = \begin{cases} 
  c - 0.5 \log \left[ 1 - \left( \sqrt{\rho^2(1 - \delta)} - \sqrt{1 - \rho^2}\delta \right)^2 \right] & \delta \geq \rho^2 \\
  \delta - 0.5 \log(2\pi) & \delta < \rho^2 
\end{cases}
\]  
(15)

with \( c = H_{\text{null}}(X) - 0.5 \log(2\pi e) \).

**Proof.** Considering the mutual information \( I(X; Z) \), we have:

\[
I(X; Z) = H(X) - H(X|Z) = H(X) - H(X - \hat{E}[X|Z]|Z) \\
\quad \geq H(X) - H(X - \hat{E}[X|Z]) \\
\quad \geq H(X) - \mathcal{N}(0, E[(X - \hat{E}[X|Z])^2]) \\
\quad \geq H(X) - 0.5 \log \left( 2\pi e \left( E[X^2] - E^2[X|Z] \right) \right)
\]  
(16)

where (i) is due to this fact that conditioning reduces entropy and (ii) is true since, for a given value of the second moment, the zero-mean normal distribution has maximum entropy; while in (iii), following the procedure in [7], the linear minimum mean squared error estimation of X given Z as \( \hat{E}[X|Z] = E[X|Z]/E[Z]^2 \) and orthogonality rule in least squares estimation \( E[\hat{E}[X|Z]X - \hat{E}[X|Z]] = 0 \) is considered. To complete the proof, the minimum of term \( E^2[X|Z]/E[Z]^2 \) should be found by considering the distortion condition. To this end, for unit variance attributes, by considering random variables \( X, Y, Z \) as vectors \( x, y, z \) in vector space \( \ell_2 \) (and therefore replacing expectation operator on the product of two random variables with inner product in \( \ell_2 \) we have [7]:

\[
\min_{E[(Y - Z)^2] \leq \delta} \frac{E^2[X|Z]}{E[Z]^2} = \min_{\|y - z\| \leq \delta} \frac{|\langle x, z \rangle|}{\|z\|^2}
\]  
(17)

By associating \( i := x, j := \frac{1}{1 - \sqrt{\rho^2}}(y - \rho x) \), and \( \hat{k} := \frac{1 - \rho}{1 - \sqrt{\rho^2}}x \) as the unit vectors along orthogonal coordinate axes, and by considering \( t := z - y = t_1i + t_2j + t_3\hat{k} \), it can be said that \( x = \hat{i}, y = \rho i + \sqrt{1 - \rho^2}j \), and \( z = (t_1 + \rho)i + (t_2 + \sqrt{1 - \rho^2})j + t_3\hat{k} \). Therefore, by substituting \( x, y, z \) in the previous minimization problem, it changes as the following optimization problem:

\[
\min_{t_1^2 + t_2^2 + t_3^2 \leq 1} \left( \frac{(t_1 + \rho)^2}{t_1^2 + t_2^2 + t_3^2 + 2t_1\rho + 2t_3\sqrt{1 - \rho^2} + 1} \right)
\]  
(18)

It is shown in [7] that when \( \rho^2 \leq \delta \), problem (18) is minimum for \( t_1 = -\rho, t_2 = t_0 = 0 \) while for the case \( \rho^2 \geq \delta \) the minimum of (18) is attained for \( t_1 = -\rho, t_2 = \sqrt{\delta(1 - \delta)/(1 - \rho^2)}, t_0 = \sqrt{\delta - (t_0^2)^2}, t_0 = 0 \). By substituting the \( t_1^2, t_2^2, t_0^2 \) in (18) we have:

\[
\min_{E[(Y - Z)^2] \leq \delta} \frac{E^2[X|Z]}{E[Z]^2} = \begin{cases} 
  0 & \delta \geq \rho^2 \\
  \left( \sqrt{\rho^2(1 - \delta)} - \sqrt{1 - \rho^2}\delta \right) & \delta < \rho^2
\end{cases}
\]  
(19)

Therefore, by considering (17) and (19) and entropy estimation \( H_{\text{null}}(X) \) and \( E[X^2] = 1 \)(zero mean unit variance attributes) the lower bounds of \( R(\delta) \) for full data case with unit variance attributes \( R_{UVLB}(\delta) \) as in equation (15) can be derived. Considering this fact that mutual information is non-negative, equation (14) would be immediate.

For the non-unit variance attributes, generally, the lower bounds would be very complicated, however, for the case \( \sigma_X^2 = \sigma_Y^2 = \sigma^2 \), the following theorem would provide a lower bound.

