Generative Adversarial Models for Learning Private and Fair Representations

Anonymous authors
Paper under double-blind review

Abstract

We present Generative Adversarial Privacy and Fairness (GAPF), a data-driven framework for learning private and fair representations. GAPF leverages recent advancements in adversarial learning to allow a data holder to learn "universal" representations that decouple a set of sensitive attributes from the rest of the dataset. Under GAPF, finding the optimal privacy mechanism is formulated as a constrained minimax game between a private/fair encoder and an adversary. We show that for appropriately chosen adversarial loss functions, GAPF provides privacy guarantees against strong information-theoretic adversaries and enforces demographic parity. We also evaluate the performance of GAPF on multi-dimensional Gaussian mixture models and real datasets, and show how a designer can certify that representations learned under an adversary with a fixed architecture perform well against more complex adversaries.

Keywords- Data Privacy, Fairness, Adversarial Learning, Generative Adversarial Networks, Minimax Games, Information Theory

1 Introduction

The use of deep learning algorithms for data analytics has recently seen unprecedented success for a variety of problems such as image classification, natural language processing, and prediction of consumer behavior, electricity use, political preferences, to name a few. The success of these algorithms hinges on the availability of large datasets, that often contain sensitive information, and thus, may facilitate learning models that inherit societal biases leading to unintended algorithmic discrimination on legally protected groups such as race or gender. This, in turn, has led to privacy and fairness concerns and a growing body of research focused on developing representations of the dataset with fairness and/or privacy guarantees. These techniques predominantly involve designing randomizing mechanisms, and in recent years, distinct approaches with provable *statistical privacy or fairness* guarantees have emerged.

On the privacy front, preserving the utility of published datasets while simultaneously providing provable privacy guarantees is a well-known challenge. On the one hand, context-free privacy solutions, such as differential privacy (Dwork et al., 2006b;a; Dwork, 2008; Dwork & Roth, 2014), provide strong worst-case privacy guarantees, but often lead to a significant reduction in utility. On the other hand, context-aware privacy solutions, such as mutual information privacy (Rebollo-Monedero et al., 2010; Calmon & Fawaz, 2012; Sankar et al., 2013; Salamatian et al., 2015; Basciftci et al., 2016), achieve an improved privacy-utility tradeoff, but assume that the data holder has access to dataset statistics.

On the fairness front, machine learning models seek to maximize predictive accuracy. Fairness concerns arise when models learned from datasets that include patterns of societal bias and discrimination inherit such biases. Thus, there is a need for actively decorrelating sensitive and non-sensitive data. In the context of publishing datasets or meaningful representations that can be "universally" used for a variety of learning tasks, modifying the training data is the most appropriate and the focus of this work. Fairness can then be achieved by carefully designing objective functions which approximate a specific fairness definition while simultaneously ensuring maximal utility (Zemel et al., 2013; Calmon et al., 2017; Ghassami et al., 2018). This, in turn, requires dataset statistics.

Recently, adversarial learning approaches for context-aware privacy and fairness have been studied extensively (Edwards & Storkey, 2015; Abadi & Andersen, 2016; Raval et al., 2017; Huang et al., 2017; Tripathy et al., 2017; Beutel et al., 2017; Madras et al., 2018; Zhang et al., 2018). They allow the data curator to cleverly add noise where it matters to decorrelate the sensitive attributes from the rest of the dataset . These approaches overcome the lack of statistical knowledge by taking a *data-driven approach* that leverages recent advancements in generative adversarial networks (GANs) (Goodfellow et al., 2014; Mirza & Osindero, 2014). However, most existing works conduct extensive empirical studies without theoretical verification with designs focused dominantly on a specific classification task.

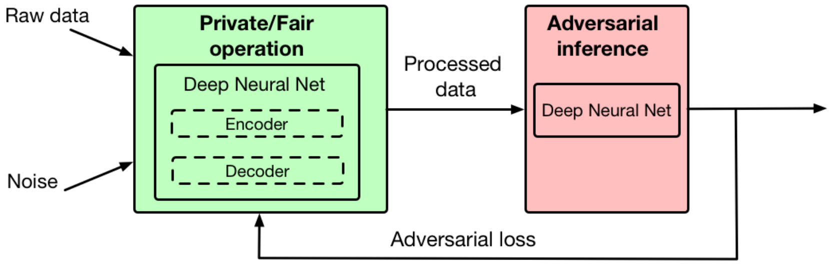


Figure 1: Generative Adversarial Model for Privacy and Fairness

This work introduces a general framework for context-aware privacy and fairness that we call *generative adversarial privacy and fairness* (GAPF) (see Figure 1). We provide precise connections to game-theoretic privacy and fairness formulations and derive game-theoretically optimal decorrelating mechanisms to compare against those learned directly from the data. We also explore how to create representations of datasets that can be used for a variety of learning tasks.

Our Contributions. We list our main contributions below.

1. We introduce GAPF as a minimax game-theoretic formulation (see Figure 1) to design decorrelating mechanisms matched to an adversarial model and ensure privacy and/or fairness.
2. We show that our framework captures a rich class of statistical and information-theoretic adversaries. This allows us to compare data-driven approaches directly against strong inferential adversaries (e.g., a maximum *a posteriori* probability (MAP) adversary with access to dataset statistics). This is in sharp contrast with recent works on data-driven privacy (surveyed in Section 5) where privacy guarantees are only offered in the context of computational adversaries. We also show that by carefully designing the loss functions in the GAPF framework, we can perturb the data to enforce demographic parity.
3. We make precise connections between data-driven privacy/fairness methods and the minimax game-theoretic GAPF formulation; this implies that when: (i) the neural networks used in the data-driven approach have sufficient capacity, (ii) the learning rate is sufficiently small, and (iii) the training data is sufficiently large, the learned decorrelating scheme converges to the game-theoretically optimal one.
4. To showcase the power of our data-driven framework, we investigate a multi-dimensional Gaussian mixture data model. We derive game-theoretically optimal decorrelating mechanisms and compare them with those that are directly learned in a data-driven fashion to show that the gap between theory and practice is negligible. Furthermore, we demonstrate the performance of GAPF on two meaningful, widely used datasets identified as the GENKI dataset (Whitehill & Movellan, 2012) and HAR (Anguita et al., 2013) dataset for specific public and sensitive features. We also use state-of-the-art entropy estimators to verify that GAPF effectively decorrelates the sensitive features from the data.
5. To validate that the representations learned under an adversary with a fixed architecture generalize against unseen (more complex) adversaries, we show that the mutual information between the learned representations and the sensitive attributes is small, implying that no adversary (regardless of their computational power) can reliably learn the sensitive labels (Cover & Thomas, 2012).

Outline. The remainder of our paper is organized as follows. We formally present our GAPF model in Section 2. In Section 3, we present results for Gaussian mixture dataset models. In Section 4, we showcase the performance of GAPF on the GENKI and HAR datasets. We review recent related works in Section 5 and conclude in Section 6. All proofs and algorithms are deferred to appendices in the accompanying supplementary materials.

2 Generative Adversarial Model for Privacy and Fairness

We consider a dataset $\mathcal{D}$ with n entries where each entry is denoted as (S, X, Y) where $S \in \mathcal{S}$ is the sensitive variable, $X \in \mathcal{X}$ is the public variable, and $Y \in \mathcal{Y}$ is the target (non-sensitive) variable (for learning). Instances of X, S, and Y are denoted by x, s and y, respectively. We assume that each entry pair (X, S, Y) is distributed according to $P(X, S, Y)$, and is independent from other entry pairs in the dataset.

Privacy and fairness. Context-aware notions of privacy models how well an adversary, with access to the public data X, can infer the sensitive features S from the data. Research on context-aware privacy focus on privacy that capture a range of adversarial capabilities ranging from a belief refining adversary using mutual information to quantify privacy to a guessing adversary using a hard-decision rule. On the other hand, recent results on fairness in learning applications guarantees that for a specific target variable Y, the prediction of a machine learning model is accurate with respect to (*w.r.t.*) Y but unbiased *w.r.t.* the sensitive variable S. The three oft-used fairness measures are demographic parity, equalized odds, and equal opportunity. Demographic parity imposes the strongest fairness requirement via complete independence of $\hat{Y}$ and S, and thus, least favors (for correlated Y and S) utility Hardt et al. (2016). Equalized odds ensures this independence conditioned on the label Y thereby ensuring equal rates for true and false positives (binary Y) for all demographics. Equal opportunity ensures equalized odds for the true positive case alone (Hardt et al., 2016).

When publishing a useful learning representation for multiple users with different learning tasks, it is not possible to identify a set of target variables *a priori*. Thus, our decorrelating mechanisms do not include Y. Formally, we define the decorrelating mechanism as a randomized mapping given by $\hat{X} = g(X)$. We note that $g(\cdot)$ can more generally depend on both X and S but for the sake of simplicity, we restrict our attention to mechanisms that only depend on X.

Let h be a decision rule used by the adversary to infer the sensitive variable S as $\hat{S} = h(g(X))$ from the representation $g(X)$. We allow for *hard decision rules* under which $h(g(X))$ is a direct estimate of S and *soft decision rules* under which $h(g(X)) = P_h(\cdot|g(X))$ is a distribution over $\mathcal{S}$. To quantify the adversary's performance, we use a loss function $\ell(h(g(X = x)), S = s)$ defined for every public-sensitive pair (x, s). Thus, the expected loss of the adversary *w.r.t.* X and S is $L(h, g) \triangleq \mathbb{E}[\ell(h(g(X)), S)]$, where the expectation is taken over $P(X, S)$ and the randomness in g and h.

Intuitively, the private/fair encoder would like to minimize the adversary's ability to learn S reliably from the published representation. This can be trivially done by releasing an $\hat{X}$ independent of X. However, such an approach provides no utility for data analysts who want to learn non-sensitive variables from $\hat{X}$. To overcome this issue, we capture the loss incurred by perturbing the original data via a distortion function $d(\hat{x}, x)$, which measures how far the original data $X = x$ is from the processed data $\hat{X} = \hat{x}$. Ensuring statistical utility in turn requires constraining the average distortion $\mathbb{E}[d(g(X), X)]$ where the expectation is taken over $P(X, S)$ and the randomness in g.

