Trade-offs and Guarantees of Adversarial Representation Learning for Information Obfuscation

Han Zhao*
D. E. Shaw & Co.
han.zhao@cs.cmu.edu

Jianfeng Chi*
Department of Computer Science
University of Virginia
jc6ub@virginia.edu

Yuan Tian
Department of Computer Science
University of Virginia
yuant@virginia.edu

Geoffrey J. Gordon
Carnegie Mellon University
Microsoft Research Montreal
geoff.gordon@microsoft.com

Abstract

Crowdsourced data used in machine learning services might carry sensitive information about attributes that users do not want to share. Various methods have been proposed to minimize the potential information leakage of sensitive attributes while maximizing the task accuracy. However, little is known about the theory behind these methods. In light of this gap, we develop a novel theoretical framework for attribute obfuscation. Under our framework, we propose a minimax optimization formulation to protect the given attribute and analyze its inference guarantees against worst-case adversaries. Meanwhile, it is clear that in general there is a tension between minimizing information leakage and maximizing task accuracy. To understand this, we prove an information-theoretic lower bound to precisely characterize the fundamental trade-off between accuracy and information leakage. We conduct experiments on two real-world datasets to corroborate the inference guarantees and validate this trade-off. Our results indicate that, among several alternatives, the adversarial learning approach achieves the best trade-off in terms of attribute obfuscation and accuracy maximization.

1 Introduction

With the growing demand for machine learning systems provided as services, a massive amount of data containing sensitive information, such as race, income level, age, etc., are generated and collected from local users. This poses a substantial challenge and it has become an imperative object of study in machine learning [18], computer vision [6, 34], healthcare [2, 3], speech recognition [30], and many other domains. In this paper, we consider a practical scenario where the prediction vendor requests crowdsourced data for a target task, e.g, scientific modeling. The data owner agrees on the data usage for the target task while she does not want her other sensitive information (*e.g.*, age, race) to be leaked. The goal in this context is then to obfuscate sensitive attributes of the sanitized data released by data owner from potential attribute inference attacks from a malicious adversary. For example, in an online advertising scenario, while the user (data owner) may agree to share her historical purchasing events, she also wants to protect her age information so that no malicious adversary can infer her age range from the shared data. Note that simply removing age attribute

*The first two authors contributed equally to this work. Work done while HZ was at Carnegie Mellon University.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

from the shared data is insufficient for this purpose, due to the redundant encoding in data, i.e., other attributes may have a high correlation with age.

Under this scenario, a line of work [4, 14, 19, 22, 24, 25, 32–34] aims to address the problem in the framework of (constrained) minimax problem. However, the theory behind these methods is little known. Such a gap between theory and practice calls for an important and appealing challenge:

> *Can we prevent the information leakage of the sensitive attribute while still maximizing the task accuracy? Furthermore, what is the fundamental trade-off between attribute obfuscation and accuracy maximization in the minimax problem?*

Under the setting of attribute obfuscation, the notion of information confidentiality should be attribute-specific: the goal is to protect specific attributes from being inferred by malicious adversaries as much as possible. Note that this is in sharp contrast with differential privacy (we systematically compare the related notions in Sec. 5 Related Work), where mechanisms are usually designed to resist worst-case membership query among all the data owners instead of preventing information leakage of the sensitive attribute [11]. From this perspective, our relaxed definition of attribute obfuscation against adversaries also allows for a more flexible design of algorithms with better accuracy.

Our Contributions In this paper, we first formally define the notion of attribute inference attack in our setting and justify why our definitions are particularly suited under our setting. Through the lens of representation learning, we formulate the problem of accuracy maximization with information obfuscation constraint as a minimax optimization problem. To provide a formal guarantee on attribute obfuscation, we prove an information-theoretic lower bound on the inference error of the protected attribute under attacks from arbitrary adversaries. To investigate the relationship between attribute obfuscation and accuracy maximization, we also prove a theorem that formally characterizes the inherent trade-off between these two concepts. We conduct experiments to corroborate our formal guarantees and validate the inherent trade-offs in different attribute obfuscation algorithms. From our empirical results, we conclude that the adversarial representation learning approach achieves the best trade-off in terms of attribute obfuscation and accuracy maximization, among various state-of-the-art attribute obfuscation algorithms.

2 Preliminaries

Problem Setup We focus on the setting where the goal of the adversary is to perform *attribute inference*. This setting is ubiquitous in sever-client paradigm where machine learning is provided as a service (MLaaS, Ribeiro et al. [27]). Formally, there are two parties in the system, namely the prediction vendor and the data owner. We consider the practical scenarios where users agree to contribute their data for specific purposes (*e.g.*, training a machine learning model) but do not want others to infer their sensitive attributes in the data, such as health information, race, gender, etc. The prediction vendor will not collect raw user data but processed user data and the target attribute for the target task. In our setting, we assume the adversary cannot get other auxiliary information than the processed user data. In this case, the adversary can be anyone who can get access to the processed user data to some extent and wants to infer other private information. For example, malicious machine learning service providers are motivated to infer more information from users to do user profiling and targeted advertisements. The goal of the data owner is to provide as much information as possible to the prediction vendor to maximize the vendor's own accuracy, but under the constraint that the data owner should also protect the private information of the data source, *i.e.*, *attribute obfuscation*. For ease of discussion, in our following analysis, we assume the the prediction vendor performs binary classification on the processed data. Extensions to multi-class classification is straightforward.

Notation We use $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{A}$ to denote the input, output and adversary's output space, respectively. Accordingly, we use X, Y, A to denote the random variables which take values in $\mathcal{X}, \mathcal{Y}$ and $\mathcal{A}$. We note that in our framework the input space $\mathcal{X}$ may or may not contain the sensitive attribute A. For two random variables X and Y, $I(X;Y)$ denotes the mutual information between X and Y. We use $H(X)$ to mean the Shannon entropy of random variable X. Similarly, we use $H(X \mid Y)$ to denote the conditional entropy of X given Y. We assume there is a joint distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y} \times \mathcal{A}$ from which the data are sampled. To make our notation consistent, we use $\mathcal{D}_{\mathcal{X}}$, $\mathcal{D}_{\mathcal{Y}}$ and $\mathcal{D}_{\mathcal{A}}$ to denote the marginal distribution of $\mathcal{D}$ over $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{A}$. Given a feature map function $f : \mathcal{X} \to \mathcal{Z}$ that maps instances from the input space $\mathcal{X}$ to feature space $\mathcal{Z}$, we define

$\mathcal{D}^f := \mathcal{D} \circ f^{-1}$ to be the induced (pushforward) distribution of $\mathcal{D}$ under f, i.e., for any event $E' \subseteq \mathcal{Z}$, $\Pr_{\mathcal{D}^f}(E') := \Pr_{\mathcal{D}}(\{x \in \mathcal{X} \mid f(x) \in E'\})$.

To simplify the exposition, we mainly discuss the setting where $\mathcal{X} \subseteq \mathbb{R}^d, \mathcal{Y} = \mathcal{A} = \{0,1\}$, but the underlying theory and methodology could easily be extended to the categorical case as well. In what follows, we first formally define both the *accuracy* of the prediction vendor for the individualized service and the *attribute inference advantage* of an adversary. It is worth pointing out that our definition of inference advantage is *attribute-specific*. In particular, we seek to keep the data useful while being robust to an adversary on protecting specific attribute information from attack.

A *hypothesis* is a function $h : \mathcal{X} \to \mathcal{Y}$. The *error* of a hypothesis h under the distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ is defined as: $\mathrm{Err}(h) := \mathbb{E}_{\mathcal{D}}\big[|Y - h(X)|\big]$. Similarly, we use $\widehat{\mathrm{Err}}(h)$ to denote the empirical error of h on a sample from $\mathcal{D}$. For binary classification problem, when $h(\mathbf{x}) \in \{0,1\}$, the above loss also reduces to the error rate of classification. Let $\mathcal{H}$ be the space of hypotheses. In the context of binary classification, we define the accuracy of a hypothesis $h \in \mathcal{H}$ as:

Definition 2.1 (Accuracy)**.** The accuracy of $h \in \mathcal{H}$ is $\textsc{Acc}(h) := 1 - \mathbb{E}_{\mathcal{D}}\big[|Y - h(X)|\big]$.

For binary classification, we always have $0 \leq \textsc{Acc}(h) \leq 1$, $\forall h \in \mathcal{H}$. Similarly, an *adversarial hypothesis* is a function of $h_A : \mathcal{X} \to \mathcal{A}$. Next we define a measure of how much advantage of attribute inference gained from a particular attack space in our framework:

Definition 2.2 (Attribute Inference Advantage)**.** The inference advantage w.r.t. attribute A under attacks from $\mathcal{H}_A$ is defined as $\textsc{Adv}(\mathcal{H}_A) := \max_{h_A \in \mathcal{H}_A} \big|\Pr_{\mathcal{D}}(h_A(X) = 1 \mid A = 1) - \Pr_{\mathcal{D}}(h_A(X) = 1 \mid A = 0)\big|$.

Again, it is straightforward to verify that $0 \leq \textsc{Adv}(\mathcal{H}_A) \leq 1$. Based on our definition, $\textsc{Adv}(\mathcal{H}_A)$ then measures maximal inference advantage that the adversary in $\mathcal{H}_A$ can gain. We can also refine the above definition to a particular hypothesis $h_A : \mathcal{X} \to \{0,1\}$ to measure its ability to steal information about A: $\textsc{Adv}(h_A) = \big|\Pr_{\mathcal{D}}(h_A(X) = 1 \mid A = 1) - \Pr_{\mathcal{D}}(h_A(X) = 1 \mid A = 0)\big|$.

Proposition 2.1. Let $h_A : \mathcal{X} \to \{0,1\}$ be a hypothesis, then $\textsc{Adv}(h_A) = 0$ iff $I(h_A(X); A) = 0$ and $\textsc{Adv}(h_A) = 1$ iff $h_A(X) = A$ almost surely or $h_A(X) = 1 - A$ almost surely.