**Theorem 3. For the optimization problem A, given random variable X and Y with standard deviation \( \sigma_X^2 = \sigma_Y^2 = \sigma^2 \) and the correlation coefficient \( \rho \), and assuming observed data \( W = (X, Y) \), we can say:**

\[
R(\delta) \geq \max \{0, R_{LB}(\delta)\}
\]  
(20)

where \( R_{LB}(\delta) \) (lower bound) is as follows:

\[
R_{LB}(\delta) = \begin{cases} 
  c - 0.5 \log(\sigma^2) & \delta \geq \sigma^2 \rho^2 \\
  c - 0.5 \log \left( \sigma^2 - \left( \sqrt{\rho^2(\sigma^2 - \delta)} - \sqrt{1 - \rho^2}\delta \right)^2 \right) & \delta < \sigma^2 \rho^2
\end{cases}
\]  
(21)

with \( c = H_{\text{null}}(X) - 0.5 \log(2\pi e) \).

**Proof.** The approach for finding the lower bound (21) is exactly similar to the procedure in Theorem 2 until arriving
at equation (17). Now, to solve equation (17) (for $\sigma^2 = \sigma^2_{\tilde{x}}$), by associating $\tilde{i} := x/\sigma$, $\tilde{j} := \frac{1}{\sigma \sqrt{1 - \rho^2}}(y - \rho x)$, and $\tilde{k} := \frac{\mathbf{z} - \mathbf{y}}{\mathbf{X} - \mathbf{Y}}$ as the unit vectors along orthogonal coordinate axes and by considering $\mathbf{t} := \mathbf{z} - \mathbf{y} = t_1 \tilde{i} + t_2 \tilde{j} + t_3 \tilde{k}$, we can say $\mathbf{x} = \sigma \tilde{i}$, $\mathbf{y} = \sigma \tilde{p} \tilde{i} + \sigma \sqrt{1 - \rho^2} \tilde{j}$, and $\mathbf{z} = (t_1 + \sigma \rho \tilde{i}) + (t_2 + \sigma \sqrt{1 - \rho^2}) \tilde{j} + t_3 \tilde{k}$. By substituting these $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ in (17), the minimization problem (17) changes as follows:

$$\min_{\mathbf{t}_1^2 + \mathbf{t}_2^2 + \mathbf{t}_3^2 \leq \delta} \left[ \frac{\sigma^2 (t_1 + \sigma \rho)^2}{t_1^2 + t_2^2 + t_3^2 + 2t_1 \sigma \rho + 2t_2 \sigma \sqrt{1 - \rho^2} + 1} \right]$$

(22)

By factoring $\sigma^2$ from numerator and denominator of the objective function in (22) and the both sides of the distortion constraint, and then using change of variables as $t'_1 = \frac{\mathbf{t}_1}{\sigma}$, $t'_2 = \frac{\mathbf{t}_2}{\sigma}$, $t'_3 = \frac{\mathbf{t}_3}{\sigma}$ and considering $\delta' = \delta/\sigma^2$ the minimization problem (22) changes as follows:

$$\min_{t'_1^2 + t'_2^2 + t'_3^2 \leq \delta'} \left[ \frac{\sigma^2 (t'_1 + \rho)^2}{t'_1^2 + t'_2^2 + t'_3^2 + 2t'_1 \rho + 2t'_2 \sqrt{1 - \rho^2} + 1} \right]$$

(23)

which is exactly similar to the minimization problem (18) except with a $\sigma^2$ factor at the numerator of the objective function. Therefore, similar to the (18), when $\rho^2 \leq \delta'$, problem (23) is minimum for $t'_1 = -\rho$, $t'_2 = t'_3 = 0$ while for the case when $\rho^2 > \delta'$ the minimum of (23) is found for $t'_1 = -\delta' \rho - \sqrt{\delta'(1 - \delta')(1 - \rho^2)}$, $t'_2 = \sqrt{\delta' - (t'_1)^2}$, $t'_3 = 0$. Thus, for the non-unit variance and when $\sigma^2_{\tilde{x}} = \sigma^2_{\tilde{y}} = \sigma^2$ we have:

$$\min_{E[YZ] \leq \delta} \mathbb{E}[XZ] = \begin{cases} 0 & \delta' \geq \rho^2 \\ \sigma^2 \left( \sqrt{\rho^2(1 - \delta')} - \sqrt{(1 - \rho^2)\delta'} \right)^2 & \delta' < \rho^2 \\ \text{null} & \delta' \leq \sigma^2 \rho^2 \\ \delta \geq \sigma^2 \rho^2 \\ \sigma^2 \left( \sqrt{\rho^2(1 - \delta')} - \sqrt{(1 - \rho^2)\delta'} \right)^2 & \delta' < \sigma^2 \rho^2 \\ \end{cases}$$

(24)

Therefore, by considering (16) and (24) and entropy estimation $H_{\text{null}}(X)$ and $\mathbb{E}[X^2] = \sigma^2_{\tilde{x}} = \sigma^2$, the lower bound (21) would be derived. Similarly, this fact that the mutual information is non-negative would complete the proof.

