Generating Fair Universal Representations Using Adversarial Models

Peter Kairouz, Jiachun Liao, Chong Huang, Maunil Vyas, Monica Welfert, and Lalitha Sankar, *Senior Member, IEEE*

***Abstract*— We present a data-driven framework for learning *fair universal representations* (FUR) that guarantee statistical fairness for any learning task that may not be known *a priori*. Our framework leverages recent advances in adversarial learning to allow a data holder to learn representations in which a set of sensitive attributes are decoupled from the rest of the dataset. We formulate this as a constrained minimax game between an encoder and an adversary where the constraint ensures a measure of usefulness (utility) of the representation. The resulting problem is that of censoring, i.e., finding a representation that is least informative about the sensitive attributes given a utility constraint. For appropriately chosen adversarial loss functions, our censoring framework precisely clarifies the optimal adversarial strategy against strong information-theoretic adversaries; it also achieves the fairness measure of demographic parity for the resulting constrained representations. We evaluate the performance of our proposed framework on both synthetic and publicly available datasets. For these datasets, we use two tradeoff measures: censoring vs. representation fidelity and fairness vs. utility for downstream tasks, to amply demonstrate that multiple sensitive features can be effectively censored even as the resulting fair representations ensure accuracy for multiple downstream tasks.**

***Index Terms*— Fair universal representations, algorithmic fairness, generative adversarial networks, minimax games.**

I. Introduction

The use of data-driven machine learning (ML) has recently seen unprecedented success in a variety of automated decision-making systems including facial recognition, natural language processing, mortgage lending, and even parole prediction. The success of these approaches hinges on the availability of large datasets that often include sensitive personal information. It has been shown that models learned from such datasets can inherit societal bias and discrimination patterns [1], [2] and learn sensitive features even when they are not explicitly used during training [3]. Concerns about the fairness, bias, and privacy of learning algorithms have led to a growing body of research focused on both defining meaningful fairness measures and designing algorithms with such guarantees.

A key challenge in algorithmic fairness is the quantification of disparate treatment and impact – legal notions developed to ensure that societal decisions neither hinder nor discriminate against specific groups. Addressing this has broadly lead to two classes of measures: (i) group fairness measures which require similar outcomes for all groups [4]; (ii) individual fairness measures which require treating similar individuals similarly [5]. Approaches combining both fairness requirements have also been considered [6], [7]. In the context of supervised learning of intended tasks (our setting here), two key group fairness measures have emerged [8]: (i) demographic parity (DemP) which requires predicted outcomes to be independent of the sensitive features, and (ii) equalized odds (EO) wherein such an independence holds only when conditioned on the true outcome. The EO measure was introduced to ensure accurate predictions within groups, a limitation of DemP [9].

Three distinct approaches have been considered to enforce fairness in learning: in-processing, pre-processing, and post-processing. In-processing approaches are most commonly used in the supervised setting where the learning objective is known (e.g., [5], [10]); the resulting trained model guarantees fairness for the specific objective. Pre-processing generally produces fair representations of data tuned for a chosen learning objective [11]–[13] while post-processing provides fairness by properly altering decision outputs [8], [14], [15].

Recently, *censoring* has emerged as a compelling pre-processing approach wherein protected features (e.g., race, gender, and their correlates) are actively decorrelated from the rest of the data to explicitly limit their effect on decisions. Censoring is inspired by information-theoretic privacy methods to limit leakage of sensitive features [3], [16]–[19] and can be achieved in practice using generative adversarial networks (GANs) [20]. Thus far, censoring for fairness has largely focused on learning fair predictors [10]–[12].

A. Our Contributions

Taking a preprocessing approach, the main contribution of this work is to use censoring to generate fair representations

Manuscript received May 18, 2021; revised November 20, 2021 and March 14, 2022; accepted April 1, 2022. Date of publication May 10, 2022; date of current version May 26, 2022. This work was supported in part by the National Science Foundation under Grant CCF-1350914, Grant CIF-1815361, Grant CIF-1901243, Grant SaTC-2031799, Grant DMS-2134256, and Grant CIF-2007688. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Matthieu R. Bloch. *(Peter Kairouz and Jiachun Liao are co-first authors.) (Corresponding author: Monica Welfert.)*

Peter Kairouz is with Google Research, Mountain View, CA 94043 USA (e-mail: kairouz@google.com).

Jiachun Liao, Chong Huang, and Maunil Vyas were with the School of Electrical, Computer, and Energy Engineering, Arizona State University, Tempe, AZ 85281 USA (e-mail: jliao15@asu.edu; chuang83@asu.edu; mrvyas@asu.edu).

Monica Welfert and Lalitha Sankar are with the School of Electrical, Computer, and Energy Engineering, Arizona State University, Tempe, AZ 85281 USA (e-mail: mwelfert@asu.edu; lsankar@asu.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/TIFS.2022.3170265, provided by the authors.

Digital Object Identifier 10.1109/TIFS.2022.3170265

1556-6021 © 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: ASU Library. Downloaded on June 10,2022 at 04:26:01 UTC from IEEE Xplore. Restrictions apply.

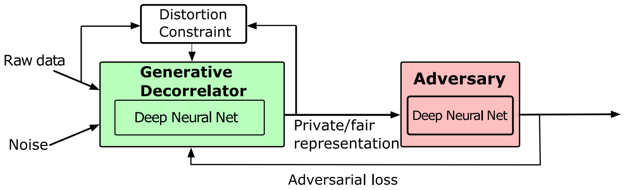


Fig. 1. Generative adversarial model for censoring/fairness.

that are *universal*. These are representations from which the sensitive features have been actively decoupled and can be universally used for a variety of *a priori* unknown learning tasks. We show that such fair universal representations (FURs) can assure DemP group fairness for all downstream predictions (from the data processing inequality). We now detail our contributions:

- We present a framework for learning FURs as a *constrained* minimax game between an encoder[1] and an adversary, where the encoder *generates a noisy representation* of the original data, subject to a distortion constraint, to thwart an adversary that actively tries to infer the sensitive features (see Fig. 1). There has been recent work on using adversarial methods to generate *transferable* fair representations [11]; our universal FUR approach, while similar in philosophy, goes a step further by enforcing a *hard* distortion constraint that allows better control of the learned representations, and therefore, better downstream utility guarantees. Algorithmically, we showcase how Lagrange penalty methods [21] can be leveraged to enforce the hard constraint in a GAN-setting.[2]
- Building on existing definitions of fair predictors, we formally define demographic parity for FRs. We use censoring (of the sensitive features) to ensure DemP FRs and provide information-theoretic assurances on both. Censoring methods, used often for assuring information-theoretic privacy of sensitive features when releasing data (e.g., [17], [19]), can also ensure DemP fairness (relative to the sensitive feature). Building on this, we formally define censored representations (Definition 2) for the setting when adversaries are limited to practical ML models and loss functions. This has a broader value in auditing fairness/censoring guarantees.
- Our prior work [17] shows that the constrained FUR minimax game captures a range of adversarial actions through the choice of loss functions and identifies the corresponding information-theoretic optimal adversarial strategies. Focusing here on high-dimensional Gaussian mixture models with independent components, we present the optimal additive Gaussian noise distribution that minimizes the adversary's probability of detecting the mixture class (sensitive feature), i.e., the game-theoretic optimal under the strongest MAP (maximum *a posteriori*) adversary. We then use normalized mutual information estimates to show that the empirical FUR framework, using neural networks and log-loss, performs very well relative to this game-theoretic optimal.
- Our most important contribution is in illustrating the utility of FURs for multiple publicly available datasets including the UCI Adult [23], UTKFace [24], GENKI [25], and HAR [26] datasets. Our visual results demonstrate our success in creating high quality representations that increasingly erase the sensitive attributes with decreasing fidelity requirements. In contrast to state-of-the-art [10]–[12], our theoretical framework and experiments are the first to include non-binary sensitive attributes, multiple downstream tasks, as well as hard distortion constraints. Our results show that one can still learn high accuracy DemP (and even EO) fair classifiers from DemP FURs. In particular, via the UCI dataset, that is often used in fair ML analyses, we showcase the advantage of our approach relative to related approaches (e.g., [12], [10], [11], [27]). Our results also straddle a wide range of values for the chosen fairness measure (DemP) and include perfect fairness, in contrast to the above works.

The remainder of our paper is organized as follows. We set up the problem and review known measures for fair predictors in Section II. In Section III, we formalize our FUR model, introduce definitions for censored and fair representations, and highlight the theoretical guarantees of this approach. In Section IV, we present theoretical results for datasets modeled as multi-dimensional Gaussian mixtures. Finally, we showcase the performance of the FUR framework on the UCI Adult, UTKFace, GENKI, and HAR datasets in Section V. Proofs for the key results are in the appendix. All proofs that build on prior results in [17] can be found in an extended version [28]. Finally, details of the FUR architecture for all datasets are in the accompanying supplementary material.

II. Preliminaries

Consider a dataset $\mathcal{D}$ with n entries where each entry is a random tuple $(S, X, Y) \in \mathcal{S} \times \mathcal{X} \times \mathcal{Y}$ where S, X, and Y are sensitive, non-sensitive, and target (non-sensitive) features, respectively, and $\hat{Y} \in \mathcal{Y}$ is a predictor of Y. Note that S and Y can be a collection of features or labels (e.g., S can be gender, race, sexual orientation, or a combination of these, while Y could be age, facial expression, etc.); for ease of writing, we use the term variable to denote both single and multiple features/labels. Instances of X, S, and Y are denoted by x, s and y, respectively. The entries (X, S, Y) of $\mathcal{D}$ are independent and identically distributed (i.i.d.) according to $P(X, S, Y)$.

Recent results on algorithmic fairness guarantee that, for a specific target Y, the prediction of a machine learning model is accurate with respect to (*w.r.t.*) Y but unbiased *w.r.t.* the sensitive S. While more than two dozen measures for fairness have been proposed, two oft-used fairness measures are demographic parity and equalized odds (and variants thereof). Demographic parity (DemP) ensures complete independence between the prediction of the target $\hat{Y}$ and the sensitive S; this notion of fairness favors utility the least, especially when

[1] referred to interchangeably throughout the paper as fair encoder or generator or decorrelator

[2] Recently TensorFlow updated its package to allow enforcing hard constraints [22] using a similar approach.

Authorized licensed use limited to: ASU Library. Downloaded on June 10,2022 at 04:26:01 UTC from IEEE Xplore. Restrictions apply.

Y and S are correlated [8]. Equalized odds (EO) enforces this independence conditioned on Y, thereby ensuring equal rates for true and false positives (when Y is binary) for all demographics. We now define DemP and EO formally (for binary S and Y as originally introduced). These definitions can be generalized to the non-binary setting and we do so in the sequel for fair representations.

Definition 1 [8]: A predictor $f(S, X) = \hat{Y}$ satisfies

- *demographic parity (DemP) w.r.t. S, if $\hat{Y} \perp S$, i.e.,*

$$\Pr(\hat{Y}=1|S=1)=\Pr(\hat{Y}=1|S=0) \quad (1)$$

- *equalized odds (EO) w.r.t. (S, Y), if $\hat{Y} \perp S|Y$, for $y \in \{0, 1\}$:*

$$\Pr(\hat{Y}=1|S=1, Y=y)=\Pr(\hat{Y}=1|S=0, Y=y). \quad (2)$$

In the following section, we present our FUR framework which includes definitions for both demographically fair and censored representations.