Proposition 2.1 justifies Definition 2.2 on how well an adversary h_A can infer about A from X: when $\textsc{Adv}(h_A) = 0$, it means that $h_A(X)$ contains no information about the sensitive attribute A. On the other hand, if $\textsc{Adv}(h_A) = 1$, then $h_A(X)$ fully predicts A (or equivalently, $1 - A$) from input X. In the latter case $h_A(X)$ also contains perfect information of A in the sense that $I(h_A(X); A) = H(A)$, i.e., the Shannon entropy of A. It is worth pointing out that Definition 2.2 is insensitive to the marginal distribution of A, and hence is more robust than other definitions such as the error rate of predicting A. In that case, if A is extremely imbalanced, even a naive predictor can attain small prediction error by simply outputting constant. We call a hypothesis space $\mathcal{H}_A$ *symmetric* if $\forall h_A \in \mathcal{H}_A, 1 - h_A \in \mathcal{H}_A$ as well. When $\mathcal{H}_A$ is symmetric, we can also relate $\textsc{Adv}(\mathcal{H}_A)$ to a binary classification problem:

Proposition 2.2. If $\mathcal{H}_A$ is symmetric, then $\textsc{Adv}(\mathcal{H}_A) + \min_{h_A \in \mathcal{H}_A} \Pr(h_A(X) = 0 \mid A = 1) + \Pr(h_A(X) = 1 \mid A = 0) = 1$.

Consider the confusion matrix between the actual sensitive attribute A and its predicted variable $h_A(X)$. The false positive rate (eqv. Type-I error) is defined as FPR = FP / (FP + TN) and the false negative rate (eqv. Type-II error) is similarly defined as FNR = FN / (FN + TP). Using the terminology of confusion matrix, it is then clear that $\Pr(h_A(X) = 0 \mid A = 1) = \mathrm{FNR}$ and $\Pr(h_A(X) = 1 \mid A = 0) = \mathrm{FPR}$. In other words, Proposition 2.2 says that if $\mathcal{H}_A$ is symmetric, then the larger the attribute inference advantage of $\mathcal{H}_A$, the smaller the minimum sum of Type-I and Type-II error under attacks from $\mathcal{H}_A$.

3 Main Results

Given a set of samples $\mathbf{S} = \{(\mathbf{x}_i, y_i, a_i)\}_{i=1}^n$ drawn i.i.d. from the joint distribution $\mathcal{D}$, how can the data owner keep the data useful while keeping the sensitive attribute A obfuscated under potential attacks from malicious adversaries? Through the lens of representation learning, we seek to find a (non-linear) feature representation $f : \mathcal{X} \to \mathcal{Z}$ from input space $\mathcal{X}$ to feature space $\mathcal{Z}$ such that f

still preserves relevant information w.r.t. the target task of inferring Y while hiding sensitive attribute A. Specifically, we can solve the following unconstrained regularized problem with $\lambda > 0$:

$$\min_{h \in \mathcal{H}, f} \max_{h_A \in \mathcal{H}_A} \quad \widehat{\mathrm{Err}}(h \circ f) - \lambda \big(\Pr_{\mathbf{S}}(h_A(f(X)) = 0 \mid A = 1) + \Pr_{\mathbf{S}}(h_A(f(X)) = 1 \mid A = 0) \big) \tag{1}$$

It is worth pointing out that the optimization formulation in (1) admits an interesting game-theoretic interpretation, where two agents f and h_A play a game whose score is defined by the objective function in (1). Intuitively, h_A seeks to minimize the sum of Type-I and Type-II error while f plays against h_A by learning transformation to removing information about the sensitive attribute A. Algorithmically, for the data owner to achieve the goal of hiding information about the sensitive attribute A from malicious adversary, it suffices to learn a representation that is independent of A:

Proposition 3.1. Let $f : \mathcal{X} \to \mathcal{Z}$ be a deterministic function and $\mathcal{H}_A \subseteq 2^{\mathcal{Z}}$ be a hypothesis class over $\mathcal{Z}$. For any joint distribution $\mathcal{D}$ over X, A, Y, if $I(f(X); A) = 0$, then $\textsc{Adv}(\mathcal{H}_A \circ f) = 0$.

Note that in this sequential game, f is the first-mover and h_A is the second. Hence without explicit constraint f possesses a first-mover advantage so that f can dominate the game by simply mapping all the input X to a constant or uniformly random noise[2]. To avoid these degenerate cases, the first term in the objective function of (1) acts as an incentive to encourage f to preserve task-related information. But will this incentive compromise the information of A? As an extreme case if the target variable Y and the sensitive attribute A are perfectly correlated, then it should be clear that there is a trade-off in achieving accuracy and preventing information leakage of the attribute. In Sec. 3.2 we shall provide an information-theoretic bound to precisely characterize such trade-off.

3.1 Formal Guarantees against Attribute Inference

In the unconstrained minimax formulation (1), the hyperparameter λ measures the trade-off between accuracy and information obfuscation. On one hand, if $\lambda \to 0$, we barely care about the information obfuscation of A and devote all the focus to maximize our accuracy. On the other extreme, if $\lambda \to \infty$, we are only interested in obfuscating the sensitive information. In what follows we analyze the true error that an optimal adversary has to incur in the limit when both the task classifier and the adversary have unlimited capacity, i.e., they can be any randomized functions from $\mathcal{Z}$ to $\{0, 1\}$. To study the true error, we hence use the population loss rather than the empirical loss in our objective function. Furthermore, since the binary classification error in (1) is NP-hard to optimize even for hypothesis class of linear predictors, in practice we consider the cross-entropy loss function as a convex surrogate loss. With a slight abuse of notation, the cross-entropy loss $\mathrm{CE}_Y(h)$ of a probabilistic hypothesis $h : \mathcal{X} \to [0, 1]$ w.r.t. Y on a distribution $\mathcal{D}$ is defined as follows:

$$\mathrm{CE}_Y(h) := -\mathbb{E}_{\mathcal{D}}[\mathbb{I}(Y = 0) \log(1 - h(X)) + \mathbb{I}(Y = 1) \log(h(X))].$$

We also use $\mathrm{CE}_A(h_A)$ to mean the cross-entropy loss of the adversary h_A w.r.t. A. Using the same notation, the optimization formulation with cross-entropy loss becomes:

$$\min_{h \in \mathcal{H}, f} \max_{h_A \in \mathcal{H}_A} \quad \mathrm{CE}_Y(h \circ f) - \lambda \cdot \mathrm{CE}_A(h_A \circ f) \tag{2}$$

Given a feature map $f : \mathcal{X} \to \mathcal{Z}$, assume that $\mathcal{H}$ contains all the possible probabilistic classifiers from the feature space $\mathcal{Z}$ to $[0, 1]$. For example, a probabilistic classifier can be constructed by first defining a function $h : \mathcal{Z} \to [0, 1]$ followed by a random coin flipping to determine the output label, where the probability of the coin being 1 is given by $h(Z)$. Under such assumptions, the following lemma shows that the optimal target classifier under f is given by the conditional distribution $h^*(Z) := \Pr(Y = 1 \mid Z)$.

Lemma 3.1. For any feature map $f : \mathcal{X} \to \mathcal{Z}$, assume that $\mathcal{H}$ contains all the probabilistic classifiers, then $\min_{h \in \mathcal{H}} \mathrm{CE}_Y(h \circ f) = H(Y \mid Z)$ and $h^*(Z) := \arg\min_{h \in \mathcal{H}} \mathrm{CE}_Y(h \circ f) = \Pr(Y = 1 \mid Z = f(X))$.

By a symmetric argument, we can also see that the worst-case (optimal) adversary under f is the conditional distribution $h_A^*(Z) := \Pr(A = 1 \mid Z)$ and $\min_{h_A \in \mathcal{H}_A} \mathrm{CE}_A(h_A \circ f) = H(A \mid Z)$.

[2]The extension of Proposition 3.1 to randomized function is staightforward as long as the randomness is independent of the sensitive attribute A.

Hence we can further simplify the optimization formulation (2) to the following form where the only optimization variable is the feature map f:

$$\min_{f} \quad H(Y \mid Z = f(X)) - \lambda H(A \mid Z = f(X)) \tag{3}$$

Since $Z = f(X)$ is a deterministic feature map, it follows from the basic properties of Shannon entropy that

$$H(Y \mid X) \leq H(Y \mid Z = f(X)) \leq H(Y), \qquad H(A \mid X) \leq H(A \mid Z = f(X)) \leq H(A)$$

which means that $H(Y \mid X) - \lambda H(A)$ is a lower bound of the optimum of the objective function in (3). However, such lower bound is not necessarily achievable. To see this, consider the simple case where $Y = A$ almost surely. In this case there exists no deterministic feature map $Z = f(X)$ that is both a sufficient statistics of X w.r.t. Y while simultaneously filters out all the information w.r.t. A except in the degenerate case where $A(Y)$ is constant. Next, to show that solving the optimization problem in (3) helps to remove sensitive information, the following theorem gives a bound of attribute inference in terms of the error that has to be incurred by the optimal adversary:

Theorem 3.1. Let f^* be the optimal feature map of (3) and define $H^* := H(A \mid Z = f^*(X))$. Then for any adversary $\widehat{A}$ such that $I(\widehat{A}; A \mid Z) = 0$, $\Pr_{\mathcal{D}^{f^*}}(\widehat{A} \neq A) \geq H^*/2\lg(6/H^*)$.

Remark Theorem 3.1 shows that whenever the conditional entropy $H^* = H(A \mid Z = f^*(X))$ is large, then the inference error of the protected attribute incurred by any (randomized) adversary has to be at least $\Omega(H^*/\log(1/H^*))$. The assumption $I(\widehat{A}; A \mid Z) = 0$ says that, given the processed feature Z, the adversary $\widehat{A}$ could not use any external information that depends on the true sensitive attribute A. As we have already shown above, the conditional entropy essentially corresponds to the second term in our objective function, whose optimal value could further be flexibly adjusted by tuning the trade-off parameter λ. As a final note, Theorem 3.1 also shows that representation learning helps to remove the information about A since we always have $H(A \mid Z = f(X)) \geq H(A \mid X)$ for any deterministic feature map f so that the lower bound of inference error by any adversary is larger after learning the representation $Z = f(X)$.