The data holder would like to find a decorrelating mechanism g that is both privacy/fairness preserving (in the sense that it is difficult for the adversary to learn S from $\hat{X}$) and utility preserving (in the sense that it does not distort the original data too much). In contrast, for a fixed choice of decorrelating mechanism g, the adversary would like to find a (potentially randomized) function h that minimizes its expected loss, which is equivalent to maximizing the negative of the expected loss. This leads to a constrained minimax game between the private/fair encoder and the adversary given by

$$\min_{g(\cdot)} \max_{h(\cdot)} \quad -L(h, g), \quad s.t. \quad \mathbb{E}[d(g(X), X)] \leq D, \tag{1}$$

where the constant $D \geq 0$ determines the allowable distortion for the encoder and the expectation is taken over $P(X, S)$ and the randomness in g and h.

Theorem 1. *Under the class of hard decision rules, i.e., $\ell(h(g(x), s))$ is the 0-1 loss function, the optimal adversarial strategy simplifies to using a maximum a posteriori (MAP) decision rule that maximizes $P(S|g(X))$. On the other hand, for a soft-decision decoding adversary, under log-loss function $\ell(h(g(X)), s) = -\log P_h(s|g(X))$, the optimal adversarial strategy h^* is the posterior belief of S given $g(X)$ and the GAPF minimax problem in equation 1 simplifies to $\min_{g(\cdot)} I(g(X); S)$ subject to $\mathbb{E}[d(g(X), X)] \leq D$, where $I(g(X); S)$ is the mutual information (MI) between $g(X)$ and S.*

Proposition 1. *Under the log-loss, GAPF enforces demographic parity subject to the distortion constraint.*

The proof of Theorem 1 and Proposition 1 are presented in Appendix A and B, respectively. Theorem 1 shows that GAPF can recover MI privacy (under a log-loss) and MAP privacy (under a 0-1 loss). Further, Proposition 1 shows that GAPF enforces demographic parity under the log-loss function. Many notions of fairness rely on computing probabilities to ensure independence of sensitive and target variables that are not easy to optimize in a data-driven fashion. We propose log-loss (modeled in practice via cross-entropy) in GAPF as a proxy for enforcing fairness. Next, we discuss how to use data-driven GAPF to learn the decorrelating mechanism directly from data.

Data-driven GAPF. Thus far, we have focused on a setting where the data holder has access to $P(X, S)$. When $P(X, S)$ is known, the data holder can simply solve the constrained minimax optimization problem in equation 1 (game-theoretic version of GAPF) to obtain a decorrelating mechanism that would perform best against a chosen type of adversary. In the absence of $P(X, S)$, we propose a data-driven version of GAPF that allows the data holder to learn decorrelating mechanisms directly from a dataset $\mathcal{D} = \{(x_{(i)}, s_{(i)})\}_{i=1}^{n}$. Under the data-driven version of GAPF, we represent the decorrelating mechanism via a generative model $g(X; \theta_p)$ parameterized by θ_p. This generative model takes X as input and outputs $\hat{X}$. In the training phase, the data holder learns the optimal parameters θ_p by competing against a *computational adversary*:

a classifier modeled by a neural network $h(g(X;\theta_p);\theta_a)$ parameterized by θ_a. In the evaluation phase, the performance of the learned decorrelating mechanism can be tested under a strong adversary that is computationally unbounded and has access to dataset statistics. We follow this procedure in the next section.

We note that in theory, the functions h and g can be arbitrary. However, in practice, we need to restrict them to a rich hypothesis class. Figure 1 shows an example of the GAPF model in which the private/fair encoder and adversary are modeled as multi-layer neural networks. For a fixed h and g, we can quantify the adversary's *empirical loss* using cross entropy $L_n(\theta_p,\theta_a) = -\frac{1}{n}\sum_{i=1}^{n} s_{(i)} \log h(g(x_{(i)};\theta_p);\theta_a) + (1 - s_{(i)})\log(1 - h(g(x_{(i)};\theta_p);\theta_a))$. The optimal parameters for encoder and adversary are the solutions to

$$\min_{\theta_p}\max_{\theta_a} \quad -L_n(\theta_p,\theta_a), \quad s.t. \quad \mathbb{E}_{\mathcal{D}}[d(g(X;\theta_p),X)] \le D, \tag{2}$$

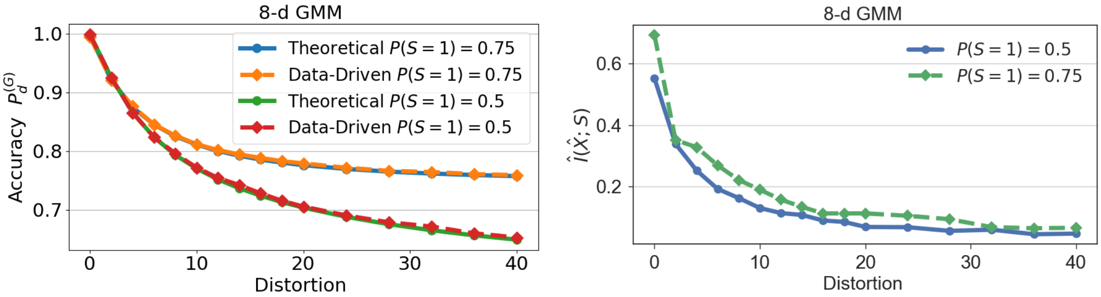
where the expectation is over $\mathcal{D}$ and the randomness in g.

The minimax optimization in equation 2 is a two-player non-cooperative game between the encoder and the adversary with strategies θ_p and θ_a, respectively. In practice, we can learn the equilibrium of the game using an iterative algorithm (see Algorithm 1 in Appendix C). We first maximize the negative of the adversary's loss function in the inner loop to compute the parameters of h for a fixed g. Then, we minimize the encoder's loss function, which is modeled as the negative of the adversary's loss function, to compute the parameters of g for a fixed h. To avoid over-fitting and ensure convergence, we alternate between training the adversary for j epochs and training the encoder for one epoch. This results in the adversary moving towards its optimal solution for small perturbations of the encoder (Goodfellow et al., 2014). Observe that the hard constraint in equation 2 makes our minimax problem different from what is extensively studied in the machine learning community. The algorithmic approach and optimization techniques that we use to solve the constrained optimization in equation 2 are detailed in the Appendix C.

3 GAPF FOR GAUSSIAN MIXTURE MODELS

In this section, we focus on a setting where $S \in \{0,1\}$ and X is an m-dimensional Gaussian mixture random vector whose mean is dependent on S. We derive game-theoretically optimal decorrelating mechanisms by considering a MAP adversary who has access to $P(X,S)$ and the mechanism. We compare the game-theoretically optimal mechanism against the data-driven decorrelating mechanism learned from a synthetic dataset by competing against a computational adversary (modeled by a multi-layer neural network). To quantify the performance of the learned decorrelating mechanism, we compute the accuracy of inferring S under a strong MAP adversary that has access to both the joint distribution of (X,S) and the decorrelating mechanism. Furthermore, we use state-of-the-art mutual information estimator (detailed in Appendix F) to demonstrate that GAPF effectively decorrelates the sensitive variables from the data. The details of the game-theoretically optimal and data-driven GAPF are included in Appendix D.

Figure 2 illustrates the performance of the learned GAPF mechanism against a strong theoretical MAP adversary for $P(S=1)=0.75$ and 0.5. It can be seen that the inference accuracy of the MAP adversary decreases as the distortion increases and asymptotically approaches (as expected) the prior on the sensitive variable. The decorrelating mechanism obtained via the data-driven approach performs very well when pitted against the MAP adversary (maximum accuracy difference around 0.3% compared to the theoretical approach). Furthermore, the estimated mutual information decreases as the distortion increases. In other words, for the data generated by Gaussian mixture model with binary sensitive variable, the data-driven version of GAPF can learn decorrelating mechanisms that perform as well as the mechanisms computed under the theoretical version of GAPF, given that the encoder has access to the statistics of the dataset.



(a) Sensitive variable classification accuracy

(b) Estimated mutual information between S and $\hat{X}$

Figure 2: Performance of learned GAPF for Gaussian mixture models

4 GAPF FOR REAL DATASETS

We apply our GAPF framework to real-world datasets to demonstrate its effectiveness. Specifically, we consider the following two datasets: the GENKI dataset comprised of greyscale face images and the UCI Human Activity Recognition (HAR) dataset containing time and frequency features of motion sensor data. The following experiments are implemented using TensorFlow (Abadi et al., 2016).

4.1 EXPERIMENT SETUP

GAPF aims to learn a mechanism that decorrelates sensitive variables from the the rest of the data with limited distortion. We train our model based on the data-driven GAPF presented in Section 2. To evaluate the performance of GAPF, we conduct the following three experiments. Firstly, we compare GAPF with random noise adding methods in terms of sensitive variable classification accuracy *vs.* distortion. Secondly, we train a target (non-sensitive) variable classifier on the processed data to show that GAPF preserves the utility of the dataset. Finally, we use state-of-the-art mutual information estimator (detailed in Appendix F) to demonstrate that GAPF effectively decorrelates the sensitive variables from the data.

The GENKI dataset consists of $1,740$ training and 200 test samples. Each data sample is a 16×16 greyscale face image with varying facial expressions. Each pixel value is normalized between 0 and 1. Both training and test datasets contain 50% male and 50% female. Among each gender, we have 50% smile and 50% non-smile faces. We consider gender as sensitive variable S and the image pixels as public variable X. The HAR dataset consists of 561 time and frequency domain features of motion sensor data collected by a smartphone from 30 subjects performing six activities (walking, walking upstairs, walking downstairs, sitting, standing, laying). Each feature is normalized between -1 and 1. We choose subject identity as sensitive variable S and the features of motion sensor data as public variable X. The dataset is randomly partitioned into $8,000$ training samples and $2,299$ test samples.

4.2 ENCODER AND ADVERSARY MODEL

For the GENKI dataset, we consider two different encoder architectures: the feedforward neural network encoder (FNNE) and the transposed convolution neural network encoder (TCNNE). The FNNE architecture uses a feedforward multi-layer neural network to combine the low-dimensional random noise and the original image together (Figure 8). The TCNNE takes a low-dimensional random noise and generates high-dimensional noise using a multi-layer transposed convolution neural network. The generated high-dimensional noise is added to each pixel of the original image to produce the processed image (Figure 8).