III. FURs via Generative Adversarial Models

Formally, the FUR model consists of two components, an encoder and an adversary, as shown in Fig. 1. The goal of the encoder $g : \mathcal{X} \times \mathcal{S} \rightarrow \mathcal{X}_r$ is to actively eliminate the dependence between S and X while that of the adversary $h : \mathcal{X}_r \rightarrow \mathcal{S}$ is to infer S. In general, $g(X, S)$ is a randomized mapping that outputs a representation $X_r = g(X, S)$. Note that S may not always be available to the curator; however, it will always affect the design of g via the adversarial training process. For brevity, we henceforth write $g(\cdot)$ to include both possibilities (just X or (X, S) as inputs). On the other hand, the role of the adversary is captured via $h(X_r)$, which is the adversarial decision rule in inferring the sensitive variable S as $\hat{S} = h(X_r = g(\cdot))$ from the representation $g(\cdot)$. In general, the hypothesis h can be a *hard decision rule* under which $h(g(\cdot))$ is a direct estimate of S or a *soft decision rule* under which $h(g(\cdot)) = P_h(\cdot|g(\cdot))$ is a distribution over $\mathcal{S}$.

To quantify the adversary's performance, we use a loss function $\ell(h(g(X = x, S = s)), S = s)$ defined for every pair (x, s). Thus, the adversary's expected loss *w.r.t.* X and S is $L(h, g) \triangleq \mathbb{E}[\ell(h(g(\cdot)), S)]$, where the expectation is taken over $P(X, S)$ and the randomness in g and h. To ensure utility, we introduce a constraint on the fidelity of X_r via a distortion function $d(x_r, x)$, which measures the goodness of $X_r = x_r$ *w.r.t.* $X = x$. Ensuring statistical utility, in turn, requires constraining the average distortion $\mathbb{E}[d(g(\cdot), X)]$, where the expectation is taken over $P(X, S)$ and the randomness in g.

A. FUR: Framework and Theoretical Results

To publish a fair representation X_r, the data curator wishes to learn an encoder g that guarantees censoring (i.e., it is difficult for the adversary to learn S from X_r), and therefore, fair X_r under DemP, as well as utility (g guarantees bounded distortion of X). In contrast, for a fixed g, the adversary would like to find a (potentially randomized) function h that minimizes its expected loss, or equivalently maximizes the negative expected loss. This leads to a constrained minimax game between the encoder and the adversary given by

$$\min_{g(\cdot)} \max_{h(\cdot)} -\mathbb{E}[\ell(h(g(\cdot)), S)], \text{ s.t. } \mathbb{E}[d(g(\cdot), X)] \leq D. \quad (3)$$

where $D \geq 0$ determines the distortion constraint on X_r. The optimization in (3) highlights that the input to g depends on whether the curator has access to both (X, S) or just X. Having access to both (X, S) in general will yield a better decorrelator (e.g., see Section V-A for the UCI dataset). Finally, without the constraint in (3), the optimal $X_r = g(\cdot) \perp S$. One can approximate this in practice via arbitrarily large distortions as we show in Proposition 2; as a setup to these results, we first define censoring and fairness for representations. Our censoring definition clarifies the representation that best limits an adversary from inferring S. We then define DemP for FRs; we combine the two definitions to show how and when adversarial learning can help ensure demographic parity.

Definition 2 (Censored Representations): A representation X_r of X is censored w.r.t. the sensitive features S against a learning adversary $h(\cdot)$, whose performance is evaluated via a loss function $\ell(h(X_r), S)$, if for an optimal adversarial strategy $h^* = \arg\min_h \mathbb{E}[\ell(h(X_r), S)]$,

$$\mathbb{E}[\ell(h^*(g(\cdot)), S)] \leq \mathbb{E}[\ell(h^*(X_r), S)], \quad (4)$$

where $g(\cdot)$ is any (randomized) function of X (or (X, S)) and the expectation is over h, g, X, and S.

The above definition suggests that the best censored representation X_r is the least informative about S to an adversary whose inferential action is captured by a loss function $\ell(\cdot, \cdot)$, i.e., the average loss is the worst for X_r than for any other arbitrary function $g(\cdot)$. While the comparison in (4) is w.r.t. the best $h^*(X_r)$ for X_r, choosing the optimal $h(\cdot)$ for any $g(\cdot)$ will only serve as a lower bound to the left side of (4).

We now define DemP for representations; we then prove that a demographically fair representation X_r guarantees that any downstream algorithm using X_r satisfies DemP *w.r.t.* S. In the following, we assume that X, and therefore, X_r are discrete random variables with arbitrarily large alphabets; however, the definition below can be extended to continuous-valued X and X_r by considering all Borel subsets and an appropriately defined measure on the space.

Definition 3 (Demographically Fair Representations): For $(X, S) \in \mathcal{X} \times \mathcal{S}$, a representation $X_r = g(X, S) \in \mathcal{X}_r$ satisfies demographic parity w.r.t. S if for any $x_r \in \mathcal{X}_r$ and $s, s' \in \mathcal{S}$

$$\Pr(X_r = x_r|S = s) = \Pr(X_r = x_r|S = s') \quad (5)$$

where $g : \mathcal{X} \times \mathcal{S} \rightarrow \mathcal{X}_r$ is any (possibly randomized) function.

Theorem 1 (Fair Learning via Fair Representation): If $X_r = g(X, S)$ satisfies demographic parity w.r.t. S, then any algorithm $f : \mathcal{X}_r \rightarrow \mathcal{Y}$ satisfies demographic parity w.r.t. S.

The proof of Theorem 1 follows from a direct application of the data-processing inequality for mutual information since $(X, S) - X_r - Y$ form a Markov chain; details can be found in [28].

Remark 1: Note that equalized odds in Def. 1 explicitly involves a downstream task, and therefore, the design of an EO

Authorized licensed use limited to: ASU Library. Downloaded on June 10,2022 at 04:26:01 UTC from IEEE Xplore. Restrictions apply.

TABLE I

THE ADVERSARIES CAPTURED BY THE FUR FRAMEWORK BY USING A VARIETY OF LOSS FUNCTIONS

	Loss function $\ell(h(g(\cdot)), s)$	Optimal adversarial strategy h^*	Adversary type
Squared loss	$(h(g(\cdot)) - S)^2$	$\mathbb{E}[S\|g(\cdot)]$	MMSE adversary
0-1 loss	$\begin{cases} 0 & \text{if } h(g(\cdot)) = S \\ 1 & \text{otherwise} \end{cases}$	$\underset{s \in \mathcal{S}}{\operatorname{argmax}}\, P(s\|g(\cdot))$	MAP adversary
Log-loss	-log $P_h(s\|g(\cdot))$	$P(s\|g(\cdot))$	Belief refining adversary
α-loss	$\frac{\alpha}{\alpha-1}\left(1 - P_h(s\|g(\cdot))^{1-\frac{1}{\alpha}}\right)$	$\frac{P(s\|g(\cdot))^{\alpha}}{\sum_{s \in \mathcal{S}} P(s\|g(\cdot))^{\alpha}}$	Generalized belief refining adversary

fair X_r needs to include a predictor explicitly. In contrast to the FUR setting considered here, such targeted representations and the ensuing fair predictors provide guarantees only for specific target Y. In this limited context, however, one can still define an X_r as ensuring EO w.r.t. to (S, Y) if $\hat{Y}(X_r) \perp S|Y$.

One simple approach to obtain a fair/censored representation X_r is by choosing $X_r = N$ where $N \perp (X, S)$. However, such an X_r has no utility (quantified, for example, via downstream task accuracy). The design of X_r has to ensure utility, and thus, there is a tradeoff between guaranteeing fairness/censoring and assuring a desired level of utility. We now quantify such tradeoffs using FUR framework.

Theorem 2: For sufficiently large distortion bound D, (3) *yields a universal representation X_r censored w.r.t. to S.*

The proof follows by observing that for sufficiently large D, X_r can be arbitrarily noisy, reducing (3) to an unconstrained optimization. For this X_r with $h^* = \arg\min_h \mathbb{E}[\ell(h(X_r), S)]$,

$$\mathbb{E}[\ell(h^*(X_r), S)] = -\min_{g(\cdot)} \max_{h(\cdot)} -\mathbb{E}[\ell(h(g(\cdot)), S)] \quad (6)$$

$$\geq \mathbb{E}[\ell(h^*(g(\cdot)), S)], \quad (7)$$

thus satisfying Definition 2.

The FUR framework in (3) places no restrictions on the adversary. Indeed, different loss functions and decision rules lead to different adversarial models (see Table I). This versatility to a large class of (inferring) adversarial models is captured by the last entry in Table I by using the recently introduced tunable *α-loss* [29], [30], defined for $\alpha \in (0, 1) \cup (1, \infty)$ as:

$$\ell_\alpha(h(g(\cdot)), s) = \frac{\alpha}{\alpha - 1}\left(1 - P_h(s|g(\cdot))^{\frac{\alpha-1}{\alpha}}\right) \quad (8)$$

with continuous extensions at $\alpha = 1$ and $\alpha = \infty$ (the loss simplifies to a constant for $\alpha = 0$). Note that the loss in (8) operates on a soft decision $P_h(\cdot|\cdot)$, the output of the adversary h. By tuning $\alpha \in [0, \infty]$, α-loss captures a variety of information-theoretic adversaries as listed in Table I (see also [29], [30]):

(i) a hard-decision adversary for $\alpha = \infty$ captured by $\ell_\infty(h(g(\cdot)), s) = 1 - \Pr[h(g(\cdot)) = s]$,[3] and

(ii) a soft-decision adversary for $\alpha = 1$ via the oft-used log-loss $\ell_1(h(g(\cdot)), s) = -\log P_h(s|g(\cdot))$ (this follows directly by applying L'Hôpital's rule).

(iii) Values of $\alpha > 1$ allow interpolating between the NP-hard to implement MAP rule ($\alpha = \infty$) and log-loss ($\alpha = 1$) and allow some robustness to noisy data [30]. On the other hand, choosing $\alpha < 1$ leads to more convex losses than log-loss ($\alpha = 1/2$ yields a soft exponential loss used in boosting algorithms) that are more sensitive to outliers [30].

For any encoder $g(\cdot)$, the following proposition (see also the last row of Table I) summarizes the optimal $P_h^*(s|g(\cdot))$ under α-loss.

Proposition 1: For a fixed g, under α-loss, the optimal adversary decision rule that minimizes the expected loss is a 'α-tilted' conditional distribution $P_h^*(s|g(\cdot)) = \frac{P(s|g(\cdot))^\alpha}{\sum_{s \in \mathcal{S}} P(s|g(\cdot))^\alpha}$. For $\alpha = 1$ and $\alpha = \infty$, we obtain the optimal strategies for log-loss and 0-1 loss, respectively, as the true conditional distribution and the maximal a posteriori (MAP) estimator. Then, (3) *reduces to $\min_{g(\cdot)} -H_\alpha^A(S|g(\cdot))$, where $H_\alpha^A(\cdot|\cdot)$[4] is the Arimoto conditional entropy.*

Proposition 1 states that if the adversary uses ℓ_∞, (3) simplifies to $\min_{g(\cdot)} P(g(\cdot)) \max_{s \in S} P(s|g(\cdot)) - 1$, i.e., we choose the most likely s for every $g(\cdot)$ (the MAP rule) [31].