IV. EMPIRICAL EXAMPLES ON PERFORMANCE OF PPAN MODEL

In this section, the PPAN model is evaluated for several cases where useful data $Y$ and private data $X$ are continuous non-Gaussian. In all the examples, both the Mechanism and Adversary networks are deep neural network each include three hidden layers with 16 nodes and rectified linear unit (ReLU) as the activation function. For both networks, the RMSprop optimizer [19] with learning rate 0.01 is used. The data set includes 8000 training and 4000 test samples where 10% of the training is used as validation data for tuning hyperparameters including minibatch size $B$, penalty coefficient $\lambda$, number of steps applied for training the adversary $\ell$, and the width of seed noise $m_{\text{noise}}$. In the following examples, the seed noise $U$ is generated from independent and identically distributed (i.i.d.) samples according to a uniform distribution on the interval $[0, 1]$.

A. Continuous non-Gaussian Synthetic Data Set

For this part, three examples are provided where useful data $Y$ and private data $X$ come from Continuous non-Gaussian Synthetic Data. As the first example, assume useful data $Y$ and private data $X$ each come from a Gaussian mixture of three random variables with 4000 samples from each distribution as:

$$[X_1, Y_1] \sim \mathcal{N} \begin{bmatrix} 0 \\ 0 \end{bmatrix}, [X_2, Y_2] \sim \mathcal{N} \begin{bmatrix} 1 \\ 0.95 \end{bmatrix}, [X_3, Y_3] \sim \mathcal{N} \begin{bmatrix} -1 \\ 0.35 \end{bmatrix}.$$ 

This example is considered for two cases including observation based on the useful data, and observation based on the full data. As the second example, the private and useful attributes $X$ and $Y$ are multivariate Uniform on the interval $[0, 1]$ with covariance matrix

$$\begin{bmatrix} 1.2 & 0.90 \\ 0.90 & 1.2 \end{bmatrix}.$$ 

Finally, in the third example the private and useful attributes $X$ and $Y$ are multivariate Laplace distribution with covariance matrix

$$\begin{bmatrix} 1.3 & 0.95 \\ 0.95 & 1.3 \end{bmatrix}.$$ 

The last two examples are considered for the case where observation is based on the full data. The results of applying the PPAN model to these examples are presented in Fig. 2. It can be seen that the results of PPAN are within the determined bounds close to the lower bound.

B. Continuous non-Gaussian Real Data Set

In this part, the PPAN model is applied to a practical case where we aim at sharing the electricity consumption of several houses (five randomly selected houses in this example) in a...
certain region with a utility provider or third party and at the same time prevent any adversary to estimate the actual pattern of household consumption out of the shared data [20]. So the model distorts the actual consumption depending on the level of privacy. To this end, the Pecan Street data set which contains hourly electricity consumption of houses in Texas, Austin is used [21]. In this example, the useful data $Y$ and the private data $X$ both are the same as electricity consumption. The result of using the PPAN model, in this case, is presented in Fig. 3.

![Figure 3](image)

Fig. 3. Test results of privacy-distortion trade of for the Pecan Street data set where $(B, \lambda, \ell, m_{\text{noise}}) = (200, 10, 3, 5)$.

From the provided examples it can be concluded that similar to the Gaussian case, the PPAN model can be used for continuous non-Gaussian attributes as well, but with this difference that the assessment of the model is done using mutual information estimation compared with the provided lower and upper bounds.

V. CONCLUSION

In this study, the privacy-preserving adversarial network (PPAN) was examined for continuous non-Gaussian attributes. Although the PPAN model showed worked well for the discrete synthetic data and continuous Gaussian attributes, no method was provided for assessing this model for continuous non-Gaussian attributes. In this work, the performance of the PPAN model was assessed using the KSG estimation of entropy and mutual information. A lower bound and an upper bound of the privacy-distortion problem was determined using entropy and mutual information estimation and the result of the PPAN was compared with these bounds. Several examples based on synthetic data sets are provided. In addition, a practical case related to sharing power consumption of household (using an actual dataset) were examined. The results showed that the performance of the PPAN for these cases is within the determined bounds close to the lower bound. As future work, this study can be expanded to find tight bounds for a more general case including time series data.