3.2 Inherent trade-off between Accuracy and Attribute Obfuscation

In this section we shall provide an information-theoretic bound to quantitatively characterize the inherent trade-off between these accuracy maximization and attribute obfuscation, due to the discrepancy between the conditional distributions of the target variable given the sensitive attribute. Our result is algorithm-independent, hence it applies to a general setting where there is a need to preserve both terms. To the best of our knowledge, this is the first information-theoretic result to precisely quantify such trade-off. Due to space limit, we defer all the proofs to the appendix.

Before we proceed, we first define several information-theoretic concepts that will be used in our analysis. For two distributions $\mathcal{D}$ and $\mathcal{D}'$, the Jensen-Shannon (JS) divergence $D_{\text{JS}}(\mathcal{D}, \mathcal{D}')$ is: $D_{\text{JS}}(\mathcal{D}, \mathcal{D}') := \frac{1}{2}D_{\text{KL}}(\mathcal{D} \parallel \mathcal{D}_M) + \frac{1}{2}D_{\text{KL}}(\mathcal{D}' \parallel \mathcal{D}_M)$, where $D_{\text{KL}}(\cdot \parallel \cdot)$ is the Kullback–Leibler (KL) divergence and $\mathcal{D}_M := (\mathcal{D} + \mathcal{D}')/2$. The JS divergence can be viewed as a symmetrized and smoothed version of the KL divergence, However, unlike the KL divergence, the JS divergence is bounded: $0 \leq D_{\text{JS}}(\mathcal{D}, \mathcal{D}') \leq 1$. Additionally, from the JS divergence, we can define a distance metric between two distributions as well, known as the JS distance [13]: $d_{\text{JS}}(\mathcal{D}, \mathcal{D}') := \sqrt{D_{\text{JS}}(\mathcal{D}, \mathcal{D}')}$. With respect to the JS distance, for any feature space $\mathcal{Z}$ and any deterministic mapping $f : \mathcal{X} \to \mathcal{Z}$, we can prove the following lemma via the celebrated data processing inequality:

Lemma 3.2. Let $\mathcal{D}_0$ and $\mathcal{D}_1$ be two distributions over $\mathcal{X}$ and let $\mathcal{D}_0^f$ and $\mathcal{D}_1^f$ be the induced distributions of $\mathcal{D}_0$ and $\mathcal{D}_1$ over $\mathcal{Z}$ by function f, then $d_{\text{JS}}(\mathcal{D}_0^f, \mathcal{D}_1^f) \leq d_{\text{JS}}(\mathcal{D}_0, \mathcal{D}_1)$.

Without loss of generality, any method aiming to predict the target variable Y defines a Markov chain as $X \xrightarrow{f} Z \xrightarrow{h} \hat{Y}$, where $\hat{Y}$ is the predicted target variable given by hypothesis h and Z is the intermediate representation defined by the feature mapping f. Hence for any distribution $\mathcal{D}_0(\mathcal{D}_1)$ of X, this Markov chain also induces a distribution $\mathcal{D}_0^{h\circ f}(\mathcal{D}_1^{h\circ f})$ of $\hat{Y}$ and a distribution $\mathcal{D}_0^f(\mathcal{D}_1^f)$ of Z. Now let $\mathcal{D}_0^Y(\mathcal{D}_1^Y)$ be the underlying true conditional distribution of Y given $A = 0(A = 1)$. Realize

that the JS distance is a metric, the following chain of triangular inequalities holds:

$$d_{\text{JS}}(\mathcal{D}_0^Y, \mathcal{D}_1^Y) \leq d_{\text{JS}}(\mathcal{D}_0^Y, \mathcal{D}_0^{h\circ f}) + d_{\text{JS}}(\mathcal{D}_0^{h\circ f}, \mathcal{D}_1^{h\circ f}) + d_{\text{JS}}(\mathcal{D}_1^{h\circ f}, \mathcal{D}_1^Y).$$

Combining the above inequality with Lemma 3.2 to show $d_{\text{JS}}(\mathcal{D}_0^{h\circ f}, \mathcal{D}_1^{h\circ f}) \leq d_{\text{JS}}(\mathcal{D}_0^f, \mathcal{D}_1^f)$, we immediately have:

$$d_{\text{JS}}(\mathcal{D}_0^Y, \mathcal{D}_1^Y) \leq d_{\text{JS}}(\mathcal{D}_0^Y, \mathcal{D}_0^{h\circ f}) + d_{\text{JS}}(\mathcal{D}_0^f, \mathcal{D}_1^f) + d_{\text{JS}}(\mathcal{D}_1^{h\circ f}, \mathcal{D}_1^Y).$$

Intuitively, $d_{\text{JS}}(\mathcal{D}_0^Y, \mathcal{D}_0^{h\circ f})$ and $d_{\text{JS}}(\mathcal{D}_1^Y, \mathcal{D}_1^{h\circ f})$ measure the distance between the predicted and the true target distribution on $A = 0/1$ cases, respectively. Formally, let $\text{Err}_a(h \circ f)$ be the prediction error of function $h \circ f$ conditioned on $A = a$. With the help of Lemma A.2, the following result establishes a relationship between $d_{\text{JS}}(\mathcal{D}_a^Y, \mathcal{D}_a^{h\circ f})$ and the accuracy of $h \circ f$:

Lemma 3.3. Let $\hat{Y} = h(f(X)) \in \{0,1\}$ be the predictor, then for $a \in \{0,1\}$, $d_{\text{JS}}(\mathcal{D}_a^Y, \mathcal{D}_a^{h\circ f}) \leq \sqrt{\text{Err}_a(h \circ f)}$.

Combine Lemma 3.2 and Lemma 3.3, we get the following key lemma that is the backbone for proving the main results in this section:

Lemma 3.4 (Key lemma)**.** Let $\mathcal{D}_0$, $\mathcal{D}_1$ be two distributions over $\mathcal{X} \times \mathcal{Y}$ conditioned on $A = 0$ and $A = 1$ respectively. Assume the Markov chain $X \xrightarrow{f} Z \xrightarrow{h} \hat{Y}$ holds, then $\forall h \in \mathcal{H}$:

$$d_{\text{JS}}(\mathcal{D}_0^Y, \mathcal{D}_1^Y) \leq \sqrt{\text{Err}_0(h \circ f)} + d_{\text{JS}}(\mathcal{D}_0^f, \mathcal{D}_1^f) + \sqrt{\text{Err}_1(h \circ f)}.$$

We emphasize that for $a \in \{0,1\}$, the term $\text{Err}_a(h \circ f)$ measures the conditional error of the predicted variable $\hat{Y}$ by the composite function $h \circ f$ over $\mathcal{D}_a$. Similarly, we can define the *conditional accuracy* for $a \in \{0,1\}$: $\text{ACC}_a(h \circ f) := 1 - \text{Err}_a(h \circ f)$. The following main theorem then characterizes a fundamental trade-off between accuracy and attribute obfuscation:

Theorem 3.2. Let $\mathcal{H}_A \subseteq 2^{\mathcal{Z}}$ be the hypothesis space of all the classifiers from $\mathcal{Z}$ to $\{0,1\}$. Assume the conditions in Lemma 3.4 hold, then $\forall h \in \mathcal{H}$, $\text{ACC}_0(h \circ f) + \text{ACC}_1(h \circ f) \leq 2 - \frac{1}{3} D_{\text{JS}}(\mathcal{D}_0^Y, \mathcal{D}_1^Y) + \text{ADV}(\mathcal{H}_A \circ f)$.

The upper bound given in Theorem 3.2 shows that when the marginal distribution of the target variable Y differ between two cases $A = 0$ or $A = 1$, then it is impossible to perfectly maximize accuracy and prevent the sensitive attribute being inferred. Furthermore, the trade-off due to the difference in marginal distributions is precisely given by the JS divergence $D_{\text{JS}}(\mathcal{D}_0^Y, \mathcal{D}_1^Y)$. Next, if we would like to decrease the advantage of adversaries, $\text{ADV}(\mathcal{H}_A \circ f)$, through learning proper feature transformation f, then the upper bound on the sum of conditional accuracy also becomes smaller, for any predictor h. Note that in Theorem 3.2 the upper bound holds for *any* adversarial hypothesis h_A in the richest hypothesis class $\mathcal{H}_A$ that contains all the possible binary classifiers. Put it another way, if we would like to maximally obfuscate information w.r.t. sensitive attribute A, then we have to incur a large joint error:

Theorem 3.3. Assume the conditions in Theorem 3.2 hold. If $\text{ADV}(\mathcal{H}_A \circ f) \leq D_{\text{JS}}(\mathcal{D}_0^Y, \mathcal{D}_1^Y)$, then $\forall h \in \mathcal{H}$, $\text{Err}_0(h \circ f) + \text{Err}_1(h \circ f) \geq \frac{1}{2}\big(d_{\text{JS}}(\mathcal{D}_0^Y, \mathcal{D}_1^Y) - \sqrt{\text{ADV}(\mathcal{H}_A \circ f)}\big)^2$.

Remark The above lower bound characterizes a fundamental trade-off between information obfuscation of the sensitive attribute and joint error of target task. In particular, up to a certain level $D_{\text{JS}}(\mathcal{D}_0^Y, \mathcal{D}_1^Y)$, the larger the inference advantage that the adversary can gain, the smaller the joint error. In light of Proposition 3.1, this means that although the data-owner, or the first-mover f, could try to maximally filter out the sensitive information via constructing f such that $f(X)$ is independent of A, such construction will also inevitably compromise the joint accuracy of the prediction vendor. It is also worth pointing out that our results in both Theorem 3.2 and Theorem 3.3 are attribute-independent in the sense that neither of the bounds depends on the marginal distribution of A. Instead, all the terms in our results only depend on the conditional distributions given $A = 0$ and $A = 1$. This is often more desirable than bounds involving mutual information, *e.g.*, $I(A, Y)$, since $I(A, Y)$ is close to 0 if the marginal distribution of A is highly imbalanced.

4 Experiments

In this section, we conduct experiments to investigate the following questions:

- **Q1** Are our formal guarantees valid for different attribute obfuscation methods and the inherent trade-offs between attribute information obfuscation and accuracy maximization exist in all methods?
- **Q2** Which attribute obfuscation algorithms achieve the best trade-offs in terms of attribute obfuscation and accuracy maximization?