On the other hand, if the adversary uses log-loss, (3) simplifies to $\min_{g(\cdot)} I(g(\cdot); S)$ for any prior on S, where $I(g(\cdot); S)$ is the mutual information (MI) between $g(\cdot)$ and S. More generally, using α-loss in (8), the optimal P_h^* in Proposition 1 simplifies the objective in (3) to $\min_{g(\cdot)} I_\alpha^{\mathrm{A}}(g(\cdot); S)$, where $I_\alpha^{\mathrm{A}}(g(\cdot); S)$ is the Arimoto MI[5] of order α. These MIs can be effective proxies for guaranteeing DemP FURs, since they are minimized only when $S \perp g(\cdot)$, thus leading to the following theorem.

Theorem 3: Under α-loss, for all α, (3) *enforces fairness subject to a distortion constraint. As the distortion increases, the ensuing fairness guarantee approaches ideal DemP.*

Proposition 1 was proved in [17]; Theorem 3 involves similar arguments to Theorem 2. All proofs can be found in [28]. Many notions of fairness (cf. Definition 1) require computing conditional probabilities for every sample x to ensure independence, and thus, are not easy to optimize in a data-driven fashion. The FUR framework via loss functions captures mutual information-like surrogates for such independence conditions; to this end, Theorem 3 justifies using α-loss (and thus, log-loss too) as a proxy for enforcing fairness. We remark that mutual information (MI) is a common surrogate fairness measure for demographic/statistical parity [6], [13], [32]. In [32], the authors use MI for both fairness and the distortion measure leading to an (non-convex) information bottleneck problem; we recover this formulation by choosing

[3] For $\alpha = \infty$, α-loss reduces to probability of error, for which the optimal rule that minimizes the expected loss is the *maximal a posteriori* (MAP) estimation, a hard decision. The same rule results when minimizing the 0-1 loss which is given by $\mathbb{I}(h(g(\cdot)) \neq s)$ where $\mathbb{I}$ is the indicator function.

[4] $H_\alpha^{\mathrm{A}}(U|V) \triangleq 1/(1-\alpha)\log(\sum_{u,v} P_{U,V}^{\alpha}(u, v))$

[5] $I_\alpha^{\mathrm{A}}(g(\cdot); S) = H_\alpha(S) - H_\alpha^{\mathrm{A}}(S|g(\cdot))$ where $H_\alpha(S) \triangleq H_\alpha^{\mathrm{A}}(S)$ is the Rényi entropy of order α and is also the unconditioned α-Arimoto entropy.

Authorized licensed use limited to: ASU Library. Downloaded on June 10,2022 at 04:26:01 UTC from IEEE Xplore. Restrictions apply.

$d(x, x_r) = -\log p(x_r|x)$. Thus the FUR framework is more general and allows choosing application-dependent meaningful fidelity measures (for example, different measures of representation similarity are used in natural language processing and healthcare data). Finally, we remark that the optimal adversarial strategy in Proposition 1 requires estimating a posterior; as described in Section III-B, in practice, one can use deep learning models for h and g to do so.

A predominant approach in the literature in the context of fair representations is to explicitly include the intended classification/prediction task, i.e., design representations that guarantee DemP for the specific task [10]–[12]. In fact, the FUR formulation in (3) can be extended to include this case by adding an additional term in the objective function to ensure high accuracy in learning Y. The resulting minimax game is given by

$$\min_{\tilde{g}(\cdot), f(\cdot)} \max_{h(\cdot)} -\mathbb{E}[\ell(h(\tilde{g}(\cdot)), S)] + \lambda\mathbb{E}[\ell'(f(\tilde{g}(\cdot)), Y)], \quad (9a)$$

$$\text{s.t. } \mathbb{E}[d(\tilde{g}(\cdot), X)] \le D, \quad (9b)$$

where $f(\cdot)$ is a classifier for a target Y, $\lambda > 0$, and $\tilde{g}(\cdot)$[6] and $h(\cdot)$ are the encoder and the adversarial classifier, respectively, as in (3). Note that the loss functions $\ell(\cdot)$ and $\ell'(\cdot)$ can be different. The setup in (9) involves an additional term ensuring fair classification and is, thus, a more constrained optimization than the FUR framework; in fact, we recover the FUR setup with $\lambda = 0$. However, even while generating intermediate representations $g(\cdot)$, (9) is primarily intended to design *fair classifiers*, and therefore, requires knowing the intended tasks on Y. In contrast, our FUR framework allows generating DemP-guaranteeing fair X_r that in turn guarantee DemP fairness to all downstream tasks on any subset of Y.

One can also design fair classifiers directly without intermediate representations by setting $\tilde{g}(\cdot) \triangleq \hat{Y}$; such classifiers $\tilde{g}$ can be designed with either DemP or EO guarantees. We first consider the more general problem of designing EO-fair predictors/classifiers $\tilde{g}(\cdot)$ for the target Y and show that our FUR framework subsumes this problem (and therefore, that of generating DemP predictors/classifiers). Let h be the adversary decision rule to infer S as $\hat{S} = h(\tilde{g}(\cdot)|Y)$ for every choice of Y via the soft predictor $\tilde{g}(\cdot) = P_{\hat{Y}|\cdot}$. Then, analogous to (3), the design of a fair predictor/classifier can be formulated as

$$\min_{\tilde{g}(\cdot)} \max_{h(\cdot)} -\mathbb{E}\big[\ell\big(h(\tilde{g}(\cdot)|Y), S\big)\big], \text{ s.t. } \mathbb{E}[\ell\big(\tilde{g}(\cdot), Y\big)] \le \epsilon, \quad (10)$$

where the expectation now includes Y too.

Theorem 4: Under α-loss, the formulation in (10) *enforces EO fairness subject to a performance constraint ϵ. As ϵ increases, the ensuing fairness guarantee approaches the ideal EO guarantees achievable by $\tilde{g}$ w.r.t. S and Y.*

The proof of Theorem 4 is similar to that for Theorem 2 and can be found in [28]. Note that the formulation in (10) also holds for generating a fair predictor/classifier satisfying DemP in Definition 1. In contrast to EO where the adversary needs both $\tilde{g}(\cdot)$ and Y as inputs, for DemP, only $\tilde{g}(\cdot)$ is input to the adversary.

The adversarial models and the resulting game-theoeretic solutions in Table I highlight the formal guarantees of the FUR framework. Recently, Sypherd *et al.* have demonstrated the robustness of training deep learning models with α-loss to both noise and class imbalances [30], thus promising to be applicable for learning FURs with GANs. The rest of the sequel focuses on data-driven GANs with $\alpha = 1$, i.e., log-loss, to highlight the value of FURs in guaranteeing fairness for multiple downstream tasks relative to the state-of-the-art fair classifiers. Future work will include enhancing such results to include tuning over α.

B. Data-Driven FUR

Thus far, we have focused on a setting where the curator has access to the statistics $P(X, S)$ thereby solving the constrained minimax optimization problem in (3) (game-theoretic version of the FUR formulation) to obtain a g that performs best against a chosen adversary. In practice, $P(X, S)$ is impossible to compute. To this end, we propose a data-driven version of the FUR formulation that allows the data holder to learn a generative decorrelator from an n-sample dataset $\mathcal{D} = \{(x_{(i)}, s_{(i)})\}_{i=1}^{n}$ via a generative model $g(X; \theta_p)$ that is parameterized by θ_p. This generative model takes X (or (X, S)) as input and outputs X_r. 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 $h(g(X; \theta_p); \theta_a)$ that is parameterized by θ_a. In the evaluation phase, the performance of the learned decorrelation scheme 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.

While, in theory, the functions h and g can be arbitrary, in practice, they are best approximated by a well-chosen rich hypothesis class. Fig. 1 illustrates a FUR model in which h and g are both modeled as deep neural networks (DNNs). For a fixed h and g, binary S and $\ell = \ell_1$ (log-loss for $\alpha = 1$), the adversary's *empirical loss* using cross entropy is given by

$$L_n(\theta_p, \theta_a) = -\frac{1}{n}\sum_{i=1}^{n} s_{(i)} \log h(g(x_{(i)}; \theta_p); \theta_a) + (1 - s_{(i)}) \log(1 - h(g(x_{(i)}; \theta_p); \theta_a)). \quad (11)$$

The optimal model parameters (θ_p, θ_a) are then solutions of

$$\min_{\theta_p} \max_{\theta_a} -L_n(\theta_p, \theta_a), \text{ s.t. } \frac{1}{n}\sum_{i=1}^{n} d(g(x_{(i)}; \theta_p), x_{(i)}) \le D. \quad (12)$$

The minimax optimization in (12) is a two-player non-cooperative game between the generative decorrelator and the adversary with strategies θ_p and θ_a, respectively. In practice, for chosen hypothesis classes for g and h (e.g., DNN architectures), we can learn the equilibrium of the game using an iterative algorithm as follows. (i) For a fixed θ_p, maximize the negative of the adversary's loss to compute the parameters of h. (ii) Then, minimize the decorrelator's loss (negative adversary loss) to compute θ_p for a fixed h.

It is crucial to note that the *hard* distortion constraint in (12) makes our minimax problem different from what has

[6] In general, $\tilde{g}(\cdot)$ can be a function of both X and S; the dependence on S is implicit when S is not directly available.

Authorized licensed use limited to: ASU Library. Downloaded on June 10,2022 at 04:26:01 UTC from IEEE Xplore. Restrictions apply.

been extensively studied in the literature. To incorporate the distortion constraint, we use the *penalty method* [21] to replace the constrained optimization problem by adding a penalty to the objective function. This is done via a penalty parameter ρ_t that captures a measure of violation of the constraint at the t^{th} iteration. The constrained optimization problem of g is then approximated by a *series of unconstrained optimization problems* with an objective

$$-L_n(\theta_p, \theta_a) + \rho_t(\max\{0, \frac{1}{n}\sum_{i=1}^{n} d(g(x_{(i)};\theta_p), x_{(i)}) - D\})^2, \tag{13}$$

where the penalty coefficient ρ_t decreases with the number of iterations t. We note that both the augmented Lagrangian and the penalty methods have similar performance in practice; we chose the penalty method but our results can also be obtained with the augmented Lagrangian method [33]. We provide detailed steps of the algorithm and the parameters for the penalty method in Appendix A; we also clarify our methodology for choosing both ρ_t and the learning rate η_t there. Finally, we note that one can easily generalize (11) to the multi-class setting (non-binary S) using the softmax function; one can also generalize (11) using α-loss.

In the following sections, we detail our results for synthetic multi-dimensional Gaussian mixture data and four publicly available datasets: UCI Adult, UTKFace, GENKI, and HAR. All code is available via GitHub at [28].

IV. FUR 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 $\mu = (\mu_1, \ldots, \mu_m)$. We assume that $X|S = 0$ and $X|S = 1$ have the same covariance Σ.