ACKNOWLEDGMENT

This work was supported by Hydro-Quebec, the Natural Sciences and Engineering Research Council of Canada, and McGill University in the framework of the NSERC/Hydro-Quebec Industrial Research Chair in Interactive Information Infrastructure for the Power Grid (IRCPJ406021-14). We thank Dr. Francisco Messina and Dr. Pablo Piantanida for their useful discussions and comments.

REFERENCES

[1] D. Ye, S. Shen, T. Zhu, B. Liu, and W. Zhou, “One parameter defense—defending against data inference attacks via differential privacy,” IEEE Transactions on Information Forensics and Security, vol. 17, pp. 1466–1480, 2022.
[2] A. E. Ouadhiri and A. Abdellahi, “Differential privacy for deep and federated learning: A survey,” IEEE Access, vol. 10, pp. 22359–22380, 2022.
[3] Q. Chen, C. Xiang, M. Xue, B. Li, N. Borisov, D. Kaarfar, and H. Zhu, “Differentially private data generative models,” arXiv preprint arXiv:1812.02274, 2018.
[4] D. Rebollo-Monedero, J. Forne, and J. Domingo-Ferrer, “From t-closeness-like privacy to postrandomization via information theory,” IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 11, pp. 1623–1636, 2010.
[5] L. Sankar, S. R. Rajagopalan, and H. V. Poor, “Utility-privacy tradeoffs in databases: An information-theoretic approach,” IEEE Transactions on Information Forensics and Security, vol. 8, no. 6, pp. 838–852, 2013.
[6] C. Huang, P. Kairouz, X. Chen, L. Sankar, and R. Rajagopalan, “Generative adversarial privacy,” arXiv preprint arXiv:1807.05306, 2018.
[7] A. Tripathy, Y. Wang, and P. Ishwar, “Privacy-preserving adversarial networks,” in 2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton), pp. 495–505, IEEE, 2019.
[8] M. Shateri, F. Messina, P. Piantanida, and F. Labeau, “Deep directed information-based learning for privacy-preserving smart meter data release,” in 2019 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm), pp. 1–7, IEEE, 2019.
[9] M. Shateri, F. Messina, P. Piantanida, and F. Labeau, “Real-time privacy-preserving data release for smart meters,” IEEE Transactions on Smart Grid, vol. 11, no. 6, pp. 5174–5183, 2020.
[10] A. S. Khwaja, A. Anpalagan, M. Naeem, and B. Venkatesh, “Smart meter data obfuscation using correlated noise,” IEEE Internet of Things Journal, vol. 7, no. 8, pp. 7250–7264, 2020.
[11] M. Gough, S. Santos, T. Alskaif, M. Javadi, R. Castro, and J. P. Catalao, “Preserving privacy of smart meter data in a smart grid environment,” IEEE Transactions on Industrial Informatics, 2021.
[12] Y. Jin and L. Lai, “Optimal accuracy-privacy trade-off of inference as service,” IEEE Transactions on Signal Processing, 2022.
[13] F. Ye, H. Cho, and S. El Rouayheb, “Mechanisms for hiding sensitive genotypes with information-theoretic privacy,” IEEE Transactions on Information Theory, vol. 68, no. 6, pp. 4090–4105, 2022.
[14] M. Shateri, F. Messina, P. Piantanida, and F. Labeau, “On the impact of side information on smart meter privacy-preserving methods,” in 2020 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm), pp. 1–6, 2020.
[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Advances in neural information processing systems, pp. 2672–2680, 2014.
[16] A. Kraskov, H. Stögbauer, and P. Grassberger, “Estimating mutual information,” Physical Review E, vol. 69, no. 6, p. 066138, 2004.
[17] D. Barber and F. Agakov, “The im algorithm: a variational approach to information maximization,” in Proceedings of the 16th International Conference on Neural Information Processing Systems, pp. 201–208, MIT Press, 2003.
[18] M. Pinsker, Information and Information Stability of Random Variables and Processes. Holden-Day series in time series analysis, Holden-Day, 1964.
[19] T. Tieleman and G. Hinton, “Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude,” COURSERA: Neural networks for machine learning, vol. 4, no. 2, pp. 26–31, 2012.
[20] M. Shateri, F. Messina, P. Piantanida, and F. Labeau, “Privacy-cost management in smart meters with mutual information-based reinforcement learning,” IEEE Internet of Things Journal, pp. 1–1, 2021.
[21] Pecan Street Inc., “Dataset: the world’s largest energy data resource,” 2019. https://dataplot.cloud/.