4.1 Datasets and Setup

In our experiments, we use: (1) Adult dataset [8]: The Adult dataset is a benchmark dataset for classification. The task is to predict whether an individual's income is greater or less than 50K/year based on census data. In our experiment we set the target task as income prediction and the malicious task done by the adversary as inferring gender, age and education, respectively. (2) UTKFace dataset [38]: The UTKFace dataset is a large-scale face benchmark dataset containing more than 20,000 images with annotations of age, gender, and ethnicity. In our experiment, we set our target task as gender classification and we use the age and ethnicity as the protected attributes. We refer readers to Sec. C in the appendix for detailed descriptions about the data pre-processing pipeline and the data distribution for each dataset.

We conduct experiments with the following methods to verify our theoretical results and provide a thorough practical comparison among these methods. 1). Privacy Partial Least Squares (PPLS) [14], 2). Privacy Linear Discriminant Analysis (PLDA) [33], 3). Minimax filter with alternative update (ALT-UP) [19], 4) Maximum Entropy Adversarial Representation Learning (MAX-ENT) [28] 5). Gradient Reversal Layer (GRL) [17] 6). Principal Component Analysis (PCA) 7). Normal Training (NORM-TRAIN), 8) Local Differential Privacy (LDP) with Laplacian mechanism, 9). Differentially Private SGD (DPSGD) [1]. Among the first seven methods, the first five are state-of-the-art minimax methods for protecting against attribute inference attacks while the latter two are normal representation learning baselines for comprehensive comparison. Although DP is not tailored to attribute obfuscation, we can still add two DP baselines to examine the accuracy and attribute obfuscation trade-off for comparison[3]. To ensure the comparison is fair among different methods, we conduct a controlled experiment by using the same network structure as the baseline hypothesis among all the methods for each dataset. For each experiment on the Adult dataset and UTKFace dataset, we repeat the experiments for ten times to report both the average performance and their standard deviations. Sec. C in the appendix provides detailed descriptions about the methods and the hyperparameter settings.

Note that in practice due to the non-convex nature of optimizing deep neural nets, we cannot guarantee to find the global optimal conditional entropy H^*. Hence in order to compute the formal guarantee given by our lower bound in Theorem 3.1, we use the cross-entropy loss of the optimal adversary found by our algorithm on inferring the sensitive attribute A. Furthermore, since our analysis only applies to representation learning based approaches, we do not have similar guarantee for DP-related methods in our context. We visualize the performances of the aforementioned algorithms on attribute obfuscation and accuracy maximization in Figure 1 and Figure 2, respectively.

4.2 Results and Analysis

Validation of Our Theory (Q1) From Figure 1, we can see that the formal guarantees are valid for all representation learning approaches. With the results in Figure 2, we also see that no methods are perfect in both achieving both attribute obfuscation and accuracy maximization: the methods with small accuracy loss comes with relative low inference errors and vice versa.

Comparison with Different Methods (Q2) Among all methods, LDP, PLDA, ALT-UP, MAX-ENT and GRL are effective in attribute obfuscation by forcing the optimal adversary to incur a large inference error in Figure 1. On the other hand, PCA and NORM-TRAIN are the least effective ones. This is expected as neither NORM-TRAIN nor PCA filters information in data about the sensitive attribute A.

[3] Some other methods [24, 31] in the literature are close variants of the above, so we do not include them here due to the space limit.

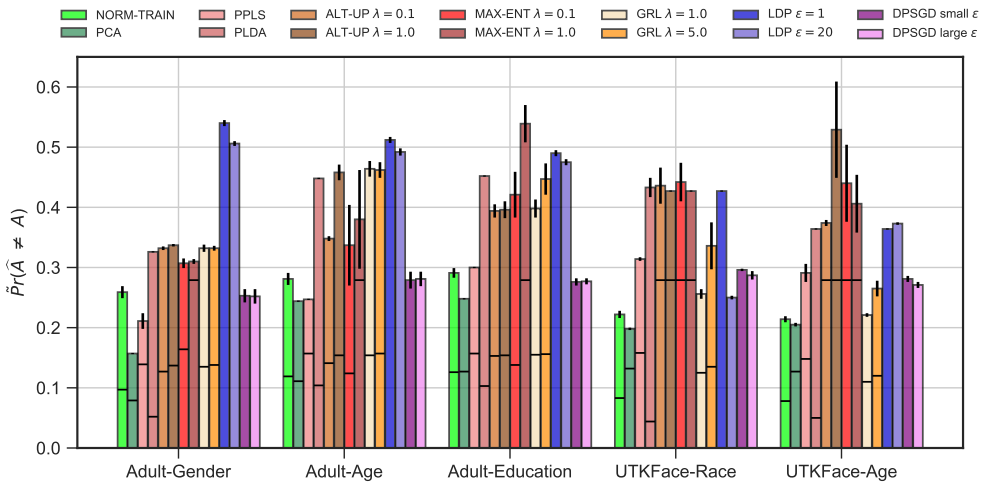

Figure 1: Performance on attribute obfuscation of different methods (the larger the better). The horizontal lines across the bars indicate the corresponding formal guarantees given by our lower bound in Theorem 3.1.

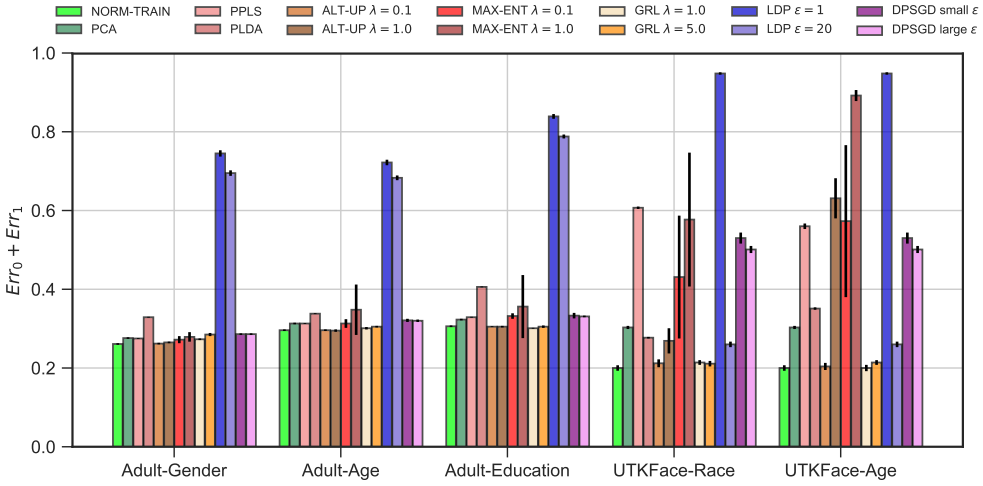

Figure 2: The joint conditional error ($\text{Err}_0 + \text{Err}_1$, the smaller the better) of different methods.

From Figure 2, we can also see a sharp contrast between DP-based methods and other methods in terms of the joint conditional error on the target task: both LDP and DPSGD could incur significant accuracy loss compared with other methods. Combining this one with our previous observation from Figure 1, we can see that DP-based methods either make data private by adding large amount of noise to filter out both target-related information and sensitive-related information available in the data, or add insufficient amount of noise so that both target-related and sensitive-related information is well preserved. As a comparison, representation learning based approaches leads to a better trade-off.

Among the representation learning methods, PLDA, ALT-UP, MAX-ENT and GRL perform the best in attribute obfuscation. Compared to PLDA and GRL, ALT-UP and MAX-ENT incur significant drops in accuracy when λ is large. It is also worth to note that different adversarial representation learning methods have different sensitivity on λ: a large λ for MAX-ENT might lead to an unstable model training process and result in a large accuracy loss. In contrast, GRL is often more stable, which is consistent to the results shown in [7].

5 Related Work

Attribute Obfuscation Various minimax formulations and algorithms have been proposed to defend against inference attack in different scenarios [4, 14, 16, 19, 22, 24, 25, 33, 34]. Bertran et al.

[4] proposed the optimization problem where the terms in the objective function are defined in terms of mutual information. Under their formulation, they analyze a trade-off between between utility loss and attribute obfuscation: under the constraint of the attribute obfuscation $I(A; Z) \leq k$, what the maximum utility loss $I(Y; X \mid Z)$ is. Compared with these works, we study the inherent trade-off between the accuracy and attribute obfuscation and provide formal guarantees to quantify worst-case inference error given the transformation.

Differential Privacy Differential privacy (DP) has been proposed and extensively investigated to protect the individual privacy of collected data [9] and DP mechanisms were used in the training of deep neural network recently [1, 26]. DP ensures the output distribution of a randomized mechanism to be statistically indistinguishable between any two neighboring datasets, and provides formal guarantees for privacy problems such as defending against the membership query attacks [21, 29]. From this perspective, DP is closely related to the well-known membership inference attack [29] instead. As a comparison, our goal of attribute obfuscation is to learn a representation such that the sensitive attributes cannot be accurately inferred. Although the two goals differ, Yeom et al. [36] show the there are deep connections between membership inference and attribute inference. An interesting direction to explore is to draw more formal connections to these two notions. Last but not least, It is also worth to mention that the notion of individual fairness may be viewed as a generalization of DP [10].

Algorithmic Fairness Recent work has shown that unfair models could lead to the leakage of users' sensitive information [35]. In particular, adversarial learning methods have been used as a tool in both fields to achieve the corresponding goals [12, 19]. However, the motivations and goals significantly differ between these two fields. Specifically, the widely adopted notion of group fairness, namely equalized odds [20], requires equalized false positive and false negative rates across different demographic subgroups. As a comparison, in applications where information leakage is a concern, we mainly want to ensure that adversaries cannot steal sensitive information from the data. Hence our goal is to give a worst case guarantee on the inference error that any adversary has at least to incur. To the best of our knowledge, our results in Theorem 3.1 is the first one to analyze the performance of attribute obfuscation in such scenarios. Furthermore, no prior theoretical results exist on discussing the trade-off between attribute obfuscation and accuracy under the setting of representation learning. Our proof techniques developed in this work could also be used to derive information-theoretic lower bounds in related problems as well [39, 40]. On a final note, the relationships of the above notions are visualized in Figure 3.