A. Game-Theoretical Approach

We consider a MAP adversary that has access to $P(X, S)$ and g. The goal is to censor X in a way that minimizes the adversary's probability of correctly inferring S from X_r. In order to have a tractable model for the encoder, we mainly focus on affine representations $X_r = g(X) = X + Z + \beta$, where for tractability reasons, we choose Z as a zero-mean multi-dimensional Gaussian random vector independent of X. This linear representation enables controlling both the mean and covariance of X_r. To quantify utility of the privatized data, we use the ℓ_2 distance between X and X_r as a distortion measure to obtain a constraint $\mathbb{E}_{X,X_r}\|X - X_r\|^2 \le D$.

We assume that $\beta = (\beta_1, \ldots, \beta_m)$ is a constant parameter vector and $Z \sim \mathcal{N}(0, \Sigma_p)$. Building on the analysis in [34], for the standard Gaussian $Q(\cdot)$ function, we can derive the adversary's detection probability $P_d^{(G)}$ as

$$P_d^{(G)} = qQ\left(-\frac{\gamma}{2} + \frac{1}{\gamma}\ln\left(\frac{1-q}{q}\right)\right) + (1-q)Q\left(-\frac{\gamma}{2} - \frac{1}{\gamma}\ln\left(\frac{1-q}{q}\right)\right), \tag{14}$$

where $\gamma = \sqrt{(2\mu)^T(\Sigma + \Sigma_p)^{-1}2\mu}$. From the constraint, we have $\mathbb{E}_{X,X_r}\|X - X_r\|^2 = \|\beta\|^2 + tr(\Sigma_p) \le D$. The mixture Gaussian classification problem, especially for the same covariance for both classes $S = 0$ and $S = 1$, is a tractable problem that has been studied in a variety of settings including communication systems [34] and machine learning [35], to name a few. The result in (14) builds directly on [34], and so, for reasons of brevity, we leave it out. The following theorem summarizes the optimal noising strategy when $X|S$ and Z are multi-dimensional i.i.d. Gaussians.

Theorem 5: Consider the representation given by $g(X) = X + Z + \beta$, where $X|S$ and Z are Gaussian random vectors with diagonal covariance matrices Σ and Σ_p, respectively, and $X|S \perp Z$. Let $\{\sigma_1^2, \ldots, \sigma_m^2\}$ and $\{\sigma_{p_1}^2, \ldots, \sigma_{p_m}^2\}$ be the diagonal entries of Σ and Σ_p, respectively. The parameters of the minimax optimal censoring mechanism g^* are

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

where λ_0^*, the dual variable enforcing the distortion constraint in (3)*, is chosen such that $\sum_{i=1}^m {\sigma_{p_i}^*}^2 = D$. For this optimal mechanism, the accuracy of the MAP adversary is given by* (14) *with $\gamma = 2\sqrt{\sum_{i=1}^m \mu_i^2/(\sigma_i^2 + {\sigma_{p_i}^*}^2)}$.*

The proof of Theorem 5 is in Appendix B. We observe that when $\sigma_i^2 > |\mu_i|/\sqrt{\lambda_0^*}$, no noise is added to the data on this dimension due to the high variance. In contrast, when $\sigma_i^2 < |\mu_i|/\sqrt{\lambda_0^*}$, the variance of the 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, require more noise to reduce the MAP adversary's inference accuracy.

We note that we could have considered a more general non-affine model for the encoder. Since synthetic datasets provide a verifiable way to formally evaluate the FUR framework, we chose a simpler affine generative model that is tractable and yields closed-form information-theoretic results, i.e., we can derive the best adversarial decoder. This in turn is helpful in comparing the data-driven approach with the game-theoretic optimal solution on these canonical data models as a much-needed sanity check. Finally, the analysis here can be generalized to correlated Gaussian distributions for each sensitive group and one expects a similar behavior as the features can be whitened when the covariance is the same for both classes.

B. Data-Driven Approach

To learn the data-driven representation $X_r = g(X) = X + Z + \beta$ using our FUR framework, we assume that g only has access to the dataset $\mathcal{D}$ with n data samples (not $P(X, S)$). Computing the optimal g^* is then a learning problem. In the training phase, we learn the parameters θ_p of g by competing against a computational adversary $h(g(\theta_p);\theta_a)$ 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 X_r,

Authorized licensed use limited to: ASU Library. Downloaded on June 10,2022 at 04:26:01 UTC from IEEE Xplore. Restrictions apply.

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 censoring mechanism.

Since the sensitive variable S is binary, we measure the training loss of the adversary network using the empirical log-loss function in (11). We model the encoder using a two-layer neural network with parameters $\theta_p = \{\beta_1, \ldots, \beta_m, \sigma_{p_1}, \ldots, \sigma_{p_m}\}$, where β_k and σ_{p_k} represent the mean and standard deviation for each dimension $k \in \{1, \ldots, m\}$, respectively. The random noise Z is drawn from a m-dimensional independent zero-mean standard Gaussian distribution such that $\{\sigma_{p_k} Z_k\}_{k=1}^{m}$ jointly have a covariance $\Sigma_p = \mathrm{diag}(\sigma_{p_1}^2, \ldots, \sigma_{p_m}^2)$. Thus, $\hat{X}_k = X_k + \beta_k + \sigma_{p_k} Z_k$. The adversary is modeled by a three-layer neural network classifier with leaky ReLU activations. Finally, as detailed in Algorithm 1, we use the penalty method to ensure the distortion constraint.

C. Illustration of Results

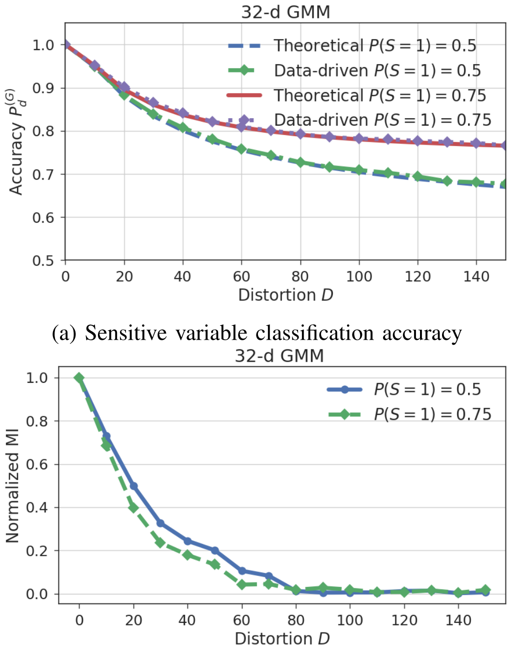
We generate two synthetic datasets to illustrate our results; each dataset has $20K$ training samples and $2K$ test samples. Each dataset is generated by sampling from an independent multi-dimensional Gaussian mixture model. The two datasets correspond to two distinct values for the prior $P(S = 1)$ as 0.75 and 0.5. Both encoder and adversary are trained via Tensorflow [36] using the Adam optimizer [37] with a learning rate of 0.005 and a minibatch size of 1000.

Fig. 2 illustrates the performance of the learned FUR scheme against a strong theoretical MAP adversary for a 32-dimensional Gaussian mixture model for both $P(S = 1) = 0.75$ and 0.5. We observe that the inference accuracy of the MAP adversary decreases as the distortion increases and asymptotically approaches (as expected) the prior $P(S = 1)$. The encoder obtained via the data-driven approach performs very well when pitted against the MAP adversary (maximum accuracy difference around 0.7% compared to the theoretical optimal). As another measure of censoring, we estimate the MI of X_r and S using k-nearest neighbor method as detailed below. Normalizing it with its theoretical maximum $I(X; S)$ (i.e., with $X_r = X$ for $D = 0$), in Fig. 2b, we show that MI $\hat{I}(X_r; S)/I(X; S)$ decreases, as expected, when the distortion increases. In other words, for Gaussian mixture data with binary S, the data-driven FUR formulation can learn decorrelation schemes that perform as well as those computed under the game-theoretical FUR formulation where the generative decorrelator has access to the data statistics.

We estimate MI using the k-nearest neighbor method [38]; in particular, for n d-dimensional FUR outputs X_r, we first estimate the entropy $\hat{H}(X_r)$ as

$$\hat{H}(X_r) = \psi(N) - \psi(k) + \log(c_d) + \frac{d}{n}\sum_{i=1}^{n} \log r_i \qquad (15)$$

where r_i is the distance of the i-th sample $\hat{x}_i$ to its k-th nearest neighbor, ψ is the digamma function (logarithmic derivative of the gamma function $\Gamma(\cdot)$), and $c_d = \frac{\pi^{d/2}}{\Gamma(1+d/2)}$. We then calculate MI as $\hat{I}(X_r; S) = \hat{H}(X_r) - P(S = 1)\hat{H}(X_r|S = 1) - P(S = 0)\hat{H}(X_r|S = 0)$, where $P(S = 1)$ and $P(S = 0)$ are empirically estimated.



(a) Sensitive variable classification accuracy

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

Fig. 2. Performance of the FUR framework for GMMs.

V. FUR for Publicly Available Datasets

We apply our FUR framework to four real-world datasets, namely, UCI Adult [23], UTKFace [24], GENKI [25], and HAR [26], briefly described below. For all four datasets, we restrict the architecture of h, g, and the downstream predictive models to neural networks. Note that for tabular datasets (e.g., UCI, HAR), boosting methods including decision trees or support vector models achieve at least comparable predictive performance [23], [39] but are out of scope of this work.

(i) The UCI Adult dataset [23] consists of 10 categorical and 4 continuous features and is used to predict a binary salary label (1: salary > 50k or 0: salary ≤ 50k). We choose gender or the tuple (gender, relationship) as the sensitive S, the remaining features except salary as non-sensitive X (Table SIV in the supplement lists all features), and salary as the target Y.

(ii) The UTKFace dataset [24] consists of more than 20k 200 × 200 color images of faces labeled by age, ethnicity, and gender. Individuals in the dataset have ages from 0 to 116 years and are divided into 5 ethnicities: White, Black, Asian, South Asian Indian, and others including Hispanic, Latino and Middle Eastern. We take gender as S, the image as X, and age and ethnicity as two target labels Y. Further, we also restrict the data to contain images for ages between 10 and 65.

(iii) The GENKI dataset [25] consists of $1,740$ training and 200 test samples. Each data sample is a 16×16 greyscale face image with varying facial expressions. We choose gender as S and the image as X.

(iv) The HAR dataset [26] consists of 561 features of motion sensor data collected by a smartphone from 30 subjects performing six activities (walking, walking upstairs, walking

Authorized licensed use limited to: ASU Library. Downloaded on June 10,2022 at 04:26:01 UTC from IEEE Xplore. Restrictions apply.

downstairs, sitting, standing, laying). Each feature is normalized between -1 and 1. We choose subject identity as S and the features of motion sensor data as X.