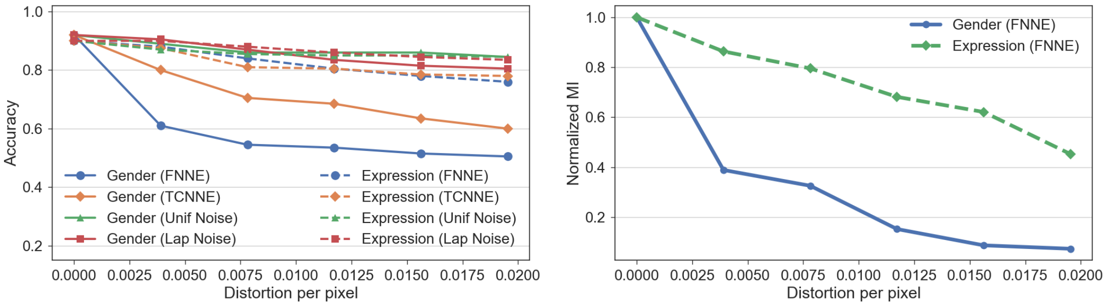
The FNNE is modeled by a four-layer feedforward neural network. We first reshape each image to a vector (256×1), and then concatenate it with a 100×1 Gaussian random noise vector. Each entry in the noise vector is sampled independently from a standard Gaussian distribution. We feed the entire vector to a four-layer fully connected (FC) neural network. Each layer has 256 neurons with a leaky ReLU activation function. Finally, we reshape the output of the last layer to a 16×16 image. To model the TCNNE, we first generate a 100×1 Gaussian random vector and use a linear projection to map the noise vector to a $4 \times 4 \times 256$ feature tensor. The feature tensor is then fed to an initial transposed convolution layer (DeCONV) with 128 filters (filter size 3×3, stride 2) and a ReLU activation, followed by another DeCONV layer with 1 filter (filter size 3×3, stride 2) and a tanh activation. The output of the DeCONV layer is added to the original image to generate the processed data. For both encoders, we add batch normalization (Ioffe & Szegedy, 2015) on each hidden layer to prevent covariance shift and help gradients to flow.

We model the adversary using state-of-the-art convolutional neural networks (CNNs). This architecture outperforms most of other models for image classification (Krizhevsky et al., 2012; Simonyan & Zisserman, 2014; Szegedy et al., 2015; He et al., 2016). In this experiment, the adversary is model by a seven-layer CNN (Figure 10 in Appendix). The processed images are fed to two convolution layers (CONV) whose sizes are $3 \times 3 \times 32$ and $3 \times 3 \times 64$, respectively. Each convolution layer is followed by ReLU activation and batch normalization. The output of each convolution layer is fed to a 2×2 maxpool layer (POOL) to extract features for classification. The second maxpool layer is followed by two fully connected layers, each contains 1024 neurons with a batch normalization and a ReLU activation. Finally, the output of the last fully connected layer is mapped to the output layer, which contains two neurons capturing the belief of the subject being a male or a female.

For the HAR dataset, we use the FNNE architecture modeled by a four-layer feedforward neural network. We first concatenate the original data with a 100×1 Gaussian random noise vector. We then feed the entire 661 vector to a Feed Forward neural network with three hidden fully connected (FC) layers. Each hidden layer has 512 neurons with a leaky ReLU activation. Finally, we use another FC layer with 561 neurons to generate the processed data. For the adversary, we use a five-layer feedforward neural network. The hidden layers have $512, 512, 256$, and 128 neurons with leaky ReLU activation, respectively. The output of the last hidden layer is mapped to the output layer, which contains 30 neurons capturing the belief of the subject's identity. For both encoder and adversary, we add a batch normalization after the output of each hidden layer.

4.3 Illustration of Results

The GENKI Dataset. Figure 3a illustrates the gender classification accuracy of the adversary for different values of distortion. It can be seen that the adversary's accuracy of classifying the sensitive variable (gender) decreases progressively as the distortion increases. Given the same distortion value, FNNE achieves lower gender classification accuracy compared to TCNNE: when the distortion is small (0.0039 per pixel), the adversary's classification accuracy is already about 80% and 62% by using the TCNNE and the FNNE architecture, respectively. When we increase the distortion to 0.0195, the classification accuracy further decreases to 60% and 50.5%. An intuitive explanation is that the FNNE uses both the noise vector and the original image to generate the processed image. However, the TCNNE generates the noise mask that is independent of the original image pixels and adds the noise mask to the original image in the final step. To demonstrate the effectiveness of the learned GAPF mechanisms, we compare the gender classification accuracy of the learned GAPF mechanisms with adding uniform or Laplace noise. Figure 3a shows that for the same distortion, the learned GAPF mechanisms achieve much lower gender classification accuracies than using uniform or Laplace noise. Furthermore, the estimated mutual information $\hat{I}(\hat{X};S)$ normalized by $\hat{I}(X;S)$ also decreases as the distortion increases (Figure 3b).



(a) Classification accuracy

(b) Estimated normalized mutual information

Figure 3: Classification accuracy and mutual information for different distortion values on GENKI dataset

To evaluate the influence of GAPF on other non-sensitive variable (Y) classification tasks, we train another CNN (see Figure 10) to perform facial expression classification on datasets processed by different decorrelating mechanisms. In Figure 3a, we observe that the facial expression classification accuracy decreases gradually as the distortion increases. However, even for a large distortion value (0.019 per pixel), the expression classification accuracy only decreases by 13% at most. We also observe that given the same distortion value, the FNNE and TCNNE achieve similar classification accuracy on facial expression. Furthermore, the estimated normalzied mutual information $\hat{I}(\hat{X};Y)/\hat{I}(X;Y)$ decreases much slower than $\hat{I}(\hat{X};S)/\hat{I}(X;S)$ as the distortion increases (Figure 3b).

Figure 3a also shows that when the distortion value is small, adding Laplace noise yields higher accuracy in both gender and expression classification than uniform noise. However, when the distortion value becomes large, the uniform noise yields higher accuracy in both gender and facial expression classification. This is due to the fact that for the same distortion (variance of the noise), the Laplace noise is very spiky when the distortion value is small. As a result, the noise added to the pixels are concentrated around a small region centered at 0. However, when the distortion value is large, the Laplace noise becomes more spread out. As a result, larger noise values are more likely to be added to the pixels and thus help reduce the classification accuracy. The processed images using FNNE is shown in Figure 4. We observe that the encoder changes mostly eyes, nose, mouth, beard, and hair.

The HAR Dataset. Figure 5a illustrates the classification accuracy of activity and identity for different values of distortion. The adversary's accuracy of classifying the sensitive variable (identity) decreases progressively as the distortion increases. When the distortion is small (2 per data sample), the adversary's classification accuracy is already around 27%. When we increase the distortion to 8, the classification accuracy further decreases to 3.8%. To study the effect of the decorrelating mechanism on the classification of non-sensitive variables, we train another adversary network to perform activity classification on the processed data through our approach. Figure 5a depicts that even for a large distortion value (8 per data sample), the activity classification accuracy only decreases by 7% at most. Furthermore, figure 5b shows that the estimated normalized mutual information also decreases as the distortion increases.

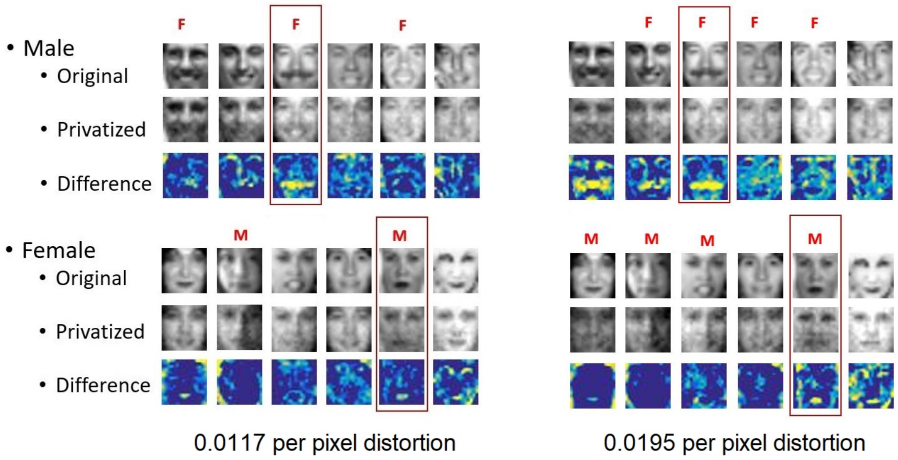
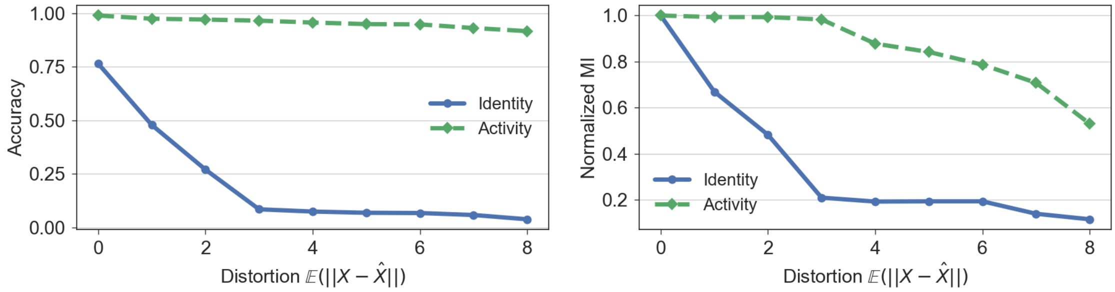


Figure 4: Perturbed images with different per pixel distortion using FNNE



(a) Classification accuracy

(b) Mutual information

Figure 5: Classification accuracy and mutual information for different distortion values on HAR dataset

5 Related Work

In the context of publishing datasets with privacy and utility guarantees, a number of similar approaches have been recently considered. We briefly review them and clarify how our work is different. DP-based obfuscators for data publishing have been considered in (Hamm, 2016; Liu et al., 2017). The author in (Hamm, 2016) considers a deterministic, compressive mapping of the input data with differentially private noise added either before or after the mapping. The mapping rule is determined by a data-driven methodology to design minimax filters that allow non-malicious entities to learn some public features from the filtered data, while preventing malicious entities from learning other sensitive features. The approach in (Liu et al., 2017) relies on using deep auto-encoders to determine the relevant feature space to add differentially private noise, thereby eliminating the need to add noise to the original data. These novel approaches leverage minimax filters and deep auto-encoders to incorporate a notion of context-aware privacy and achieve better privacy-utility tradeoffs while using DP to enforce privacy. However, DP can still incur a significant utility loss since it assumes worst-case dataset statistics. Our approach models a rich class of randomization-based mechanisms via a generative model that allows the privatizer to tailor the noise to the dataset.