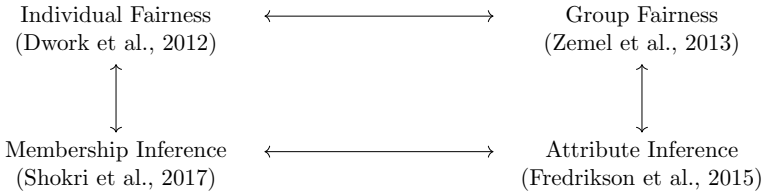

Figure 3: Relationships between different notions of fairness and inference attack.

6 Conclusion

We develop a theoretical framework for analyzing attribute obfuscation through adversarial representation learning. Specifically, the framework suggests using adversarial learning techniques to obfuscate the sensitive attribute and we also analyze the formal guarantees of such techniques in the limit of worst-case adversaries. We also prove an information-theoretic lower bound to quantify the inherent trade-off between accuracy and obfuscation of attribute information. Following our formulation, we conduct experiments to corroborate our theoretical results and to empirically compare different state-of-the-art attribute obfuscation algorithms. Experimental results show that the adversarial representation learning approaches are effective against attribute inference attacks and often achieve the best trade-off in terms of attribute obfuscation and accuracy maximization. We believe our work takes an important step towards better understanding the trade-off between accuracy maximization and attribute obfuscation, and it also helps inspire the future design of attribute obfuscation algorithms with adversarial learning techniques.

Acknowledgements

HZ and GG would like to acknowledge support from the DARPA XAI project, contract #FA87501720152 and a Nvidia GPU grant. JC and YT would like to acknowledge support from NSF OAC 2002985 and NSF CNS 1943100.

Broader Impact

In the process of data collection and information sharing, the data might contain sensitive information that the users are unwilling to disclose. This poses severe challenges for regulations such as GDPR [15] that aims to control the uses and purposes of the collected and shared data. Our work takes a step towards better understanding the trade-off therein and suggests a practical method to mitigate the potential information leakage in such high-stakes scenarios. That being said, the adversarial learning techniques might inevitably lead to degradation in target performance, and more work is needed to explore the best trade-off that could be achieved.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318. ACM, 2016.

[2] Brett K Beaulieu-Jones, Zhiwei Steven Wu, Chris Williams, Ran Lee, Sanjeev P Bhavnani, James Brian Byrd, and Casey S Greene. Privacy-preserving generative deep neural networks support clinical data sharing. *BioRxiv*, page 159756, 2018.

[3] Brett K Beaulieu-Jones, William Yuan, Samuel G Finlayson, and Zhiwei Steven Wu. Privacy-preserving distributed deep learning for clinical data. *arXiv preprint arXiv:1812.01484*, 2018.

[4] Martin Bertran, Natalia Martinez, Afroditi Papadaki, Qiang Qiu, Miguel Rodrigues, Galen Reeves, and Guillermo Sapiro. Adversarially learned representations for information obfuscation and inference. In *International Conference on Machine Learning*, pages 614–623, 2019.

[5] Chris Calabro. *The exponential complexity of satisfiability problems*. PhD thesis, UC San Diego, 2009.

[6] Edward Chou, Josh Beal, Daniel Levy, Serena Yeung, Albert Haque, and Li Fei-Fei. Faster cryptonets: Leveraging sparsity for real-world encrypted inference. *arXiv preprint arXiv:1811.09953*, 2018.

[7] Constantinos Daskalakis and Ioannis Panageas. The limit points of (optimistic) gradient descent in min-max optimization. In *Advances in Neural Information Processing Systems*, pages 9236–9246, 2018.

[8] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

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

[10] 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*, pages 214–226. ACM, 2012.

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

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

[13] Dominik Maria Endres and Johannes E Schindelin. A new metric for probability distributions. *IEEE Transactions on Information theory*, 2003.

[14] Miro Enev, Jaeyeon Jung, Liefeng Bo, Xiaofeng Ren, and Tadayoshi Kohno. Sensorsift: balancing sensor data privacy and utility in automated face understanding. In *Proceedings of the 28th Annual Computer Security Applications Conference*, pages 149–158. ACM, 2012.

[15] EU. General Data Protection Regulation. `https://en.wikipedia.org/wiki/General_Data_Protection_Regulation`, 2018.

[16] Clément Feutry, Pablo Piantanida, Yoshua Bengio, and Pierre Duhamel. Learning anonymized representations with adversarial neural networks. *arXiv preprint arXiv:1802.09386*, 2018.

[17] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[18] Ran Gilad-Bachrach, Nathan Dowlin, Kim Laine, Kristin Lauter, Michael Naehrig, and John Wernsing. Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy. In *International Conference on Machine Learning*, pages 201–210, 2016.

[19] Jihun Hamm. Minimax filter: Learning to preserve privacy from inference attacks. *The Journal of Machine Learning Research*, 18(1):4704–4734, 2017.

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

[21] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V Pearson, Dietrich A Stephan, Stanley F Nelson, and David W Craig. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS genetics*, 4(8), 2008.

[22] Yitong Li, Timothy Baldwin, and Trevor Cohn. Towards robust and privacy-preserving text representations. *arXiv preprint arXiv:1805.06093*, 2018.

[23] Jianhua Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

[24] Seyed Ali Osia, Ali Taheri, Ali Shahin Shamsabadi, Kleomenis Katevas, Hamed Haddadi, and Hamid R Rabiee. Deep private-feature extraction. *IEEE Transactions on Knowledge and Data Engineering*, 32(1):54–66, 2018.

[25] Seyed Ali Osia, Ali Shahin Shamsabadi, Sina Sajadmanesh, Ali Taheri, Kleomenis Katevas, Hamid R Rabiee, Nicholas D Lane, and Hamed Haddadi. A hybrid deep learning architecture for privacy-preserving mobile analytics. *IEEE Internet of Things Journal*, 2020.

[26] Nicolas Papernot, Martín Abadi, Ulfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. *arXiv preprint arXiv:1610.05755*, 2016.

[27] Mauro Ribeiro, Katarina Grolinger, and Miriam AM Capretz. Mlaas: Machine learning as a service. In *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, pages 896–902. IEEE, 2015.

[28] Proteek Chandan Roy and Vishnu Naresh Boddeti. Mitigating information leakage in image representations: A maximum entropy approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2586–2594, 2019.

[29] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2017.

[30] Brij Mohan Lal Srivastava, Aurélien Bellet, Marc Tommasi, and Emmanuel Vincent. Privacy-preserving adversarial representation learning in asr: Reality or illusion? *arXiv preprint arXiv:1911.04913*, 2019.

[31] Ji Wang, Jianguo Zhang, Weidong Bao, Xiaomin Zhu, Bokai Cao, and Philip S Yu. Not just privacy: Improving performance of private deep learning in mobile cloud. In *KDD*, 2018.

[32] Ye Wang, Yuksel Ozan Basciftci, and Prakash Ishwar. Privacy-utility tradeoffs under constrained data release mechanisms. *arXiv preprint arXiv:1710.09295*, 2017.

[33] Jacob Whitehill and Javier Movellan. Discriminately decreasing discriminability with learned image filters. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2488–2495. IEEE, 2012.

[34] Zhenyu Wu, Zhangyang Wang, Zhaowen Wang, and Hailin Jin. Towards privacy-preserving visual recognition via adversarial training: A pilot study. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 606–624, 2018.

[35] Mohammad Yaghini, Bogdan Kulynych, and Carmela Troncoso. Disparate vulnerability: On the unfairness of privacy attacks against machine learning. *arXiv preprint arXiv:1906.00389*, 2019.

[36] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pages 268–282. IEEE, 2018.

[37] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[38] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5810–5818, 2017.

[39] Han Zhao and Geoffrey J Gordon. Inherent tradeoffs in learning fair representations. In *Advances in neural information processing systems*, 2019.

[40] Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *International Conference on Machine Learning*, pages 7523–7532, 2019.

Appendix

In this appendix we provide the missing proofs of theorems and claims in our main paper. We also describe detailed experimental settings here.

A Technical Tools

In this section we list the lemmas and theorems used during our proof.

Lemma A.1 (Theorem 2.2, [5])**.** Let $H_2^{-1}(s)$ be the inverse binary entropy function for $s \in [0,1]$, then $H_2^{-1}(s) \geq s/2\lg(6/s)$.

Lemma A.2 (Lin [23])**.** Let $\mathcal{D}$ and $\mathcal{D}'$ be two distributions, then $D_{\text{JS}}(\mathcal{D}, \mathcal{D}') \leq \frac{1}{2}\|\mathcal{D} - \mathcal{D}'\|_1$.

Theorem A.1 (Data processing inequality)**.** Let $X \perp Y \mid Z$, then $I(X;Z) \geq I(X;Y)$.

B Missing Proofs

Proposition 2.1. Let $h_A : \mathcal{X} \to \{0,1\}$ be a hypothesis, then $\textsc{Adv}(h_A) = 0$ iff $I(h_A(X);A) = 0$ and $\textsc{Adv}(h_A) = 1$ iff $h_A(X) = A$ almost surely or $h_A(X) = 1 - A$ almost surely.

Proof. We first prove the first part of the proposition. By definition, $\textsc{Adv}(h_A) = 0$ iff $\Pr_{\mathcal{D}}(h_A(X) = 1 \mid A = 1) = \Pr_{\mathcal{D}}(h_A(X) = 1 \mid A = 0)$, which is also equivalent to $h_A(X) \perp\!\!\!\perp A$. It then follows that $h_A(X) \perp\!\!\!\perp A \Leftrightarrow I(h_A(X);A) = 0$.

For the second part of the proposition, again, by definition of $\textsc{Adv}(h_A)$, it is clear to see that we either have $\Pr_{\mathcal{D}}(h_A(X) = 1 \mid A = 1) = 1$ and $\Pr_{\mathcal{D}}(h_A(X) = 1 \mid A = 0) = 0$, or $\Pr_{\mathcal{D}}(h_A(X) = 1 \mid A = 1) = 0$ and $\Pr_{\mathcal{D}}(h_A(X) = 1 \mid A = 0) = 1$. Hence we discuss by these two cases. For ease of notation, we omit the subscript $\mathcal{D}$ from $\Pr_{\mathcal{D}}$ when it is obvious from the context which probability distribution we are referring to.