We use the accuracy of predicting S as the measure of censoring. We evaluate the fairness guarantees of X_r by computing the DemP obtained on tasks using Y. To this end, we compute the following maximal difference as a proxy for DemP in Definition 1 (includes non-binary Y and S):

$$\Delta_{\text{DemP}}(y) = \max_{s,s' \in \mathcal{S}} |P(\hat{Y} = y|S = s) - P(\hat{Y} = y|S = s')| \tag{16}$$

with smaller values of $\Delta_{\text{DemP}}(y)$ suggesting better DemP fairness guarantees. For binary Y, $\Delta_{\text{DemP}}(y)$ in (16) simplifies to a single value that we denote as Δ_{DemP}. In our experiments, we use the empirical frequencies to estimate $P(\hat{Y} = y|S = s)$ for a chosen (y, s). We illustrate both censoring and fairness results for the abovementioned datasets in the following subsections. Experimental and model details are in the supplement.

A. Illustration of Results for UCI Adult Dataset

For the UCI Adult dataset with both categorical and continuous features as shown in Table SIV in the supplement, we consider two cases:

(i) Case I: binary S by choosing 'gender' as sensitive feature

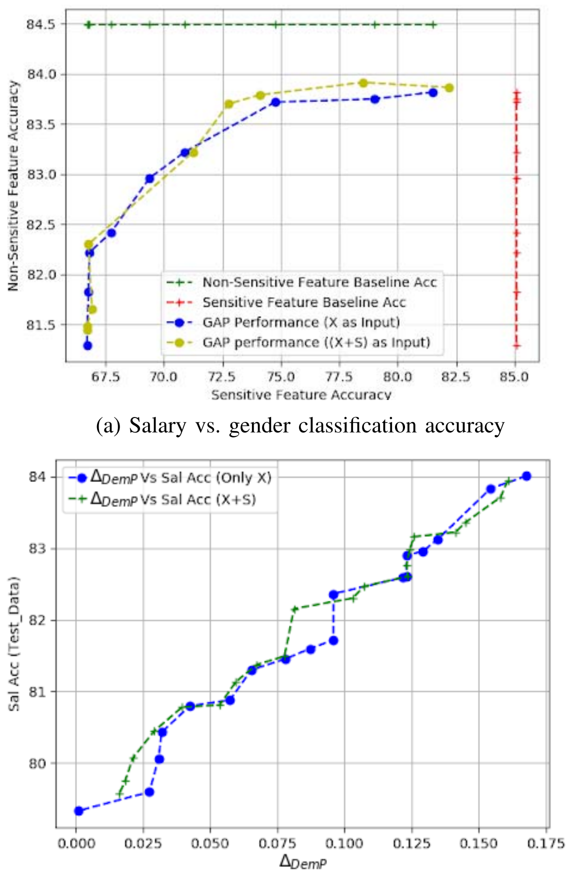
(ii) Case II: non-binary S by considering both 'gender' and 'relationship' as sensitive. For UCI, 'relationship' has 6 distinct values, and therefore, S has 12 possibilities.

For both cases, 'salary' is the binary target $Y \in \{0, 1\}$, with $Y = 1$ denoting salary $> 50K$. Since the two values for $\Delta_{\text{DemP}}(y)$ in (16) are the same for binary Y, we write Δ_{DemP} when illustrating results.

1) Case I: Binary Sensitive Feature: Fig. 3 illustrates the censoring and fairness performance of the generated X_r for the UCI dataset. For censoring, the performance is evaluated via the tradeoff between the classification accuracies of salary (utility of X_r) and gender (censoring of S). Note that salary accuracy is evaluated as a downstream task via a separately learned classifier that uses X_r while gender accuracy is a measure of performance of the neural network adversary h in the FUR model. We evaluate fairness via the tradeoff between salary accuracy and Δ_{DemP}. We consider two possible inputs to the encoder $g(\cdot)$ in (3), i.e., only X or both (X, S).

As illustrated in Fig. 3a, the baseline[7] salary and gender accuracies for the UCI dataset are about 84.5% and 85%, respectively. Further, for the FUR X_r and downstream $\hat{Y}$:

(i) the smallest gender accuracy achievable is about 66%, 20% below its baseline, while the lowest salary accuracy is about 82%, 2.5% below its baseline. Since the likelihood of a male in the original test data is 66%, with increasing distortion, the FUR gender accuracy is as good as a random guess, i.e., the generated X_r hides gender effectively while maintaining high salary accuracy.



(a) Salary vs. gender classification accuracy

(b) Salary classification accuracy vs. Δ_{DemP}

Fig. 3. Results for UCI Adult: Case I. In Fig. 3a, the green and red lines denote the baseline performances for the target Y (salary) and sensitive S (gender), respectively; in Fig. 3b, the value of Δ_{DemP} for the original test data is 0.2. In both plots, each point corresponds to a specific value of achieved test distortion; for Figs. 3a and 3b, the achieved test distortion for the blue points ranges over (0.69, 4.1) and (0.69, 4.4), respectively, with decreasing distortion from left to right for each plot. The achieved test distortion for the yellow-green points ranges over (0.87, 4.2) and (0.87, 4.9), respectively.

(ii) For the same gender accuracy, using both (X, S) seems useful only in the high utility setting (salary accuracy $\geq 83\%$). From Fig. 3b, we make the following two observations:

(i) salary classification accuracy and Δ_{DemP} have an approximately affine relationship, and when $\Delta_{\text{DemP}} \approx 0$, the salary accuracy is $\geq 79\%$, i.e., the FUR framework is effective in approaching perfect DemP with a small reduction in utility;

(ii) the FURs X_r generated from either X or (S, X) lead to similar fairness guarantees. For $\Delta_{\text{DemP}} = 0.06$, state-of-the-art approaches in [12] and [11] achieve 2% and 2.5% higher salary accuracy than ours, respectively; however, our approach is distinct in achieving $\Delta_{\text{DemP}} \approx 0$ with salary accuracy $\geq 79\%$.

From Fig. 3a, we see that gender accuracy saturates at 67% while achieving a salary accuracy of at least 81% for a specific value of distortion bound D, and therefore, test distortion; in turn, this choice of D corresponds in Fig. 3b to $\Delta_{\text{DemP}} \approx 0.06$. Further reducing Δ_{DemP} requires further increasing D, thus lowering the salary accuracy to 79% for $\Delta_{\text{DemP}} \approx 0$. This is because classification accuracy captures an average measure of correctness and is dominated by the performance over the majority class. On the other hand,

[7] The baseline performances are the salary and gender accuracies as well as Δ_{DemP} obtained from the original uncensored test dataset.

Authorized licensed use limited to: ASU Library. Downloaded on June 10,2022 at 04:26:01 UTC from IEEE Xplore. Restrictions apply.

Δ_{DemP} captures the difference in performance of the intended classifier on each of the two classes. Thus, enforcing fairness via Δ_{DemP} reduces salary accuracy thereby highlighting the tradeoff between guaranteeing fairness and utility.

We can also evaluate the fairness performance of the generated X_r by using the EO measure in Definition 1. Thus, for $Y \in \{0, 1\}$ where $Y = 1$ when salary $> 50K$, $S \in \{0, 1\}$ (female:1 and male:0), and $\hat{Y} \in \{0, 1\}$, we write $\Delta_{\text{EO}}(y), \forall y \in \{0, 1\}$ as:

$$\Delta_{\text{EO}}(y) \triangleq \left| P(\hat{Y}=y|S=0, Y=y) - P(\hat{Y}=y|S=1, Y=y) \right|. \tag{17}$$

Note that for binary Y, as is the case here, (17) is the same as the definition of EO in (2). From Fig. 4, which plots salary accuracy vs. DemP or EO measures of fairness, we observe that while the salary accuracy is above 82.4%, the values of $\Delta_{\text{EO}}(1)$ and $\Delta_{\text{EO}}(0)$ decrease to 0.0007 and 0.0254, respectively. To understand the significance of these results, we compare against the state-of-the-art in [11], wherein fair salary classifiers for both DemP and EO measures, referred to as LAFTR-DP[8] and LAFTR-EO, respectively, are learned for the UCI dataset. For the LAFTR-DP, the authors also compute the resulting EO of the DemP classifier. As a preamble to the following comparisons, we note that fair predictors, trained on specific tasks, will do at least as well as the same predictors learned on fair representations.

We make the following observations: (i) when $\Delta_{\text{EO}}(1) + \Delta_{\text{EO}}(0) = 0.04$,[9] our salary accuracy is 1.3% smaller than that achieved by LAFTR-DP (cf. Fig. 2(b) in [11]), but our minimal achieved value of $\Delta_{\text{EO}}(1) + \Delta_{\text{EO}}(0)$ is only 72% of that achieved by LAFTR-DP and is the same as the value achieved by LAFTR-EO, which uses EO as the fairness metric to train a salary classifier; (ii) the decrease of $\Delta_{\text{EO}}(1) + \Delta_{\text{EO}}(0)$ is even larger than Δ_{DemP}. That is, even though the representation is generated to satisfy DemP, it can also provide competitive downstream EO fairness guarantees. This, in turn, justifies the rationality of generating fair representations under DemP.

2) Case II: Non-Binary Sensitive Feature: Figs. 5 and 6 illustrate the censoring and fairness performances of the generated X_r in hiding 'gender' and 'relationship', respectively, while preserving 'salary' information. Fig. 5 illustrates the tradeoff between salary and sensitive feature S accuracies when S is either gender, or relationship, or both. From Fig. 5, we observe that while the salary accuracy is above 79%, the classification accuracies of gender and/or relationship are about 66% (Fig. 5a), 45% (Fig. 5b) and 41% (Fig. 5c), respectively. Note that the probabilities of male, husband, and the combination (male, husband) are 66%, 40% and 40%, respectively, in the original test data. Therefore, while the salary accuracy is preserved at 79%, the inferences of gender, relationship, and combination (gender, relationship) approach random guessing with these priors. Thus, our FUR framework can effectively hide one or more sensitive features. However, suppressing multiple correlated sensitive features comes at a cost of a reduction in salary accuracy. Thus, comparing Figs. 3a and 5a, we see a maximal reduction of 3% in salary accuracy for a given gender accuracy.[10]

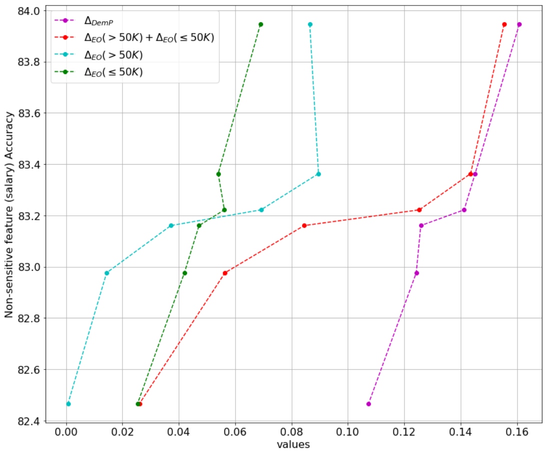


Fig. 4. Evaluation of equalized odds fairness metric under Case I for the UCI Adult dataset. The EO measures $\Delta_{\text{EO}}(1)$ and $\Delta_{\text{EO}}(0)$ are defined in (17). The red curve plotting $\Delta_{\text{EO}} = \Delta_{\text{EO}}(1) + \Delta_{\text{EO}}(0)$ matches Δ_{EO} in Figure 2(b) of [11]. Each point corresponds to a specific value of achieved test distortion ranging over (0.59, 2.01), with distortion decreasing from the left to the right for each plot.