Our work is closely related to adversarial neural cryptography (Abadi & Andersen, 2016), learning censored representations (Edwards & Storkey, 2015), privacy preserving image sharing (Raval et al., 2017), and privacy-preserving adversarial networks (Tripathy et al., 2017) in which adversarial learning is used to learn how to protect communications by encryption or hide/remove sensitive information. Similar to these problems, our model includes a minimax formulation and uses adversarial neural networks to learn privatization schemes. However, in (Edwards & Storkey, 2015; Raval et al., 2017), the authors use non-generative auto-encoders to remove sensitive information. Instead, we use a GANs-like approach to learn privatization schemes that prevent an adversary from inferring the sensitive variable. Furthermore, we propose using mutual information as a criterion to certify that the representations we learned adversarially against an attacker with a fixed architecture generalize against unseen attackers with (possibly) more complex architecture.

Fair representations using information-theoretic objective functions and constrained optimization have been proposed in (Calmon et al., 2017; Ghassami et al., 2018). However, both approaches require the knowledge of dataset statistics, which are very difficult to obtain for real datasets. We overcome the issue of statistical knowledge by taking a *data-driven approach*, i.e., learning the representation from the data directly via adversarial models. In contrast to in-processing approaches that modify learning algorithms to ensure fair predictions (e..g, using linear programs in (Dwork et al., 2012; Fish et al., 2016) or via adversarial learning approach in (Zhang et al., 2018)), we focus on a pre-processing approach to ensure fairness for a variety of learning tasks. Our work is closely related to learning fair representation adversarially (Madras et al., 2018) in which adversarial learning is used to learn a fair representation of the original data by explicitly modeling the target label and the sensitive attributes. However, for publishing a dataset or a useful representation, it is difficult to identify a set of target variables *a priori* since the representation of the dataset may be used by different users with different learning tasks. Thus, we use a distortion constraint to enforce the utility of the published dataset and impose demographic parity to enable a variety of machine learning tasks. We also go beyond in formulating a game-theoretic setting with constrained optimization, which provides a specific privacy/fairness guarantee for a fixed distortion. Finally, we also compare the performance of the privatization schemes learned in an adversarial fashion with the game-theoretically optimal ones for canonical synthetic data models thereby providing formal verification of fair/private mechanisms that are learned by competing against computational adversaries.

We use conditional generative models to represent the decorrelating schemes. Generative models have recently received a lot of attention in the machine learning community (Goodfellow et al., 2014; Mirza & Osindero, 2014). Ultimately, deep generative models hold the promise of discovering and efficiently internalizing the statistics of the target signal to be generated. State-of-the-art generative models are trained in an adversarial fashion: the generated signal is fed into a discriminator which attempts to distinguish whether the data is real (i.e., sampled from the true underlying distribution) or synthetic (i.e., generated from a low dimensional noise sequence). Training generative models in an adversarial fashion has proven to be successful in computer vision and enabled several exciting applications. Analogous to how the generator is trained in GANs, we train the encoder in an adversarial fashion by making it compete with an attacker.

6 Conclusion

We have introduced a novel generative adversarial privacy/fairness framework for designing data-driven context-aware privacy or fairness mechanisms for publishing a dataset or a useful learning representation with verifiable guarantees. GAPF allows the data holder to learn the decorrelating mechanism directly from the dataset (to be published) without requiring access to the dataset statistics. Under GAPF, finding the optimal privacy mechanism is formulated as a game between two players: a private/fair encoder and an adversary. We have shown that for appropriately chosen loss functions, GAPF can provide guarantees against strong information-theoretic adversaries, such as a guessing MAP and belief-refining MI adversaries. It can also enforce demographic parity by using a log-loss function. We have also validated the performance of GAPF on synthetic Gaussian mixture models and the real datasets.

There are several fundamental questions that we seek to address. An immediate one is to develop techniques to rigorously benchmark data-driven results for large datasets against computable theoretical guarantees. More broadly, it will be interesting to investigate the robustness and convergence speed of the decorrelating mechanisms learned in a data-driven fashion. In this paper, we connect our objective function in GAPF with demographic parity. Since there is no single metric for fairness, this leaves room for designing objective functions that link to other fairness metrics such as equalized odds and equal opportunity. Finally, it will be also interesting to compare our approach to a context-free notion of privacy such as DP.

References

Martín Abadi and David G Andersen. Learning to protect communications with adversarial neural cryptography. *arXiv preprint arXiv:1610.06918*, 2016.

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge Luis Reyes-Ortiz. A public domain dataset for human activity recognition using smartphones. In *ESANN*, 2013.

Y. O. Basciftci, Y. Wang, and P. Ishwar. On privacy-utility tradeoffs for constrained data release mechanisms. In *2016 Information Theory and Applications Workshop (ITA)*, pp. 1–6, Jan 2016. doi: 10.1109/ITA.2016.7888175.

Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*, 2017.

F. P. Calmon and N. Fawaz. Privacy against statistical inference. In *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*, pp. 1401–1408, 2012.

Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pp. 3992–4001, 2017.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

Cynthia Dwork. Differential privacy: A survey of results. In *International Conference on Theory and Applications of Models of Computation*, pp. 1–19. Springer, 2008.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, August 2014. ISSN 1551-305X. doi: 10.1561/0400000042. URL `http://dx.doi.org/10.1561/0400000042`.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 486–503. Springer, 2006a.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006b.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226. ACM, 2012.

Jonathan Eckstein and W Yao. Augmented lagrangian and alternating direction methods for convex optimization: A tutorial and some illustrative computational results. *RUTCOR Research Reports*, 32, 2012.

Harrison Edwards and Amos Storkey. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*, 2015.

Benjamin Fish, Jeremy Kun, and Ádám D Lelkes. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pp. 144–152. SIAM, 2016.

Robert G Gallager. *Stochastic processes: theory for applications*. Cambridge University Press, 2013.

AmirEmad Ghassami, Sajad Khodadadian, and Negar Kiyavash. Fairness in supervised learning: An information theoretic approach. *arXiv preprint arXiv:1801.04378*, 2018.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Jihun Hamm. Minimax filter: Learning to preserve privacy from inference attacks. *arXiv preprint arXiv:1610.03577*, 2016.

Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pp. 3315–3323, 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Chong Huang, Peter Kairouz, Xiao Chen, Lalitha Sankar, and Ram Rajagopal. Context-aware generative adversarial privacy. *Entropy*, 19(12):656, 2017.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical review E*, 69(6):066138, 2004.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Jiachun Liao, Oliver Kosut, Lalitha Sankar, and Flavio P Calmon. Privacy under hard distortion constraints. *arXiv preprint arXiv:1806.00063*, 2018.

Walter E Lillo, Mei Heng Loh, Stefen Hui, and Stanislaw H Zak. On solving constrained optimization problems with neural networks: A penalty method approach. *IEEE Transactions on neural networks*, 4 (6):931–940, 1993.

Changchang Liu, Supriyo Chakraborty, and Prateek Mittal. Deeprotect: Enabling inference-based access control on mobile sensing applications. *arXiv preprint arXiv:1702.06159*, 2017.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309*, 2018.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Tan Nguyen and Scott Sanner. Algorithms for direct 0–1 loss optimization in binary classification. In *International Conference on Machine Learning*, pp. 1085–1093, 2013.

Nisarg Raval, Ashwin Machanavajjhala, and Landon P Cox. Protecting visual secrets using adversarial nets. In *CVPR Workshop Proceedings*, 2017.

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*, 22(11):1623–1636, November 2010. ISSN 1041-4347. doi: 10.1109/TKDE.2009.190.

S. Salamatian, A. Zhang, F. P. Calmon, S. Bhamidipati, N. Fawaz, B. Kveton, P. Oliveira, and N. Taft. Managing your private and public data: Bringing down inference attacks against your privacy. 9(7): 1240–1255, 2015. doi: 10.1109/JSTSP.2015.2442227. URL `http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7118663`.

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*, 8(6):838–852, 2013.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, et al. Going deeper with convolutions. Cvpr, 2015.

Ardhendu Tripathy, Ye Wang, and Prakash Ishwar. Privacy-preserving adversarial networks. *arXiv preprint arXiv:1712.07008*, 2017.

Jacob Whitehill and Javier Movellan. Discriminately decreasing discriminability with learned image filters. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 2488–2495. IEEE, 2012.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pp. 325–333, 2013.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. *arXiv preprint arXiv:1801.07593*, 2018.

Supplementary Material

A Proof of Theorem 1

Our minimax formulation places no restrictions on the adversary. Indeed, different loss functions and decision rules lead to different adversarial models. In what follows, we will discuss a variety of loss functions under hard and soft decision rules, and show how our GAPF framework can recover several popular information theoretic privacy notions.