1. If $\Pr(h(X) = 1 \mid A = 1) = 1$ and $\Pr(h(X) = 1 \mid A = 0) = 0$, then we know that:
$$\begin{aligned}\Pr(h_A(X) \neq A) &= \Pr(A = 0)\Pr(h_A(X) \neq A \mid A = 0) + \Pr(A = 1)\Pr(h_A(X) \neq A \mid A = 1)\\ &= \Pr(A = 0)\Pr(h_A(X) = 1 \mid A = 0) + \Pr(A = 1)\Pr(h_A(X) = 0 \mid A = 1)\\ &= \Pr(A = 0)\cdot 0 + \Pr(A = 1)\cdot 0\\ &= 0.\end{aligned}$$
2. If $\Pr(h_A(X) = 1 \mid A = 1) = 0$ and $\Pr(h_A(X) = 1 \mid A = 0) = 1$, similarly, we have:
$$\begin{aligned}\Pr(h_A(X) \neq 1 - A) &= \Pr(A = 0)\Pr(h_A(X) \neq 1 - A \mid A = 0) + \Pr(A = 1)\Pr(h_A(X) \neq 1 - A \mid A = 1)\\ &= \Pr(A = 0)\Pr(h_A(X) = 0 \mid A = 0) + \Pr(A = 1)\Pr(h_A(X) = 1 \mid A = 1)\\ &= \Pr(A = 0)\cdot 0 + \Pr(A = 1)\cdot 0\\ &= 0.\end{aligned}$$

Combining the above two parts completes the proof. ■

Proposition 2.2. If $\mathcal{H}_A$ is symmetric, then $\textsc{Adv}(\mathcal{H}_A) + \min_{h_A \in \mathcal{H}_A} \Pr(h_A(X) = 0 \mid A = 1) + \Pr(h_A(X) = 1 \mid A = 0) = 1$.

Proof. By definition, we have:
$$\begin{aligned}1 - \textsc{Adv}(\mathcal{H}_A) &:= 1 - \max_{h_A \in \mathcal{H}_A} \textsc{Adv}(h_A)\\ &= \min_{h_A \in \mathcal{H}_A} 1 - \left|\Pr(h_A(X) = 1 \mid A = 1) - \Pr(h_A(X) = 1 \mid A = 0)\right|\\ &= \min_{h_A \in \mathcal{H}_A} 1 - \left(\Pr(h_A(X) = 1 \mid A = 1) - \Pr(h_A(X) = 1 \mid A = 0)\right)\\ &= \min_{h \in \mathcal{H}} \Pr(h_A(X) = 0 \mid A = 1) + \Pr(h_A(X) = 1 \mid A = 0),\end{aligned}$$

where the third equality holds due to the fact that $\max_{h_A \in \mathcal{H}_A} \left| \Pr(h_A(X) = 1 \mid A = 1) - \Pr(h_A(X) = 1 \mid A = 0) \right| = \max_{h_A \in \mathcal{H}_A} \big(\Pr(h_A(X) = 1 \mid A = 1) - \Pr(h_A(X) = 1 \mid A = 0) \big)$. To see this, for any specific h_A such that the term inside the absolute value is negative, we can find $1 - h_A \in \mathcal{H}_A$ such that it becomes positive, due to the assumption that $\mathcal{H}_A$ is symmetric. ■

Proposition 3.1. Let $f : \mathcal{X} \to \mathcal{Z}$ be a deterministic function and $\mathcal{H}_A \subseteq 2^{\mathcal{Z}}$ be a hypothesis class over $\mathcal{Z}$. For any joint distribution $\mathcal{D}$ over X, A, Y, if $I(f(X); A) = 0$, then $\textsc{Adv}(\mathcal{H}_A \circ f) = 0$.

Proof. First, by the celebrated data-processing inequality, $\forall h_A \in \mathcal{H}_A$:

$$0 \leq I(h_A(f(X)); A) \leq I(f(X); A) = 0.$$

By Proposition 2.1, this means that $\forall h_A \in \mathcal{H}_A$, $\textsc{Adv}(h_A) = 0$, which further implies that $\textsc{Adv}(\mathcal{H}_A \circ f) = 0$ by definition. ■

Lemma 3.1. For any feature map $f : \mathcal{X} \to \mathcal{Z}$, assume that $\mathcal{H}$ contains all the probabilistic classifiers, then $\min_{h \in \mathcal{H}} \mathrm{CE}_Y(h \circ f) = H(Y \mid Z)$ and $h^*(Z) := \arg\min_{h \in \mathcal{H}} \mathrm{CE}_Y(h \circ f) = \Pr(Y = 1 \mid Z = f(X))$.

Proof. Let $\mathcal{D}^f$ be the induced (pushforward) distribution of $\mathcal{D}$ under the map $f : \mathcal{X} \to \mathcal{Z}$. By the definition of cross-entropy loss, we have:

$$\begin{aligned}
\mathrm{CE}_Y(h \circ f) &= -\mathbb{E}_{\mathcal{D}} \left[\mathbb{I}(Y = 0) \log(1 - h(f(X))) + \mathbb{I}(Y = 1) \log(h(f(X))) \right] \\
&= -\mathbb{E}_{\mathcal{D}^f} \left[\mathbb{I}(Y = 0) \log(1 - h(Z)) + \mathbb{I}(Y = 1) \log(h(Z)) \right] \\
&= -\mathbb{E}_Z \mathbb{E}_{Y|Z} \left[\mathbb{I}(Y = 0) \log(1 - h(Z)) + \mathbb{I}(Y = 1) \log(h(Z)) \right] \\
&= -\mathbb{E}_Z \left[\Pr(Y = 0 \mid Z) \log(1 - h(Z)) + \Pr(Y = 1 \mid Z) \log(h(Z)) \right] \\
&= \mathbb{E}_Z \left[D_{\mathrm{KL}}(\Pr(Y \mid Z) \,||\, h(Z)) \right] + H(Y \mid Z) \\
&\geq H(Y \mid Z).
\end{aligned}$$

It is also clear from the above proof that the minimum value of the cross-entropy loss is achieved when $h(Z)$ equals the conditional probability $\Pr(Y = 1 \mid Z)$, i.e., $h^*(Z) = \Pr(Y = 1 \mid Z = f(X))$. ■

Theorem 3.1. Let f^* be the optimal feature map of (3) and define $H^* := H(A \mid Z = f^*(X))$. Then for any adversary $\widehat{A}$ such that $I(\widehat{A}; A \mid Z) = 0$, $\Pr_{\mathcal{D}^{f^*}}(\widehat{A} \neq A) \geq H^* / 2 \lg(6 / H^*)$.

Proof. To prove this theorem, let E be the binary random variable that takes value 1 iff $A \neq \widehat{A}$, i.e., $E = \mathbb{I}(A \neq \widehat{A})$. Now consider the joint entropy of $A, \widehat{A}$ and E. On one hand, we have:

$$H(A, \widehat{A}, E) = H(A, \widehat{A}) + H(E \mid A, \widehat{A}) = H(A, \widehat{A}) + 0 = H(A \mid \widehat{A}) + H(\widehat{A}).$$

Note that the second equation holds because E is a deterministic function of A and $\widehat{A}$, that is, once A and $\widehat{A}$ are known, E is also known, hence $H(E \mid A, \widehat{A}) = 0$. On the other hand, we can also decompose $H(A, \widehat{A}, E)$ as follows:

$$H(A, \widehat{A}, E) = H(\widehat{A}) + H(A \mid \widehat{A}, E) + H(E \mid \widehat{A}).$$

Combining the above two equalities yields

$$H(A \mid \widehat{A}, E) + H(E \mid \widehat{A}) = H(A \mid \widehat{A}).$$

Furthermore, since conditioning cannot increase entropy, we have $H(E \mid \widehat{A}) \leq H(E)$, which further implies

$$H(A \mid \widehat{A}) \leq H(E) + H(A \mid \widehat{A}, E).$$

Now consider $H(A \mid \widehat{A}, E)$. Since $A \in \{0, 1\}$, by definition of the conditional entropy, we have:

$$H(A \mid \widehat{A}, E) = \Pr(E = 1) H(A \mid \widehat{A}, E = 1) + \Pr(E = 0) H(A \mid \widehat{A}, E = 0) = 0 + 0 = 0.$$

To lower bound $H(A \mid \widehat{A})$, realize that

$$I(A; \widehat{A}) + H(A \mid \widehat{A}) = H(A) = I(A; Z) + H(A \mid Z).$$

Since $\widehat{A}$ is a randomized function of Z such that $A \perp \widehat{A} \mid Z$, due to the celebrated data-processing inequality, we have $I(A;\widehat{A}) \le I(A;Z)$, which implies

$$H(A \mid \widehat{A}) \ge H(A \mid Z).$$

Combine everything above, we have the following chain of inequalities hold:

$$H(A \mid Z) \le H(A \mid \widehat{A}) \le H(E) + H(A \mid \widehat{A}, E) = H(E),$$

which implies

$$\Pr_{\mathcal{D}^{f^*}}(A \neq \widehat{A}) = \Pr_{\mathcal{D}^{f^*}}(E = 1) \ge H_2^{-1}(H(A \mid Z)),$$

where $H_2^{-1}(\cdot)$ is the inverse function of the binary entropy $H(t) := -t\log t - (1-t)\log(1-t)$ when $t \in [0,1]$. To conclude the proof, we apply Lemma A.1 to further lower bound the inverse binary entropy function by

$$H_2^{-1}(H(A \mid Z)) \ge H(A \mid Z)/2\lg(6/H(A \mid Z)),$$

completing the proof. ■

Lemma 3.2. Let $\mathcal{D}_0$ and $\mathcal{D}_1$ be two distributions over $\mathcal{X}$ and let $\mathcal{D}_0^f$ and $\mathcal{D}_1^f$ be the induced distributions of $\mathcal{D}_0$ and $\mathcal{D}_1$ over $\mathcal{Z}$ by function f, then $d_{\mathrm{JS}}(\mathcal{D}_0^f, \mathcal{D}_1^f) \le d_{\mathrm{JS}}(\mathcal{D}_0, \mathcal{D}_1)$.