For Case II, Figs. 6a and 6b illustrate the tradeoffs between the salary accuracy and Δ_{DemP} for S chosen as gender or relationship or both. We observe that while salary accuracy is above 94% of the baseline performance, the value of Δ_{DemP} is dropped to 25% for gender and to about 34% for both relationship and their combination. In short, X_r works well in decorrelating gender and relationship both separately and jointly without affecting downstream classifier performance. From Fig. 6b, we observe that the value of Δ_{DemP} for the combination is almost the same as that for relationship. In addition, comparing the results in Figs. 3b and 6a, for any given Δ_{DemP} for gender, the salary accuracy in Case II is about 1% lower than that in Case I; this can be viewed as the cost of eliminating a potentially sensitive feature (relationship) that is also correlated with the target feature (see also, footnote 10). Finally, comparing the results in Figs. 3b and 6b, for any given salary accuracy, Δ_{DemP} for gender in Case II is about 0.25 higher than that in Case I; this can be viewed as the effect of using non-binary sensitive features on Δ_{DemP}, now defined as the maximum over all values taken by the non-binary sensitive feature.

B. Illustration of Results for UTKFace Dataset

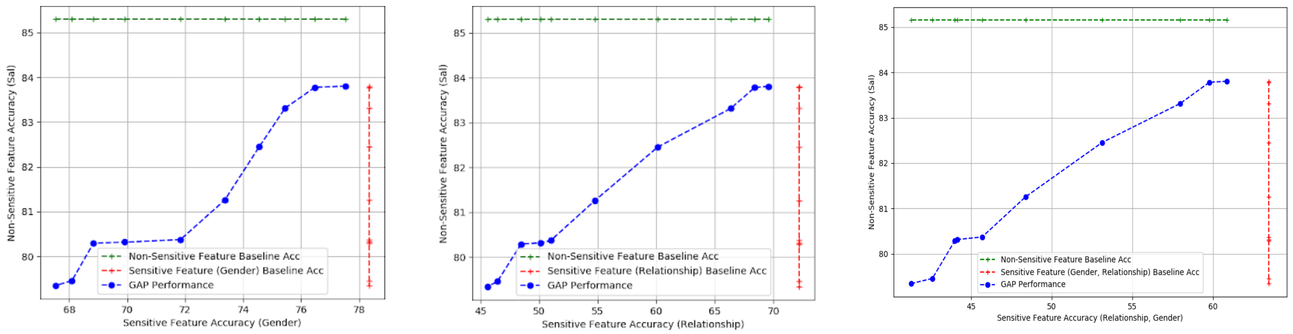
In the UTKFace dataset, the face images are the non-sensitive X. We choose 'gender' as the sensitive S; focusing on multiple downstream tasks, we consider both ethnicity classification and age regression, for which we choose 'ethnicity' or 'age' as the target variable Y, respectively. For the two downstream applications, the corresponding supports of Y are $\mathcal{Y} = \{\text{White, Black, Asian, South Asian Indian}\}$ and $\mathcal{Y} = \{i \in \mathbb{Z} : 10 \leq i \leq 65\} = [10, 65]$, respectively. We use the maximum of the DemP measure (defined in (16)) over the support Y, i.e., $\Delta_{\text{DemP}} = \max_{y \in \mathcal{Y}} \Delta_{\text{DemP}}(y)$, as the achieved fairness level.

[8] Learned Adversarially Fair and Transferable Representations (LAFTR)

[9] [11] introduced an EO measure as $\Delta_{\text{EO}} \triangleq \Delta_{\text{EO}}(1) + \Delta_{\text{EO}}(0)$

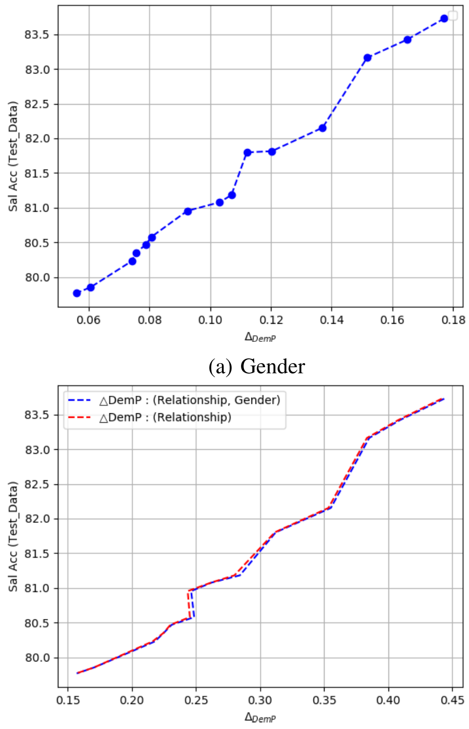
[10] In Figs., 3a and 5a, the baseline performances are different because for Case II, the feature variable X does not contain 'relationship'.

Authorized licensed use limited to: ASU Library. Downloaded on June 10,2022 at 04:26:01 UTC from IEEE Xplore. Restrictions apply.



(a) Gender (b) Relationship (c) (Gender, Relationship)

Fig. 5. Tradeoff between classification accuracy of non-sensitive feature (salary) and sensitive features (gender and/or relationship) under Case II for the UCI Adult dataset. Note that we use the classification accuracy obtained from the original testing dataset as the baseline performance, which is denoted by the green and red lines for the target variable (salary) and the sensitive variable (gender or/and relationship), respectively. In every plot, each point corresponds to a specific value of achieved test distortion (over all features except gender and relationship) ranging over (0.58, 2.1), with distortion decreasing from the left to the right for each plot.



(a) Gender

(b) Relationship (or) (Relationship, Gender)

Fig. 6. Case II for UCI Adult: Tradeoffs between salary accuracy and the Δ_{DemP} of gender and/or relationship. For the original test data, Δ_{DemP} for gender, relationship and the pair (gender, relationship) is 0.2, 0.438 and 0.443, respectively. In every plot, each point corresponds to a specific value of achieved test distortion (over all features except gender and relationship) ranging over (0.58, 2.1), with distortion decreasing from the left to the right for each plot. .

Fig. 7 illustrates the output X_r for 16 typical[11] faces in the UTKFace dataset for increasing per-pixel distortion. From Fig. 7, we observe that: (i) for a small per-pixel distortion (e.g., 0.003), gender-distinguishing features such as lip color are smoothed out; and (ii) at higher per-pixel distortion (e.g., 0.006), the FUR framework can generate a face with an opposite gender (see the highlighted examples in Fig. 7) thereby completely obfuscating this sensitive feature; (iii) when the average per-pixel distortion is too large (e.g., 0.01), the representations generated are often too blurred.

Figs. 8a and 9 show the tradeoffs between gender classification accuracy and appropriate measures for ethnicity classification and age regression, respectively. In Fig. 8a, while gender classification accuracy is about 62% and decreases about 35% from the baseline performance, the classification accuracy of ethnicity is above 74% and only decreases 14% from its baseline performance. Note that in the original testing data, the highest marginal probabilities for gender and ethnicity are 54.6% (likelihood of male) and 43.2% (likelihood of White), respectively. Therefore, gender accuracy is better than a random guess by only 7.4% while ethnicity accuracy is better than a random guess by 30.8%, i.e., the generated X_r hides gender information well while maintaining ethnicity. For age regression, we use the mean absolute error (MAE), i.e., the average absolute difference between the predicted age and the true age, as the utility measure. In Fig. 9a, we observe that while the classification accuracy for gender is about 62%, which is a 35% decrease from the baseline performance of 94%, the increase in the MAE is 1.5 which is about a 20% increase from the baseline performance of 7.2 years. Fig. 9b shows the cumulative distribution function (CDF) of the difference between the true and predicted age for various distortions, from which we can see that the drop of the cumulative probability is at most 1%. Thus, the generated FUR guarantees reliable performance for both age and ethnicity prediction; thus, constraining the distortion of the generated X_r can be effective in guaranteeing utility for multiple tasks.

In Figs. 8b and 10, we illustrate the tradeoff between the utility measure and Δ_{DemP} of the generated X_r in ethnicity classification and age regression, respectively. In Fig. 8b, we observe that while achieving about 86% of the baseline classification accuracy, the Δ_{DemP} is reduced to 0.03, which is 20% of the $\Delta_{\text{DemP}} = 0.14$ in the original testing data. Table II shows the decrease of Δ_{DemP} for each of the four ethnicities as the distortion increases. In Fig. 10a, while preserving 86% of the utility baseline performance, the Δ_{DemP}, i.e., the maximal value of demographic parity measure over the 56 age

[11] The 16 typical faces covers the 8 possible combination of 2 gender (male and female) and 4 ethnicities (White, Black, Asian and South Asian Indian) and includes young, adult and old faces.

Authorized licensed use limited to: ASU Library. Downloaded on June 10,2022 at 04:26:01 UTC from IEEE Xplore. Restrictions apply.

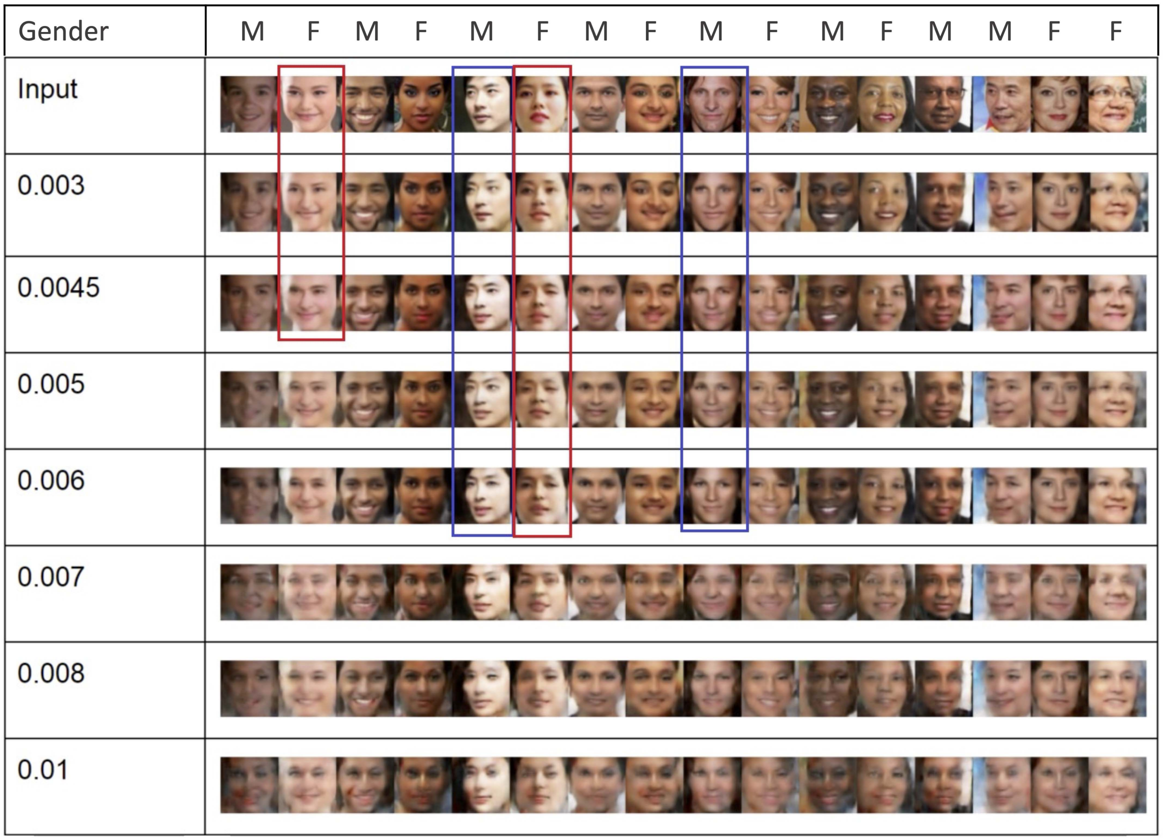


Fig. 7. The encoded face images for different values of per-pixel distortions for the UTKFace dataset. Set of vertical faces highlighted in boxes makes explicit how the sensitive feature (gender) is changed with increasing distortion. The ground truth gender values for the images are shown in the top-most row.