Hard Decision Rules. When the adversary adopts a hard decision rule, $h(g(X))$ is an estimate of S. Under this setting, we can choose $\ell(h(g(X)), S)$ in a variety of ways. For instance, if S is continuous, the adversary can attempt to minimize the difference between the estimated and true sensitive variable values. This can be achieved by considering a squared loss function

$$\ell(h(g(X)), S) = (h(g(X)) - S)^2, \tag{3}$$

which is known as the ℓ_2 loss. In this case, one can verify that the adversary's optimal decision rule is $h^* = \mathbb{E}[S|g(X)]$, which is the conditional mean of S given $g(X)$. Furthermore, under the adversary's optimal decision rule, the minimax problem in equation 1 simplifies to

$$\min_{g(\cdot)} -\text{mmse}(S|g(X)) = -\max_{g(\cdot)} \text{mmse}(S|g(X)),$$

subject to the distortion constraint. Here $\text{mmse}(S|g(X))$ is the resulting minimum mean square error (MMSE) under $h^* = \mathbb{E}[S|g(X)]$. Thus, under the ℓ_2 loss, GAPF provides privacy guarantees against an MMSE adversary. On the other hand, when S is discrete (e.g., age, gender, political affiliation, etc), the adversary can attempt to maximize its classification accuracy. This is achieved by considering a 0-1 loss function (Nguyen & Sanner, 2013) given by

$$\ell(h(g(X)), S) = \begin{cases} 0 & \text{if } h(g(X)) = S \\ 1 & \text{otherwise} \end{cases}. \tag{4}$$

In this case, one can verify that the adversary's optimal decision rule is the *maximum a posteriori probability* (MAP) decision rule: $h^* = \arg\max_{s\in\mathcal{S}} P(s|g(X))$, with ties broken uniformly at random. Moreover, under the MAP decision rule, the minimax problem in equation 1 reduces to

$$\min_{g(\cdot)} -(1 - \max_{s\in\mathcal{S}} P(s, g(X))) = \min_{g(\cdot)} \max_{s\in\mathcal{S}} P(s, g(X)) - 1, \tag{5}$$

subject to the distortion constraint. Thus, under a 0-1 loss function, the GAPF formulation provides privacy guarantees against a MAP adversary.

Soft Decision Rules. Instead of a *hard decision* rule, we can also consider a broader class of *soft decision* rules where $h(g(X))$ is a distribution over $\mathcal{S}$; i.e., $h(g(X)) = P_h(s|g(X))$ for $s \in \mathcal{S}$. In this context, we can analyze the performance under a log-loss

$$\ell(h(g(X)), s) = \log \frac{1}{P_h(s|g(X))}. \tag{6}$$

In this case, the objective of the adversary simplifies to

$$\max_{h(\cdot)} -\mathbb{E}[\log \frac{1}{P_h(s|g(X))}] = -H(S|g(X)),$$

and that the maximization is attained at $P_h^*(s|g(X)) = P(s|g(X))$. Therefore, the optimal adversarial decision rule is determined by the true conditional distribution $P(s|g(X))$, which we assume is known to the data holder in the game-theoretic setting. Thus, under the log-loss function, the minimax optimization problem in equation 1 reduces to

$$\min_{g(\cdot)} -H(S|g(X)) = \min_{g(\cdot)} I(g(X); S) - H(S),$$

subject to the distortion constraint. Thus, under the log-loss in equation 6, GAPF is equivalent to using MI as the privacy metric (Calmon & Fawaz, 2012).

The 0-1 loss captures a strong guessing adversary; in contrast, log-loss or information-loss models a belief refining adversary. Next, we consider a more general α-loss function (Liao et al., 2018) that allows continuous interpolation between these extremes via

$$\ell(h(g(X)), s) = \frac{\alpha}{\alpha - 1} \left(1 - P_h(s|g(X))^{1-\frac{1}{\alpha}}\right), \tag{7}$$

for any $\alpha > 1$. It is easy to see that for large α ($\alpha \to \infty$), this loss approaches that of the 0-1 (MAP) adversary. As α decreases, the convexity of the loss function encourages the estimator $\hat{S}$ to be probabilistic, as it increasingly rewards correct inferences of lesser and lesser likely outcomes (in contrast to a hard decision rule by a MAP adversary of the most likely outcome) conditioned on the revealed data. As $\alpha \to 1$, equation 7 yields the logarithmic loss, and the optimal belief $P_{\hat{S}}$ is simply the posterior belief. Denoting $H_{\alpha}^{\mathrm{a}}(S|g(X))$ as the Arimoto conditional entropy of order α, one can verify that

$$\max_{h(\cdot)} -\mathbb{E}\left[\frac{\alpha}{\alpha-1}\left(1-P_h(s|g(X))^{1-\frac{1}{\alpha}}\right)\right] = -H_{\alpha}^{\mathrm{a}}(S|g(X)),$$

which is achieved by a 'α-tilted' conditional distribution

$$P_h^*(s|g(X)) = \frac{P(s|g(X))^{\alpha}}{\sum\limits_{s\in\mathcal{S}} P(s|g(X))^{\alpha}}.$$

Under this choice of a decision rule, the objective of the minimax optimization in equation 1 reduces to

$$\min_{g(\cdot)} -H_{\alpha}^{\mathrm{a}}(S|g(X)) = \min_{g(\cdot)} I_{\alpha}^{\mathrm{a}}(g(X);S) - H_{\alpha}(S), \tag{8}$$

where I_{α}^{a} is the Arimoto mutual information and H_{α} is the Rényi entropy. Note that as $\alpha \to 1$, we recover the classical MI privacy setting and when $\alpha \to \infty$, we recover the 0-1 loss.

B Proof of Proposition 1

Let's consider an arbitrary target variable Y which a user is interested in learning from the data. The objective of the learning task is to train a good model that takes $\hat{X}$ to predict Y. Thus, we have the Markov chain: $S \to X \to \hat{X} \to \hat{Y}$, where $\hat{Y}$ is an estimate of Y from the trained machine learning model. According to data processing inequality, we have $I(S;\hat{X}) \geq I(S;\hat{Y})$. By Theorem 1, for the log-loss function, the objective of GAPF is equivalent to minimizing $I(S;\hat{X})$, which is an upperbound on $I(S;\hat{Y})$. Notice that demographic parity requires S and $\hat{Y}$ to be independent, which is equivalent to $I(S;\hat{Y}) = 0$. Since mutual information is non-negative, GAPF enforces demographic parity by minimizing an upperbound of $I(S;\hat{Y})$ subject to the distortion constraint under the log-loss function.

C Alternate Minimax Algorithm

In this section, we present the alternate minimax algorithm to learn the GAPF mechanism from a dataset.

To incorporate the distortion constraint into the learning algorithm, we use the *penalty method* (Lillo et al., 1993) and *augmented Lagrangian method* (Eckstein & Yao, 2012) to replace the constrained optimization problem by a series of unconstrained problems whose solutions asymptotically converge to the solution of the constrained problem. Under the penalty method, the unconstrained optimization problem is formed by adding a penalty to the objective function. The added penalty consists of a penalty parameter ρ_t multiplied by a measure of violation of the constraint. The measure of violation is non-zero when the constraint is violated and is zero if the constraint is not violated. Therefore, in Algorithm 1, the constrained optimization problem of the encoder can be approximated by a series of unconstrained optimization problems with the loss function

$$\ell(\theta_p^t, \theta_a^{t+1}) = -\frac{1}{M}\sum_{i=1}^{M} \ell(h(g(x_{(i)};\theta_p^t);\theta_a^{t+1}), s_{(i)}) + \rho_t \max\{0, \frac{1}{M}\sum_{i=1}^{M} d(g(x_{(i)};\theta_p^t), x_{(i)}) - D\}, \tag{9}$$

where ρ_t is a penalty coefficient which increases with the number of iterations t. For convex optimization problems, the solution to the series of unconstrained problems will eventually converge to the solution of the original constrained problem (Lillo et al., 1993).

The augmented Lagrangian method is another approach to enforce equality constraints by penalizing the objective function whenever the constraints are not satisfied. Different from the penalty method, the augmented Lagrangian method combines the use of a Lagrange multiplier and a quadratic penalty term. Note that this method is designed for equality constraints. Therefore, we introduce a slack variable δ to convert the inequality distortion constraint into an equality constraint. Using the augmented Lagrangian method, the constrained optimization problem of the encoder can be replaced by a series of unconstrained problems

Algorithm 1 Alternating minimax privacy preserving algorithm

```
*Input:* dataset $\mathcal{D}$, distortion parameter $D$, iteration number $T$

*Output:* Optimal encoder parameter $\theta_p$

**procedure** ALERNATE MINIMAX($\mathcal{D}, D, T$)

    Initialize $\theta_p^1$ and $\theta_a^1$

    **for** $t = 1, ..., T$ **do**

        Random minibatch of $M$ datapoints $\{x_{(1)}, ..., x_{(M)}\}$ drawn from full dataset

        Generate $\{\hat{x}_{(1)}, ..., \hat{x}_{(M)}\}$ via $\hat{x}_{(i)} = g(x_{(i)}, s_{(i)}; \theta_p^t)$

        Update the adversary parameter $\theta_a^{t+1}$ by stochastic gradient ascend for $j$ epochs

            $\theta_a^{t+1} = \theta_a^t + \alpha_t \nabla_{\theta_a^t} \frac{1}{M} \sum_{i=1}^{M} -\ell(h(\hat{x}_{(i)}; \theta_a^t), s_{(i)}), \quad \alpha_t > 0$

        Compute the descent direction $\nabla_{\theta_p^t} l(\theta_p^t, \theta_a^{t+1})$, where

            $\ell(\theta_p^t, \theta_a^{t+1}) = -\frac{1}{M} \sum_{i=1}^{M} \ell(h(g(x_{(i)}, s_{(i)}; \theta_p^t); \theta_a^{t+1}), s_{(i)})$

        subject to $\frac{1}{M} \sum_{i=1}^{M} [d(g(x_{(i)}, s_{(i)}; \theta_p^t), x_{(i)})] \leq D$

        Perform line search along $\nabla_{\theta_p^t} l(\theta_p^t, \theta_a^{t+1})$ and update

            $\theta_p^{t+1} = \theta_p^t - \alpha_t \nabla_{\theta_p^t} \ell(\theta_p^t, \theta_a^{t+1})$

        Exit if solution converged

    **return** $\theta_p^{t+1}$
```

with the loss function given by

$$\begin{aligned}\ell(\theta_p^t, \theta_a^{t+1}, \delta) = &-\frac{1}{M} \sum_{i=1}^{M} \ell(h(g(x_{(i)}; \theta_p^t); \theta_a^{t+1}), s_{(i)}) + \frac{\rho_t}{2} (\frac{1}{M} \sum_{i=1}^{M} d(g(x_{(i)}; \theta_p^t), x_{(i)}) + \delta - D)^2 \\ &- \lambda_t (\frac{1}{M} \sum_{i=1}^{M} d(g(x_{(i)}; \theta_p^t), x_{(i)}) + \delta - D),\end{aligned} \tag{10}$$

where ρ_t is a penalty coefficient which increases with the number of iterations t and λ_t is updated according to the rule $\lambda_{t+1} = \lambda_t - \rho_t (\frac{1}{M} \sum\limits_{i=1}^{M} d(g(x_{(i)}; \theta_p^t), x_{(i)}) + \delta - D)$. For convex optimization problems, the solution to the series of unconstrained problems formulated by the augmented Lagrangian method also converges to the solution of the original constrained problem (Eckstein & Yao, 2012).