Proof. Let B be a uniform random variable taking value in $\{0,1\}$ and let the random variable Z_B with distribution $\mathcal{D}_B^f$ (resp. X_B with distribution $\mathcal{D}_B$) be the mixture of $\mathcal{D}_0^f$ and $\mathcal{D}_1^f$ (resp. $\mathcal{D}_0$ and $\mathcal{D}_1$) according to B. It is easy to see that $\mathcal{D}_B = (\mathcal{D}_0 + \mathcal{D}_1)/2$, and we have:

$$\begin{aligned}
I(B;X_B) &= H(X_B) - H(X_B \mid B) \\
&= -\sum \mathcal{D}_B \log \mathcal{D}_B + \frac{1}{2}\left(\sum \mathcal{D}_0 \log \mathcal{D}_0 + \sum \mathcal{D}_1 \log \mathcal{D}_1\right) \\
&= -\frac{1}{2}\sum \mathcal{D}_0 \log \mathcal{D}_B - \frac{1}{2}\sum \mathcal{D}_1 \log \mathcal{D}_B + \frac{1}{2}\left(\sum \mathcal{D}_0 \log \mathcal{D}_0 + \sum \mathcal{D}_1 \log \mathcal{D}_1\right) \\
&= \frac{1}{2}\sum \mathcal{D}_0 \log \frac{\mathcal{D}_0}{\mathcal{D}_B} + \frac{1}{2}\sum \mathcal{D}_1 \log \frac{\mathcal{D}_1}{\mathcal{D}_B} \\
&= \frac{1}{2}D_{\mathrm{KL}}(\mathcal{D}_0 \parallel \mathcal{D}_B) + \frac{1}{2}D_{\mathrm{KL}}(\mathcal{D}_1 \parallel \mathcal{D}_B) \\
&= D_{\mathrm{JS}}(\mathcal{D}_0, \mathcal{D}_1).
\end{aligned}$$

Similarly, we have:

$$D_{\mathrm{JS}}(\mathcal{D}_0^f, \mathcal{D}_1^f) = I(B;Z_B).$$

Since $\mathcal{D}_0^f$ (resp. $\mathcal{D}_1^f$) is induced by f from $\mathcal{D}_0$ (resp. $\mathcal{D}_1$), by linearity, $\mathcal{D}_B^f$ is also induced by f from $\mathcal{D}_B$. Hence $Z_B = f(X_B)$ and the following Markov chain holds:

$$B \to X_B \to Z_B.$$

Apply the data processing inequality, we have

$$D_{\mathrm{JS}}(\mathcal{D}_0, \mathcal{D}_1) = I(B;X_B) \ge I(B;Z_B) = D_{\mathrm{JS}}(\mathcal{D}_0^f, \mathcal{D}_1^f).$$

Taking square root on both sides of the above inequality completes the proof. ■

Lemma 3.3. Let $\hat{Y} = h(f(X)) \in \{0,1\}$ be the predictor, then for $a \in \{0,1\}$, $d_{\mathrm{JS}}(\mathcal{D}_a^Y, \mathcal{D}_a^{h\circ f}) \le \sqrt{\mathrm{Err}_a(h\circ f)}$.

Proof. For $a \in \{0,1\}$, by definition of the JS distance:

$$\begin{aligned}
d_{\mathrm{JS}}^2(\mathcal{D}_a^Y, \mathcal{D}_a^{h\circ f}) &= D_{\mathrm{JS}}(\mathcal{D}_a^Y, \mathcal{D}_a^{h\circ f}) \\
&\le \|\mathcal{D}_a^Y - \mathcal{D}_a^{h\circ f}\|_1/2 && \text{(Lemma A.2)} \\
&= \big(|\Pr(Y=0 \mid A=a) - \Pr(h(f(X))=0 \mid A=a)| \\
&\qquad + |\Pr(Y=1 \mid A=a) - \Pr(h(f(X))=1 \mid A=a)|\big)/2 \\
&= |\Pr(Y=1 \mid A=a) - \Pr(h(f(X))=1 \mid A=a)| \\
&= |\mathbb{E}[Y \mid A=a] - \mathbb{E}[h(f(X)) \mid A=a]| \\
&\le \mathbb{E}[|Y - h(f(X))| \mid A=a] \\
&= \mathrm{Err}_a(h\circ f),
\end{aligned}$$

where the expectation is taken over the joint distribution of X, Y. Taking square root at both sides then completes the proof. ■

Theorem 3.2. Let $\mathcal{H}_A \subseteq 2^{\mathcal{Z}}$ be the hypothesis space of all the classifiers from $\mathcal{Z}$ to $\{0,1\}$. Assume the conditions in Lemma 3.4 hold, then $\forall h \in \mathcal{H}$, $\textsc{Acc}_0(h\circ f) + \textsc{Acc}_1(h\circ f) \le 2 - \frac{1}{3} D_{\mathrm{JS}}(\mathcal{D}_0^Y, \mathcal{D}_1^Y) + \textsc{Adv}(\mathcal{H}_A \circ f)$.

Proof. Before we delve into the details, we first give a high-level sketch of the main idea. The proof could be basically partitioned into two parts. In the first part, we will show that when $\mathcal{H}_A$ contains all the measurable prediction functions, $\textsc{Adv}(\mathcal{H}_A \circ f)$ could be used to upper bound $D_{\mathrm{JS}}(\mathcal{D}_0^f, \mathcal{D}_1^f)$. The second part combines Lemma 3.3 and Lemma 3.2 to complete the proof.

In this part we first show that $D_{\mathrm{JS}}(\mathcal{D}_0^f, \mathcal{D}_1^f) \le \textsc{Adv}(\mathcal{H}\circ f)$:

$$\begin{aligned}
D_{\mathrm{JS}}(\mathcal{D}_0^f, \mathcal{D}_1^f) &\le \frac{1}{2}\|\mathcal{D}_0^f - \mathcal{D}_1^f\|_1 \\
&= d_{\mathrm{TV}}(\mathcal{D}_0^f, \mathcal{D}_1^f) \\
&= \sup_{A\in\mathscr{B}} |\mathcal{D}_0^f(A) - \mathcal{D}_1^f(A)|,
\end{aligned}$$

where $d_{\mathrm{TV}}(\cdot,\cdot)$ denotes the total variation distance and $\mathscr{B}$ is the sigma algebra that contains all the measurable subsets of $\mathcal{Z}$. On the other hand, when $\mathcal{H}_A$ contains all the measurable functions in $2^{\mathcal{Z}}$, we have:

$$\begin{aligned}
\textsc{Adv}(\mathcal{H}_A \circ f) &= \max_{h_A\in\mathcal{H}_A} |\Pr(h_A(Z)=1 \mid A=0) - \Pr(h_A(Z)=1 \mid A=1)| \\
&= \max_{h_A\in\mathcal{H}_A} |\mathcal{D}_0(h_A^{-1}(1)) - \mathcal{D}_1(h_A^{-1}(1))| \\
&= \sup_{A\in\mathscr{B}} |\mathcal{D}_0^f(A) - \mathcal{D}_1^f(A)|,
\end{aligned}$$

where the last equality follows from the fact that $\mathcal{H}_A$ is complete and contains all the measurable functions. Combine the above two parts we immediately have $D_{\mathrm{JS}}(\mathcal{D}_0^f, \mathcal{D}_1^f) \le \textsc{Adv}(\mathcal{H}_A \circ f)$.

Now using the key lemma, we have:

$$\begin{aligned}
d_{\mathrm{JS}}(\mathcal{D}_0^Y, \mathcal{D}_1^Y) &\le d_{\mathrm{JS}}(\mathcal{D}_0^Y, \mathcal{D}_0^{h\circ f}) + d_{\mathrm{JS}}(\mathcal{D}_0^f, \mathcal{D}_1^f) + d_{\mathrm{JS}}(\mathcal{D}_1^{h\circ f}, \mathcal{D}_1^Y) \\
&\le \sqrt{\mathrm{Err}_0(h\circ f)} + \sqrt{\textsc{Adv}(\mathcal{H}_A\circ f)} + \sqrt{\mathrm{Err}_1(h\circ f)} \\
&= \sqrt{1 - \textsc{Acc}_0(h\circ f)} + \sqrt{\textsc{Adv}(\mathcal{H}_A\circ f)} + \sqrt{1 - \textsc{Acc}_1(h\circ f)} \\
&\le \sqrt{3(1 - \textsc{Acc}_0(h\circ f) + 1 - \textsc{Acc}_1(h\circ f) + \textsc{Adv}(\mathcal{H}_A\circ f))} \\
&= \sqrt{3\big(2 - (\textsc{Acc}_0(h\circ f) + \textsc{Acc}_1(h\circ f) - \textsc{Adv}(\mathcal{H}_A\circ f))\big)}.
\end{aligned}$$

Taking square at both sides and then rearrange the terms then completes the proof. ■

Theorem 3.3. Assume the conditions in Theorem 3.2 hold. If $\textsc{Adv}(\mathcal{H}_A \circ f) \leq D_{\text{JS}}(\mathcal{D}_0^Y, \mathcal{D}_1^Y)$, then $\forall h \in \mathcal{H}$, $\text{Err}_0(h \circ f) + \text{Err}_1(h \circ f) \geq \frac{1}{2}\big(d_{\text{JS}}(\mathcal{D}_0^Y, \mathcal{D}_1^Y) - \sqrt{\textsc{Adv}(\mathcal{H}_A \circ f)}\big)^2$.