TABLE II

Demographic Parity Fairness (Indicated by $\Delta_{\mathrm{DEMP}}(\cdot)$) of Ethnicity Classification on the UTKFace Dataset

Distortion	0	0.003	0.0045	0.005	0.006	0.007	0.008	0.01
Δ_{DemP}(White)	0.061	0.055	0.04	0.03	0.03	0.02	0.02	0.01
Δ_{DemP}(Black)	0.109	0.021	0.02	0.05	0.03	0.05	0.03	0.03
Δ_{DemP}(Asian)	0.14	0.082	0.07	0.07	0.06	0.07	0.06	0.03
Δ_{DemP}(Indian)	0.031	0.006	0.01	0	0.01	0	0.01	0.01

values, decreases to 0.015, which is less than 33% of the $\Delta_{\mathrm{DemP}} = 0.046$ in the original testing data. Fig. 10b shows the demographic measure $\Delta_{\mathrm{DemP}}(y)$, $y \in [10, 65]$, for various distortions; we observe that when the pixel distortion is 0.01, even while $\Delta_{\mathrm{DemP}} = 0.015$, $\Delta_{\mathrm{DemP}}(y) = 0$ for 17 distinct ages. That is, the predictions of these 17 ages are completely independent of gender and DemP is achieved for those predictions.

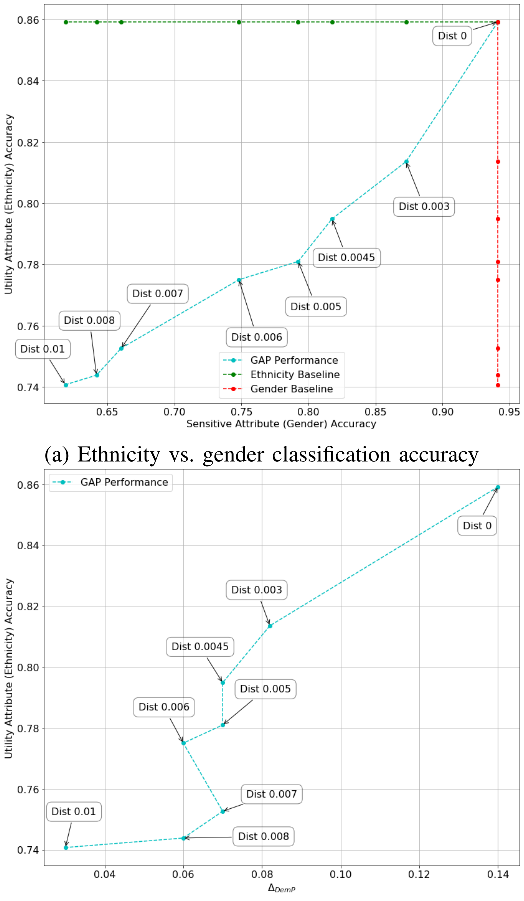
C. Illustration of Results for GENKI Dataset

For the GENKI dataset, we consider the following two approaches to decorrelating the data (X, S): the feedforward neural network decorrelator (FNND) and the transposed convolution neural network decorrelator (TCNND). Specific architectural details for both can be found in the supplement. Fig. 11a illustrates the gender classification accuracy of the adversary for different values of distortion. It can be seen that the adversary's accuracy of classifying S (gender) decreases as the distortion increases. Given the same distortion value, the FNND achieves lower gender classification accuracy compared to the TCNND. An intuitive explanation for this is that the FNND uses both the noise vector and the original image to generate the processed image, while the TCNND generates the noise mask independently of the original image and then adds this mask to the original image in the final step.

1) Censoring vs. Differential Privacy Guarantees: For this dataset, in addition to highlighting the role of GAN-like architectures to learn fair representations, we also explore the effect of censoring on assuring privacy of the sensitive (here, gender) feature. Differential privacy (DP) has emerged as the gold standard for data privacy [40]. Thus, one way to censor and privatize data is to add noise with differential privacy guarantees [40]. Since the dataset is continuous valued, we consider two types of additive DP noise mechanisms at the pixel level: Gaussian and Laplacian. We vary the variances of the Laplace and Gaussian noise to then obtain a specific local DP guarantee[12] building on [41]. We compute the resulting

[12] DP, by definition, guarantees that the output of a differentially private (randomizing) mechanism cannot aid in distinguishing between two neighboring (defined appropriately) datasets. Local DP is stronger than DP in that it provides such a guarantee for any pair of inputs.

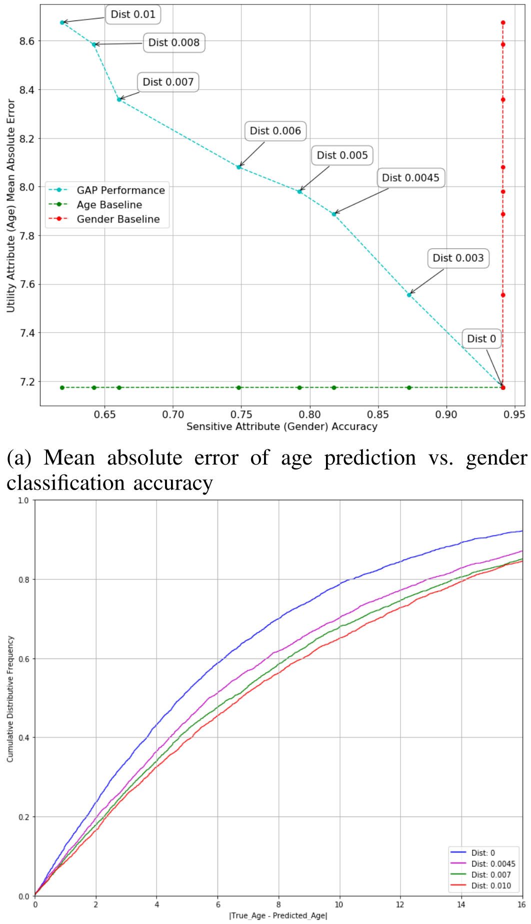
Authorized licensed use limited to: ASU Library. Downloaded on June 10,2022 at 04:26:01 UTC from IEEE Xplore. Restrictions apply.



(a) Ethnicity vs. gender classification accuracy

(b) Ethnicity classification accuracy vs. Δ_{DemP}

Fig. 8. Ethnicity classification accuracy vs. gender classification and Δ_{DemP} for the UTKFace dataset. In Fig. 8b, the x-axis is the maximal value of DemP in (16) over the four ethnicities and 'dist' indicates the per pixel distortion.



(a) Mean absolute error of age prediction vs. gender classification accuracy

(b) The CDF of the difference between the true and predicted age

Fig. 9. Utility of age regression on the UTKFace dataset. Note that 'dist' indicates the per pixel distortion.

DP guarantees provided by independent Laplace and Gaussian noise-adding mechanisms for different distortion values. The details are provided in the supplement (see Section SII-B). In Fig. 11a, we compare the gender classification accuracy of the learned FUR schemes with those obtained by adding Laplace or Gaussian noise. We see that for the same distortion, the learned FUR schemes achieve much lower gender classification accuracy. In Table III, we observe that even when a large amount of noise is added to each pixel, the privacy risk (ϵ) is still significantly high. Furthermore, such noise levels deteriorate the expression classification accuracy (cf. Fig. 11a). It is worth noting that the distortion constraint for the FUR framework is an average over the entire image.

2) Evaluating Adversarial Performance via Mutual Information Estimation: Our FUR framework offers a scalable way to find a (local) equilibrium in the constrained min-max optimization for certain adversarial attacks (e.g. inference attacks on S using a neural network). Yet the privatized data, through our approach, should be immune to any general attacks and should ultimately achieve the goal of decreasing the correlation between the X_r and S. To this end, we use the estimated MI, using the k-nearest neighbor method as detailed in Section IV-C, to verify that our framework protects S.

One noteworthy difficulty is that X, and therefore, X_r are usually high dimensional objects (each image in the GENKI dataset has 256 dimensions), so it is almost impossible to calculate the empirical entropy based on raw data due to the sample complexity. Thus, we train a neural network that classifies S from the learned data representation to reduce the dimension of the data. We choose the layer before the softmax outputs of the adversary (denoted by $\hat{X}_g$) to be the feature embedding that has a much lower dimension than the original X_r which still captures the information about S. We use this $\hat{X}_g$ as a surrogate for X_r in (15) to first estimate $\hat{H}(\hat{X}_g)$ and then compute $\hat{I}(\hat{X}_g; S)$ as the approximate MI between X_r and S. We similarly extract an $\hat{X}_f$ by training a neural network that now classifies Y; we then compute $\hat{H}(\hat{X}_f)$ from which we obtain $\hat{I}(\hat{X}_f; Y)$ as the approximate MI between X_r and Y. The details of the common neural network architecture we use to extract $\hat{X}_f$ and $\hat{X}_g$ can be found in the supplement (see Section SII-C).

3) Utility-Fairness Tradeoffs: To evaluate the value of the representation generated by the FUR framework for the GENKI dataset, we consider the task of classifying the facial expression (non-sensitive feature Y) as smiling or non-smiling (i.e., the task that the GENKI dataset was intended for). To this

Authorized licensed use limited to: ASU Library. Downloaded on June 10,2022 at 04:26:01 UTC from IEEE Xplore. Restrictions apply.