D GAPF FOR GAUSSIAN MIXTURE MODELS

In this section, we focus on a setting where $S \in \{0, 1\}$ and X is an m-dimensional Gaussian mixture random vector whose mean is dependent on S. Let $P(S = 1) = q$. Let $X|S = 0 \sim \mathcal{N}(-\mu, \Sigma)$ and $X|S = 1 \sim \mathcal{N}(\mu, \Sigma)$, where $\mu = (\mu_1, ..., \mu_m)$, and without loss of generality, we assume that $X|S = 0$ and $X|S = 1$ have the same covariance Σ.

We consider a MAP adversary who has access to $P(X, S)$ and the privacy mechanism. The encoder's goal is to privatize X in a way that minimizes the adversary's probability of correctly inferring S from

$\hat{X}$. In order to have a tractable model for the encoder, we mainly focus on linear (precisely affine) GAPF mechanisms $\hat{X} = g(X) = X + Z + \beta$, where Z is an independently generated noise vector. This linear GAPF mechanism enables controlling both the mean and covariance of the privatized data. To quantify utility of the privatized data, we use the ℓ_2 distance between X and $\hat{X}$ as a distortion measure to obtain a distortion constraint $\mathbb{E}_{X,\hat{X}} \|X - \hat{X}\|^2 \leq D$.

D.1 Game-Theoretical Approach

Consider the setup where both the encoder and the adversary have access to $P(X, S)$. Further, let Z be a zero-mean multi-dimensional Gaussian random vector. Although other distributions can be considered, we choose additive Gaussian noise for tractability reasons.

Without loss of generality, we assume that $\beta = (\beta_1, ..., \beta_m)$ is a constant parameter vector and $Z \sim \mathcal{N}(0, \Sigma_p)$. Following similar analysis in (Gallager, 2013), we can show that the adversary's probability of detection is given by

$$P_d^{(G)} = qQ\left(-\frac{\alpha}{2} + \frac{1}{\alpha}\ln\left(\frac{1-q}{q}\right)\right) + (1-q)Q\left(-\frac{\alpha}{2} - \frac{1}{\alpha}\ln\left(\frac{1-q}{q}\right)\right), \tag{11}$$

where $\alpha = \sqrt{(2\mu)^T(\Sigma + \Sigma_p)^{-1}2\mu}$. Furthermore, since $\mathbb{E}_{X,\hat{X}}[d(\hat{X}, X)] = \mathbb{E}_{X,\hat{X}}\|X - \hat{X}\|^2 = \mathbb{E}\|Z + \beta\|^2 = \|\beta\|^2 + tr(\Sigma_p)$, the distortion constraint implies that $\|\beta\|^2 + tr(\Sigma_p) \leq D$. To make the problem more tractable, we assume both X and Z are independent multi-dimensional Gaussian random vectors with diagonal covariance matrices.

Theorem 2. *Consider GAPF mechanisms given by $g(X) = X + Z + \beta$, where X and Z are multi-dimensional Gaussian random vectors with diagonal covariance matrices Σ and Σ_p. Let $\{\sigma_1^2, ..., \sigma_m^2\}$ and $\{\sigma_{p_1}^2, ..., \sigma_{p_m}^2\}$ be the diagonal entries of Σ and Σ_p, respectively. The parameters of the minimax optimal privacy mechanism are*

$$\beta_i{}^* = 0, \quad \sigma_{p_i}^*{}^2 = \left(\frac{|\mu_i|}{\sqrt{\lambda_0^*}} - \sigma_i^2, 0\right)^+, \forall i = \{1, 2, ..., m\},$$

where λ_0^* is chosen such that $\sum_{i=1}^m \left(\frac{|\mu_i|}{\sqrt{\lambda_0^*}} - \sigma_i^2\right)^+ = D$. For this optimal mechanism, the accuracy of the MAP adversary is given by equation 11 with $\alpha = 2\sqrt{\sum_{i=1}^m \frac{\mu_i^2}{\sigma_i^2 + \left(\frac{|\mu_i|}{\sqrt{\lambda_0^*}} - \sigma_i^2\right)^+}}$.

Proof. Let us consider $\hat{X} = X + Z + \beta$, where $\beta \in \mathbb{R}$ and Σ_p is a diagonal covariance whose diagonal entries is given by $\{\sigma_{p_1}^2, ..., \sigma_{p_m}^2\}$. Given the MAP adversary's optimal inference accuracy in equation 11, the objective of the encoder is to

$$\begin{aligned} \min_{\beta, \Sigma_p} \quad & P_{\text{d}}^{(\text{G})} \\ s.t. \quad & \|\beta\|^2 + tr(\Sigma_p) \leq D. \end{aligned} \tag{12}$$

Define $\frac{1-q}{q} = \eta$. The gradient of $P_{\text{d}}^{(\text{G})}$ *w.r.t.* α is given by

$$\frac{\partial P_{\text{d}}^{(\text{G})}}{\partial \alpha} = \tilde{p}\left(-\frac{1}{\sqrt{2\pi}} e^{-\frac{\left(-\frac{\alpha}{2} + \frac{1}{\alpha}\ln\eta\right)^2}{2}}\right)\left(-\frac{1}{2} - \frac{1}{\alpha^2}\ln\eta\right) \tag{13}$$

$$+ (1-\tilde{p})\left(-\frac{1}{\sqrt{2\pi}} e^{-\frac{\left(-\frac{\alpha}{2} - \frac{1}{\alpha}\ln\eta\right)^2}{2}}\right)\left(-\frac{1}{2} + \frac{1}{\alpha^2}\ln\eta\right)$$

$$= \frac{1}{2\sqrt{2\pi}}\left(\tilde{p}e^{-\frac{\left(-\frac{\alpha}{2} + \frac{1}{\alpha}\ln\eta\right)^2}{2}} + (1-\tilde{p})e^{-\frac{\left(-\frac{\alpha}{2} - \frac{1}{\alpha}\ln\eta\right)^2}{2}}\right) \tag{14}$$

$$+ \frac{\ln\eta}{\alpha^2\sqrt{2\pi}}\left(\tilde{p}e^{-\frac{\left(-\frac{\alpha}{2} + \frac{1}{\alpha}\ln\eta\right)^2}{2}} - (1-\tilde{p})e^{-\frac{\left(-\frac{\alpha}{2} - \frac{1}{\alpha}\ln\eta\right)^2}{2}}\right).$$

Note that

$$\frac{\tilde{p}e^{-\frac{\left(-\frac{\alpha}{2} + \frac{1}{\alpha}\ln\eta\right)^2}{2}}}{(1-\tilde{p})e^{-\frac{\left(-\frac{\alpha}{2} - \frac{1}{\alpha}\ln\eta\right)^2}{2}}} = \frac{\tilde{p}}{1-\tilde{p}} e^{\frac{\left(-\frac{\alpha}{2} - \frac{1}{\alpha}\ln\eta\right)^2 - \left(-\frac{\alpha}{2} + \frac{1}{\alpha}\ln\eta\right)^2}{2}} = \frac{\tilde{p}}{1-\tilde{p}} e^{\frac{2\ln\eta}{2}} = \frac{\tilde{p}}{1-\tilde{p}} e^{\ln\eta} = 1. \tag{15}$$

Therefore, the second term in equation 14 is 0. Furthermore, the first term in equation 14 is always positive. Thus, $P_{\mathrm{d}}^{(\mathrm{G})}$ is monotonically increasing in α. As a result, the optimization problem in equation 12 is equivalent to

$$\begin{aligned} \min_{\beta,\Sigma_p} \quad & (2\mu)^T(\Sigma+\Sigma_p)^{-1}2\mu \\ s.t. \quad & \|\beta\|^2 + tr(\Sigma_p) \le D. \end{aligned} \tag{16}$$

The objective function in equation 16 can be written as

$$2\begin{bmatrix}\mu_1 & \mu_2 & \cdots & \mu_m\end{bmatrix}\begin{bmatrix} \frac{1}{\sigma_1^2+\sigma_{p_1}^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2^2+\sigma_{p_2}^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_m^2+\sigma_{p_m}^2} \end{bmatrix} 2\begin{bmatrix}\mu_1 \\ \mu_2 \\ \vdots \\ \mu_m\end{bmatrix} = \sum_{i=1}^{m} \frac{4\mu_i^2}{\sigma_i^2+\sigma_{p_i}^2}.$$

Thus, the optimization problem in equation 16 is equivalent to

$$\begin{aligned} \min_{\beta,\sigma_{p_1}^2,\ldots,\sigma_{p_m}^2} \quad & \sum_{i=1}^{m} \frac{\mu_i^2}{\sigma_i^2+\sigma_{p_i}^2} \\ s.t. \quad & \|\beta\|^2 + tr(\Sigma_p) \le D \\ & \sigma_{p_i}^2 \ge 0 \quad \forall i \in \{1,2,...m\}. \end{aligned} \tag{17}$$

Since a non-zero β does not affect the objective function but result in positive distortion, the optimal mechanism satisfies $\beta = (0,...,0)$. Furthermore, the Lagrangian of the above optimization problem is given by

$$L(\sigma_{p_1}^2,...,\sigma_{p_m}^2,\lambda) = \sum_{i=1}^{m} \frac{\mu_i^2}{\sigma_i^2+\sigma_{p_i}^2} + \lambda_0\Big(\sum_{i=1}^{m}\sigma_{p_i}^2 - D\Big) - \sum_{i=1}^{m}\lambda_i\sigma_{p_i}^2, \tag{18}$$

where $\lambda = \{\lambda_0,...,\lambda_m\}$ denotes the Lagrangian multipliers associated with the constraints. Taking the derivatives of $L(\sigma_{p_1}^2,...,\sigma_{p_m}^2,\lambda)$ with respect to $\sigma_{p_i}^2, \forall i \in \{1,...,m\}$, we have

$$\frac{\partial L(\sigma_{p_1}^2,...,\sigma_{p_m}^2,\lambda)}{\partial \sigma_{p_i}^2} = -\frac{\mu_i^2}{(\sigma_i^2+\sigma_{p_i}^2)^2} + \lambda_0 - \lambda_i. \tag{19}$$