Proof. Similarly, using the key lemma, we have:

$$\begin{aligned} d_{\text{JS}}(\mathcal{D}_0^Y, \mathcal{D}_1^Y) &\leq d_{\text{JS}}(\mathcal{D}_0^Y, \mathcal{D}_0^{h\circ f}) + d_{\text{JS}}(\mathcal{D}_0, \mathcal{D}_1) + d_{\text{JS}}(\mathcal{D}_1^{h\circ f}, \mathcal{D}_1^Y) \\ &\leq \sqrt{\text{Err}_0(h \circ f)} + \sqrt{\textsc{Adv}(\mathcal{H}_A \circ f)} + \sqrt{\text{Err}_1(h \circ f)} \end{aligned}$$

Under the assumption that $\textsc{Adv}(\mathcal{H}_A \circ f) \leq D_{\text{JS}}(\mathcal{D}_0^Y, \mathcal{D}_1^Y)$, we have $d_{\text{JS}}(\mathcal{D}_0^Y, \mathcal{D}_1^Y) \geq \sqrt{\textsc{Adv}(\mathcal{H}_A \circ f)}$, hence by AM-GM inequality:

$$\sqrt{2\big(\text{Err}_0(h \circ f) + \text{Err}_1(h \circ f)\big)} \geq \sqrt{\text{Err}_0(h \circ f) + \text{Err}_1(h \circ f)} \geq d_{\text{JS}}(\mathcal{D}_0^Y, \mathcal{D}_1^Y) - \sqrt{\textsc{Adv}(\mathcal{H}_A \circ f)}.$$

Taking square at both sides then completes the proof. ■

C Detailed Experiments

In this section, we provide more details of the experiments. First we provide the details of different existing methods we evaluate. Then we elaborate more dataset description, model architecture and training parameters in different experiments.

C.1 Details on Methods

We provide a detailed description of each method here:

1). Privacy Partial Least Squares (PPLS): It learns $n \times X_d$ matrix for the feature transformation. The matrix is learned by maximizing the covariance of the learned representation and target attribute while minimizing the covariance of the learned representation and sensitive attribute.

2). Privacy Linear Discriminant Analysis (PLDA): It learns $n \times X_d$ matrix for the feature transformation. The matrix is learned by maximizing the Fisher's linear discriminability of the learned representation and target attribute while minimizing the Fisher's linear discriminability of the learned representation and sensitive attribute.

3). Minimax filter with alternative update (ALT-UP): The representation is learned via optimizing Equation 2 in an alternative way, first we update the parameters of the feature transformation module and the target attribute classifier, and then accordingly update the sensitive attribute classifier.

4). Maximum Entropy Adversarial Representation Learning (MAX-ENT): The objective equation is the slightly different from ALT-UP. The latter term contains additional entropy term to maximize unpredictability of the sensitive attribute.

5). Gradient Reversal Layer (GRL): The objective equation is the same as ALT-UP, and we train the feature transformation module by adding a gradient reversal layer between the feature transformation module and the sensitive attribute classifier.

6). Principal Component Analysis (PCA): It generates a $n \times X_d$ matrix for the feature transformation where the rows of the matrix are the n largest eigenvectors of the input dataset X.

7). Normal Training (NORM-TRAIN): It is equivalent to normal training by setting $\lambda = 0$ in Equation 2.

8). Local Differential Privacy (LDP): Standard Laplace mechanism of local differential privacy, where the noise is added to the raw representation for erasing the information of the sensitive attribute.

9). Differentially private SGD (DPSGD): It is one of the state-of-the-art differential privacy methods on deep learning. It adds Gaussian noise to the gradients when training the model.

C.2 Details on UCI Adult Dataset Evaluation

UCI Adult dataset is a benchmark machine learning dataset for income prediction. Each data record contains 14 categorical or numerical attributes, such as occupation, education and gender, to predict

whether individual annual income exceeds \$50K/year. The dataset is divided into training set (24130 examples), validation (6032 examples), and test set (15060 examples). We choose gender, age, and education as the sensitive attributes, respectively.

Table 1: Data distribution of income (Y) and gender (A) in UCI Adult dataset.

	$Y=0$	$Y=1$
$A=0$	20988	9539
$A=1$	13026	1669

Table 2: Data distribution of income (Y) and age (A) in UCI Adult dataset.

	$Y=0$	$Y=1$
$A=0$	18042	2473
$A=1$	15972	8735

Table 3: Data distribution of income (Y) and education (A) in UCI Adult dataset.

	$Y=0$	$Y=1$
$A=0$	20447	4248
$A=1$	13567	6960

We process each sensitive attribute as binary label for each experiment: for age label, 0 if the person is no greater than 35 years old and 1 otherwise; for education label, 0 if the person has not entered college or receive higher education than college, and 1 otherwise. In the mean time, we also remove corresponding sensitive attribute from the input, so the dimension of input data for each experiment is different. The input dimensions for income-gender experiment, income-age experiment, and income-education experiment are 113, 104 and 99, respectively. Table 1, Table 2 and Table 3 summarize the data distribution of UCI Adult dataset for protecting different sensitive attributes.

We use the two-layer ReLU-based neural net for f and one-layer neural net for h. The output dimensions of f are 64. We train all methods using SGD with the initial learning late 0.001 and momentum 0.9 for 40 epochs. In the DP-SGD experiment, we set the noise multiplier as 0.45 and 4.0 for small noise and large noise, respectively, and set the clipping norm as 1.0. (ϵ, δ) for DPSGD small noise and DPSGD large noise are $(33.7, 10^{-5})$ and $(0.572, 10^{-5})$, respectively. Among all methods, we report the one achieving the best performance on the target task in the validation set. We run the experiments for ten random seeds and compute the average.

C.3 Details on UTKFace Dataset Evaluation

UTKFace dataset is a large scale face dataset with annotations of age (range from 0 to 116 years old), gender (male and female), and ethnicity (White, Black, Asian, Indian, and Others). It contains 23,705 64×64 aligned and cropped RGB face images and we split the dataset into training set (15171 examples), validation set (3793 examples) and test set (4741 examples), respectively. We further process age label and ethnicity label as binary labels: 0 if the person is not greater than 35 years old for age label (is white for ethnicity label), and 1 if the the person is greater than 35 years old for age label (is non-white for ethnicity label). Table 4 and Table 5 summarize the data distribution of UTKFace dataset for protecting different sensitive attributes.

Table 4: Data distribution of gender (Y) and race (A) in UTKFace dataset.

	$Y=0$	$Y=1$
$A=0$	5477	4601
$A=1$	6914	6713

Table 5: Data distribution of gender (Y) and age (A) in UTKFace dataset.

	$Y=0$	$Y=1$
$A=0$	6889	8218
$A=1$	5502	3096

Since NORM-TRAIN, ALT-UP, GRL and DP can directly enjoy the benefits of using the state-of-the-art neural network architecture as feature extraction module, so we use the feature extraction module of Wide Residual Network [37] for the (non-linear) feature transformation module, while

PPLS, PLDA, and PCA learn 12288×2048 matrix filter for f. We train all methods using SGD with the initial learning late 0.01 and momentum 0.9 for 30 epochs. The learning rate is decayed by a factor of 0.1 for every 10 epochs. In the DP-SGD experiment, we set the noise multiplier as 0.45 and 1.0 for small noise and large noise, respectively, and set the clipping norm as 1.0. (ϵ, δ) for DPSGD small noise and DPSGD large noise are $(25.7, 10^{-5})$ and $(2.7, 10^{-5})$, respectively. Among all methods, we report the one achieving the best performance on the target task in the validation set. We run the experiments for ten times and compute the average.

D Additional Experimental Results

In this section, we present additional experimental results to gain more insights into how the trade-off parameter λ affects the performances of different adversarial presentation learning methods. We varies the values of λ and report the accuracies of both tasks using the Adult dataset when the sensitive attribute is gender. Note that all hyperparameter settings follow the previous experiments. The results are shown in Table 6. We can see that the overall trend is that when λ increases, the accuracies for both tasks decrease. Compared to ALT-UP and GRL, the training of MAX-ENT is unstable when λ is large.

Table 6: Performances of different adversarial representation learning methods when λ changes.

Gender	ALT-UP	λ	0	0.1	1	5
		TAR. ACC.	0.8501±0.0010	0.8496±0.0013	0.8483±0.0010	0.8456±0.0014
		SEN. ACC.	0.7408±0.0096	0.6682±0.0026	0.6627±0.0021	0.6737±0.0005
	GRL	λ	0	0.1	1	5
		TAR. ACC.	0.8501±0.0010	0.8465±0.0017	0.8449±0.0010	0.8387±0.0019
		SEN. ACC.	0.7408±0.0096	0.6677±0.0060	0.6677±0.0039	0.6764±0.0054
	MAX-ENT	λ	0	0.1	1	5
		TAR. ACC.	0.8501±0.0010	0.8450±0.0038	0.8411±0.0055	0.7891±0.0449
		SEN. ACC.	0.7408±0.0096	0.6928±0.0084	0.6897±0.0038	0.5695±0.1679
Age	ALT-UP	λ	0	0.1	1	5
		TAR. ACC.	0.8467±0.0011	0.8468±0.0009	0.8472±0.0011	0.8451±0.0008
		SEN. ACC.	0.7190±0.010	0.6516±0.0038	0.5422±0.0133	0.5573±0.0438
	GRL	λ	0	0.1	1	5
		TAR. ACC.	0.8467±0.0011	0.8444±0.0009	0.8445±0.0012	0.8422±0.0013
		SEN. ACC.	0.7190±0.010	0.6486±0.0067	0.5361±0.0134	0.5381±0.0133
	MAX-ENT	λ	0	0.1	1	5
		TAR. ACC.	0.8467±0.0011	0.8379±0.0056	0.8194±0.0345	0.7795±0.0406
		SEN. ACC.	0.7190±0.0100	0.6633±0.0669	0.6201±0.0820	0.5400±0.0316
Education	ALT-UP	λ	0	0.1	1	5
		TAR. ACC.	0.8494±0.0008	0.8498±0.0004	0.8497±0.0012	0.8494±0.0015
		SEN. ACC.	0.7088±0.0080	0.6062±0.0108	0.6044±0.0145	0.5462±0.0358
	GRL	λ	0	0.1	1	5
		TAR. ACC.	0.8494±0.0008	0.8525±0.0010	0.8518±0.0007	0.8500±0.0013
		SEN. ACC.	0.7088±0.0080	0.6082±0.0119	0.6015±0.0154	0.5528±0.0260
	MAX-ENT	λ	0	0.1	1	5
		TAR. ACC.	0.8494±0.0008	0.8365±0.0033	0.8253±0.0376	0.8087±0.0468
		SEN. ACC.	0.7088±0.0080	0.5790±0.0383	0.5484±0.0001	0.5386±0.0305