Notice that the objective function in equation 16 is decreasing in $\sigma_{p_i}^2, \forall i \in \{1,...,m\}$. Thus, the optimal solution ${\sigma_{p_i}^*}^2$ satisfies $\sum_{i=1}^{m} {\sigma_{p_i}^*}^2 = D$. By the KKT conditions, we have

$$\frac{\partial L(\sigma_{p_1}^2,...,\sigma_{p_m}^2,\lambda)}{\partial \sigma_{p_i}^2}\Big|_{\sigma_{p_i}^2={\sigma_{p_i}^*}^2,\lambda=\lambda^*} = -\frac{\mu_i^2}{(\sigma_i^2+{\sigma_{p_i}^*}^2)^2} + \lambda_0^* - \lambda_i^* = 0. \tag{20}$$

Since $\lambda_i^*, i \in \{0,1,...,m\}$ is dual feasible, we have $\lambda_i^* \ge 0, i \in \{0,1,...,m\}$. Therefore

$$\lambda_0^* \ge \frac{\mu_i^2}{(\sigma_i^2+{\sigma_{p_i}^*}^2)^2}.$$

If $\lambda_0^* > \frac{\mu_i^2}{\sigma_i^4}$, we have $\lambda_0^* > \frac{\mu_i^2}{(\sigma_i^2+{\sigma_{p_i}^*}^2)^2}$. This implies $\lambda_i^* > 0$. Thus, by complementary slackness, ${\sigma_{p_i}^*}^2 = 0$. On the other hand, if $\lambda_0^* < \frac{\mu_i^2}{\sigma_i^4}$, we have ${\sigma_{p_i}^*}^2 > 0$. Furthermore, by the complementary slackness condition, $\lambda_i^*{\sigma_{p_i}^*}^2 = 0, \forall {\sigma_{p_i}^*}^2$. This implies $\lambda_i^* = 0, \forall {\sigma_{p_i}^*}^2 > 0$. As a result, for all ${\sigma_{p_i}^*}^2 > 0$, we have

$$\frac{|\mu_i|}{\sqrt{\lambda_0^*}} = \sigma_i^2 + {\sigma_{p_i}^*}^2. \tag{21}$$

Therefore, ${\sigma_{p_i}^*}^2 = \max\{\frac{|\mu_i|}{\sqrt{\lambda_0^*}} - \sigma_i^2, 0\} = \left(\frac{|\mu_i|}{\sqrt{\lambda_0^*}} - \sigma_i^2\right)^+$ with $\sum_{i=1}^{m} {\sigma_{p_i}^*}^2 = D$. Substitute this optimal solution into equation 11 with $\alpha = \sqrt{(2\mu)^T(\Sigma+\Sigma_p)^{-1}2\mu}$, we obtain the accuracy of the MAP adversary. □

We observe that the when σ_i^2 is greater than some threshold $\frac{|\mu_i|}{\sqrt{\lambda_0^*}}$, no noise is added to the data on this dimension due to the high variance. When σ_i^2 is smaller than $\frac{|\mu_i|}{\sqrt{\lambda_0^*}}$, the amount of noise added to this dimension is proportional to $|\mu_i|$; this is intuitive since a large $|\mu_i|$ indicates the two conditionally Gaussian distributions are further away on this dimension, and thus, distinguishable. Thus, more noise needs to be added in order to reduce the MAP adversary's inference accuracy.

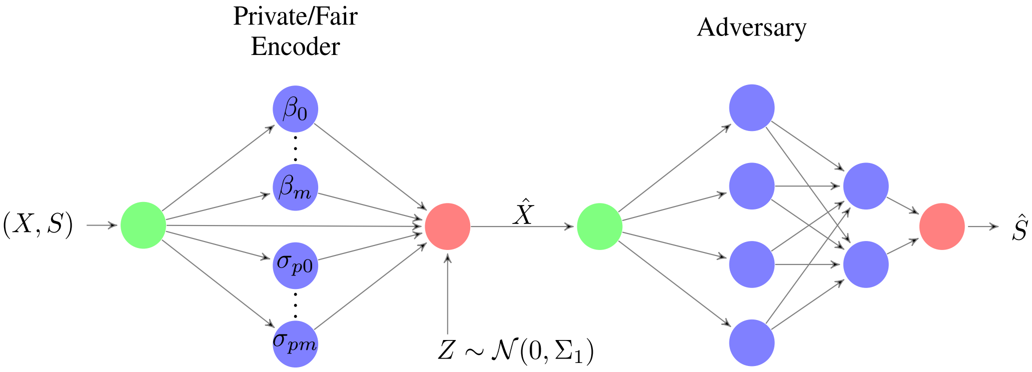


Figure 6: Neural network structure of linear GAPF for Gaussian mixture data

D.2 Data-driven Approach

For the data-driven linear GAPF mechanism, we assume the encoder only has access to the dataset $\mathcal{D}$ with n data samples but not the actual distribution of (X, S). Computing the optimal privacy mechanism becomes a learning problem. In the training phase, the data holder learns the parameters of the GAPF mechanism by competing against a computational adversary modeled by a multi-layer neural network. When convergence is reached, we evaluate the performance of the learned mechanism by comparing with the one obtained from the game-theoretic approach. To quantify the performance of the learned GAPF mechanism, we compute the accuracy of inferring S under a strong MAP adversary that has access to both the joint distribution of (X, S) and the privacy mechanism.

Since the sensitive variable S is binary, we measure the training loss of the adversary network by the empirical log-loss function

$$L_n(\theta_p, \theta_a) = -\frac{1}{n}\sum_{i=1}^{n} s_{(i)} \log h(g(x_{(i)}; \theta_p); \theta_a) + (1 - s_{(i)}) \log(1 - h(g(x_{(i)}; \theta_p); \theta_a)). \tag{22}$$

For a fixed encoder parameter θ_p, the adversary learns the optimal θ_a^* by maximizing equation 22. For a fixed θ_a, the encoder learns the optimal θ_p^* by minimizing $-L_n(h(g(X; \theta_p); \theta_a), S)$ subject to the distortion constraint $\mathbb{E}_{X,\hat{X}}\|X - \hat{X}\|^2 \leq D$.

As shown in Figure 6, the encoder is modeled by a two-layer neural network with parameters $\theta_p = \{\beta_0, ..., \beta_m, \sigma_{p0}, ..., \sigma_{pm}\}$, where β_k and σ_{pk} represent the mean and standard deviation for each dimension $k \in \{1, ..., m\}$, respectively. The random noise Z is drawn from a m-dimensional independent zero-mean standard Gaussian distribution with covariance Σ_1. Thus, we have $\hat{X}_k = X_k + \beta_k + \sigma_{pk} Z_k$. The adversary, whose goal is to infer S from privatized data $\hat{X}$, is modeled by a three-layer neural network classifier with leaky ReLU activations.

To incorporate the distortion constraint into the learning process, we add a penalty term to the objective of the encoder. Thus, the training loss function of the encoder is given by

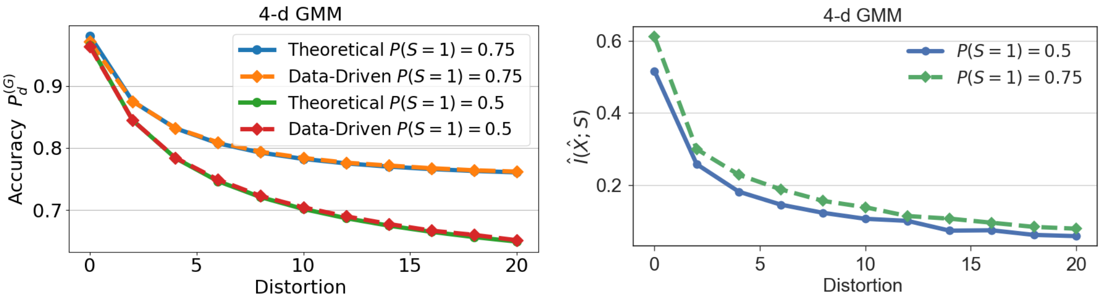
$$L(\theta_p, \theta_a) = L_n(\theta_p, \theta_a) + \rho \max\{0, \frac{1}{n}\sum_{i=1}^{n} d(g(x_{(i)}; \theta_p), x_{(i)}) - D\}, \tag{23}$$

where ρ is a penalty coefficient which increases with the number of iterations. The added penalty consists of a penalty parameter ρ multiplied by a measure of violation of the constraint. This measure of violation is non-zero when the constraint is violated. Otherwise, it is zero.

D.3 Illustration of Results

We use synthetic data generated by Gaussian mixture model as our first attempt to evaluate the performance of the learned GAPF mechanisms. Each dataset contains $20K$ training samples and $2K$ test samples. Each data entry is sampled from an independent multi-dimensional Gaussian mixture model. We consider two categories of synthetic datasets with $P(S = 1)$ equal to 0.75 and 0.5, respectively. Both the encoder and the adversary in the GAPF framework are trained on Tensorflow (Abadi et al., 2016) using Adam optimizer with a learning rate of 0.005 and a minibatch size of 1000. The distortion constraint is enforced by the penalty method as detailed in supplement B (see equation 9).

Figure 2 illustrates the performance of the learned GAPF mechanism against a strong theoretical MAP adversary for $q = 0.75$ and $q = 0.5$. It can be seen that the inference accuracy of the MAP adversary reduces as the distortion increases and asymptotically approaches (as expected) the prior on the sensitive variable. This is because noise adding mechanisms cannot further reduce the accuracy of the MAP adversary than the prior on S. We also observe that the privacy mechanism obtained via the data-driven approach performs very well when pitted against the MAP adversary (maximum accuracy difference around 0.3% compared to the theoretical approach). In other words, for the Gaussian mixture data model with binary sensitive variable, the data-driven version of GAPF can learn privacy mechanisms that perform as well as the mechanisms computed under the theoretical version of GAPF, which assumes that the encoder has access to the underlying distribution of the dataset. Figure 7a compares the inference accuracy of the MAP adversary for GAPF mechanisms obtained from both game-theoretical and data-drive approach under different distortion values. The synthetic dataset used in this simulation is sampled from a Gaussian mixture model with $P(S = 1) = 0.5$.



(a) Classification accuracy

(b) Mutual information

Figure 7: Mutual information of GMM model

E GAPF Architecture for GENKI Dataset

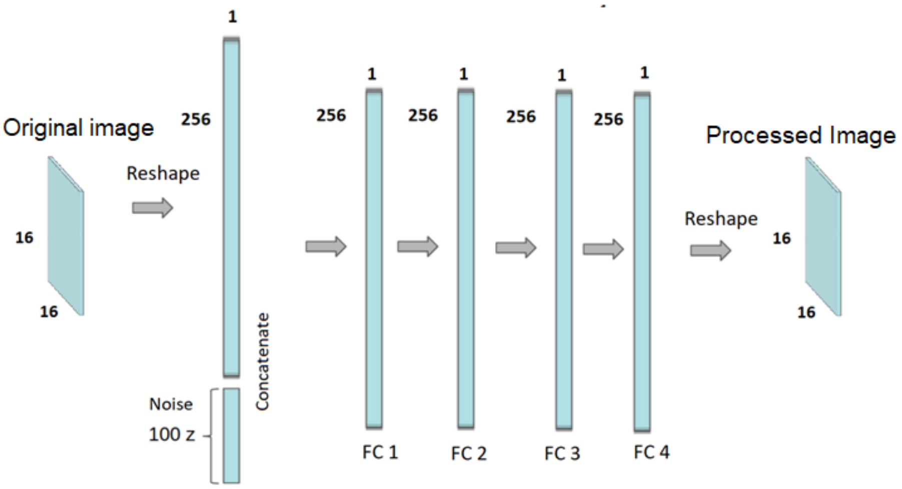


Figure 8: Feedforward neural network encoder

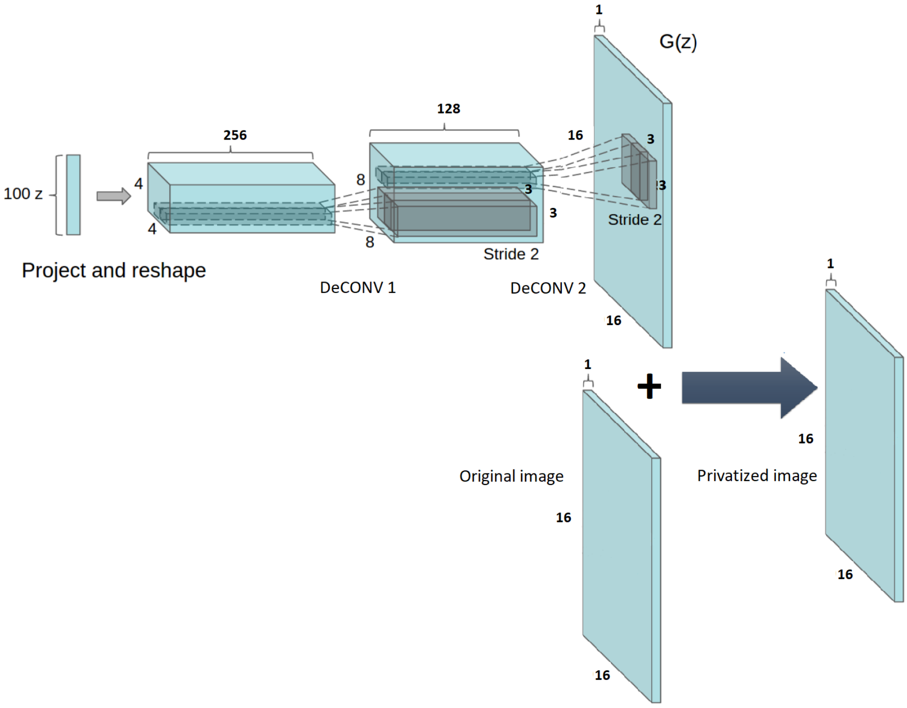


Figure 9: Transposed convolution neural network encoder

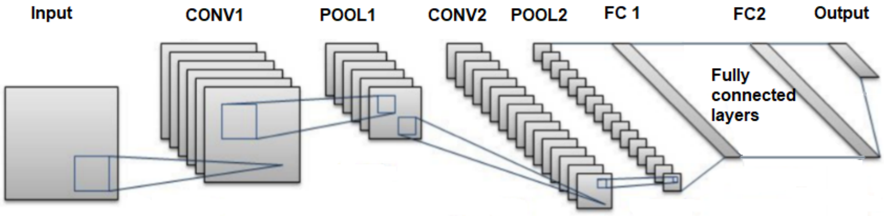


Figure 10: Convolutional neural network adversary

F Mutual Information

Our GAPF framework offers a scalable way to find a (local) equilibrium in the constrained min-max optimization, under certain attacks (e.g. attacks based on a neural network). Yet the privatized data, through our approach, should be immune to any general attacks and ultimately achieving the goal of decreasing the correlation between the privatized data and the sensitive labels. Therefore we use the mutual information to certify that the sensitive data indeed is protected via our framework, from the perspective of information theory.

We calculate the mutual information from data-driven perspective, i.e. estimating the empirical mutual information $\hat{I}(X;Y) = \hat{H}(X) - \hat{H}(X|Y)$, where $\hat{H}$ characterizes the empirical entropy, X is a variable representing released information, and Y is a variable indicating sensitive information. This empirical entropy can be obtained using the classical nearest k-th neighbor method(Kraskov et al., 2004)

$$\hat{H}(X) = \psi(N) - \psi(k) + \log(c_d) + \frac{d}{N}\sum_{i=1}^{N} \log r_i \tag{24}$$

where r_i is the distance of the i-th sample x_i to its k-th nearest neighbor, ψ is the digamma function, $c_d = \frac{\pi^{d/2}}{\Gamma(1+d/2)}$ in Euclidean norm, and N is the number of samples.

In the experiment of GENKI dataset, the sensitive label is gender that is characterized by a binary variable Y. We can express the empirical MI as

$$\hat{I}(X;Y) = \hat{H}(X) - \big(P(Y=1)\hat{H}(X|Y=1) + P(Y=0)\hat{H}(X|Y=0)\big), \tag{25}$$

where $P(Y=1)$ and $P(Y=0)$ can be approximated by the sample frequency.

The empirical mutual information $\hat{I}(X;S)$ for Gaussian Mixture Model(GMM) is displayed in Figure 7b, where X is 4-d GMM with 2 components, and Y is a binary variable that labels the corresponding 2 components.

One noteworthy difficulty is that X usually lives in high dimensions (e.g. each image has 256 dimensions in GENKI dataset) which is almost impossible to calculate the empirical entropy based on original data, because it will need a huge amount of samples. Thus, we train a neural network that classifies the gender of raw images, to perform the role of dimension reduction. More specifically, the layer before the softmax outputs is considered to be a feature embedding that has a much lower dimension than original X, which also captures the information of a image that correlated with its gender. We denote the feature embeddings to be X_f as a surrogate of X. The resulting approximate MI is

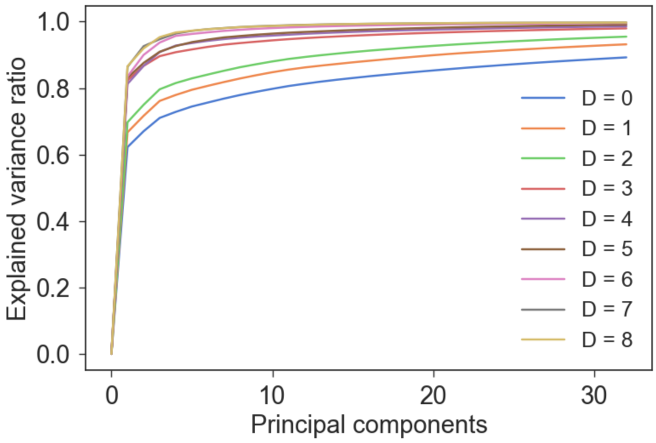
$$\hat{I}(X_f;Y) = \hat{H}(X_f) - \hat{H}(X_f|Y) = \hat{H}(X_f) - \big(P(Y=1)\hat{H}(X_f|Y=1) + P(Y=0)\hat{H}(X_f|Y=0)\big).$$

Following the same manner, the MI between privatized data $\hat{X}$ and Y is approximated by $\hat{I}(\hat{X}_f;Y)$, where $\hat{X}_f$ is the feature embedding that represents a privatized image $\hat{X}$.

Given the GENKI dataset that has the 16×16 gray-scale images, we construct a convolutional neural network initialized by two conv blocks, then followed by two fully connected (FC) layers, and lastly ended with two neurons having the softmax activations. In each conv block, we have a convolution layer consisting of filters with the size equals 3×3 and the stride equals 1, a 2×2 max-pooling layer with the stride equals 2, and a ReLU activation function. Those two conv blocks have 32 and 64 filters respectively. We flatten out the output of second conv block yielding a 256 dimension vector. Then it passes through the first FC layer with batch normalization and ReLU activation to get a 8 dimension vector, followed with the second FC layer to output a 2 dimensional vector that applied with the softmax function. The aforementioned 8-dim vector is the feature embedding vector X_f in our empirical MI estimation.

The Figure 3b shows the change of normalized mutual information, i.e. $\frac{\hat{I}(\hat{X}_f;Y)}{\hat{I}(X_f;Y)}$, with respect to the increasing distortions for the GENKI dataset.

Calculating mutual information for HAR dataset has a slightly different challenge, because the alphabetic size of values that the sensitive label(i.e. identity) can take is 30. Thus, it requires at least 30 neurons prior to the output layer of the corresponding classification task. In fact we pose 128 neurons before the final softmax output layer in order to get a reasonably good classification accuracy. Using the 128-dim vector as our feature embedding to calculate mutual information is almost impossible due to the curse of dimensionality. Therefore, we apply PCA, shown in Figure 11, and pick first 12 component to circumvent this issue. The resulting 12-dim vector is considered to be an approximate feature embedding that encapsulates the major information of the processed data.



(a) Top 32 principal components out of the 561 features with different distortion D

Figure 11: PCA for processed data in HAR