TABLE III
DIFFERENTIAL PRIVACY RISK FOR DIFFERENT DISTORTION VALUES

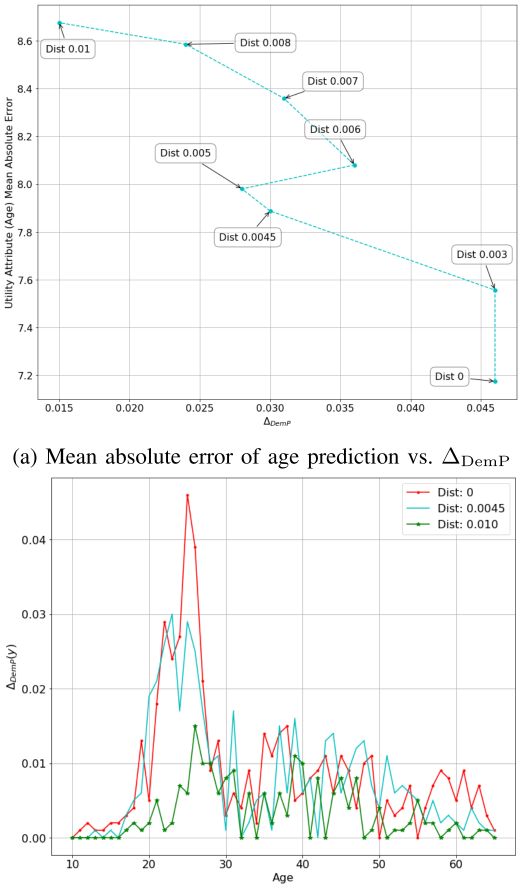
Distortion D	1	2	3	4	5	100	1000
Laplace Mechanism ϵ	5792.61	4096	3344.36	2896.31	2590.53	579.26	183.17
Gaussian Mechanism ($\delta = 10^{-6}$) ϵ	1918.24	1354.08	1107.57	959.18	857.76	191.82	60.66

TABLE IV
ERROR RATES FOR EXPRESSION CLASSIFICATION USING REPRESENTATION LEARNED BY FNND FOR THE GENKI DATASET

Expression	Original Data		D = 1		D = 3		D = 5	
Classification	Male	Female	Male	Female	Male	Female	Male	Female
False Positive Rate	0.04	0.14	0.1	0.18	0.18	0.16	0.16	0.14
False Negative Rate	0.16	0.02	0.2	0.08	0.26	0.12	0.24	0.24

TABLE V
ERROR RATES FOR EXPRESSION CLASSIFICATION USING REPRESENTATION LEARNED BY TCNND FOR THE GENKI DATASET

Expression	Original Data		D = 1		D = 3		D = 5	
Classification	Male	Female	Male	Female	Male	Female	Male	Female
False Positive Rate	0.04	0.14	0.04	0.16	0.06	0.12	0.08	0.16
False Negative Rate	0.16	0.02	0.2	0.08	0.2	0.14	0.18	0.16

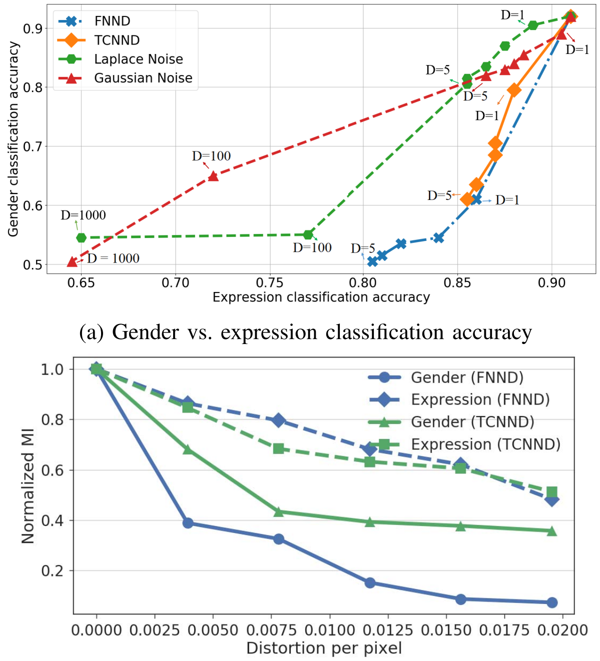


(a) Mean absolute error of age prediction vs. Δ_{DemP}

(b) The demographic parity measure for various distortions

Fig. 10. Achieved demographic parity for the age regression task on the UTKFace dataset. Note that in Fig. 10a, the x-axis is the maximal value of DemP in (16) over the chosen age range (10-65) and 'dist' indicates the per pixel distortion.

end, we train another CNN (see Fig. S4 in the supplement for architecture details) to perform facial expression classification on datasets processed by different decorrelation schemes.



(a) Gender vs. expression classification accuracy

(b) Normalized mutual information estimation

Fig. 11. The tradeoff between classification accuracy and mutual information estimation for the GENKI dataset.

The trained model is then tested on the original test data. In Fig. 11a, we observe that the expression classification accuracy decreases gradually as the distortion increases. However, even for a large distortion value (5 per image), the expression classification accuracy only decreases by 10%. To make meaningful comparisons using MI, we normalize $\hat{I}(\hat{X}_f; Y)$ by $\hat{I}(X_f; Y)$ where X_f is the low-dimensional representation of X ($=X_r$ for $D = 0$); we similarly normalize $\hat{I}(\hat{X}_g; S)$ by $\hat{I}(X_g; S)$ where X_g is defined similarly. As shown in Fig. 11b, the estimated normalized MI $\hat{I}(\hat{X}_f; Y)/\hat{I}(X_f; Y)$ decreases at a much slower rate than $\hat{I}(\hat{X}_g; S)/\hat{I}(X_g; S)$ as the distortion increases thus verifying that X_r preserves relatively

Authorized licensed use limited to: ASU Library. Downloaded on June 10,2022 at 04:26:01 UTC from IEEE Xplore. Restrictions apply.

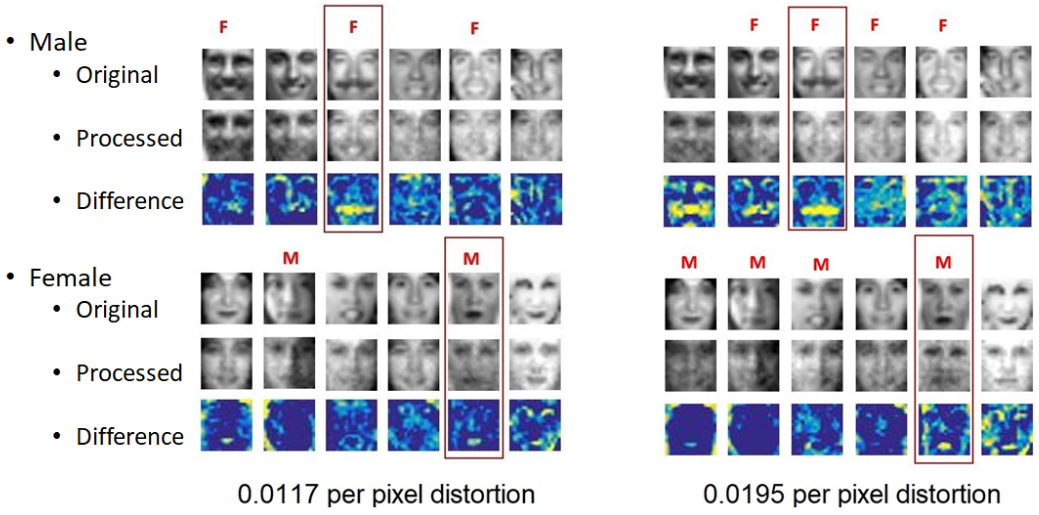
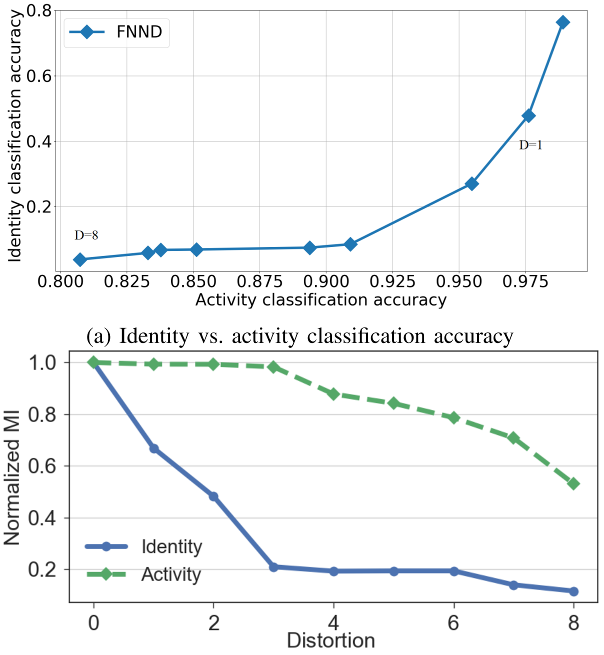


Fig. 12. Perturbed images with different per pixel distortion using FNND.



(a) Identity vs. activity classification accuracy

(b) Normalized mutual information estimation

Fig. 13. The tradeoff between classification accuracy and mutual information estimation for the HAR dataset.

more information about Y than it does about S at every distortion level.

Tables IV and V present different error rates for the facial expression classifiers trained using data representations created by different decorrelator architectures. For the GENKI dataset, a smiling expression is considered as the positive label for the expression classification task. We observe that as distortion increases, the difference across the two sensitive groups (male vs. female) for each error rate decreases. This implies the classifier's decision is less biased with respect to S when trained using X_r. When $D = 5$, the differences are quite small. In particular, the FNND architecture performs better in enforcing fairness but suffers from a slightly higher error rate relative to TCNND. The images processed by FNND are shown in Fig. 12. As highlighted in the figure, the dominant features that the decorrelator changes are those that capture gender, namely eyes, nose, mouth, beard, and hair.

D. Illustration of Results for HAR Dataset

For the HAR dataset, we recall that the X features are motion sensor data from 30 subjects performing six different activities; the goal is to classify activity (intended task Y) without revealing subject identity (sensitive S). In Fig. 13a, for different values of distortion, we illustrate the accuracy in classifying activity against that of identification. We see that classification accuracy of identity decreases as the distortion increases. In fact, when distortion is small ($D = 2$), the identity accuracy is down to 27%. If we increase the distortion to 8, the identity accuracy further decreases to 3.8%. However, we note that for an even large distortion value ($D = 8$), the activity classification accuracy only decreases by at most 18%. Finally, in Fig. 13b, we demonstrate that the estimated normalized mutual information $\hat{I}(\hat{X}_f; Y)/\hat{I}(X_f; Y)$ decreases at a much slower rate than $\hat{I}(\hat{X}_g; S)/\hat{I}(X_g; S)$ as the distortion increases, thereby assuring that the adversarial model chosen assures censoring of identity without compromising accuracy of the intended task from X_r. Details on the architecture for the FUR model, the classifier, and for obtaining the representations used to estimate MI can be found in the supplement.

VI. Conclusion

We have introduced an adversarial learning framework with verifiable guarantees for learning generative models that can create censored and fair universal representations for datasets with known sensitive features. The novelty of our approach is in producing representations that are fair with respect to the sensitive features for any *a priori* unknown downstream learning task. We have shown that our FUR framework allows the data holder to learn the fair encoding scheme (a randomized mapping that decorrelates the sensitive and non-sensitive features) directly from the dataset without requiring access to dataset statistics. One of our key results is

Authorized licensed use limited to: ASU Library. Downloaded on June 10,2022 at 04:26:01 UTC from IEEE Xplore. Restrictions apply.

that for appropriately chosen loss functions, the minimax game generating FURs can provide guarantees against strong information-theoretic adversaries, such as MAP (0-1 loss) and MI (log-loss) adversaries. We have also shown that our framework also allows approaching (ideal) demographic parity fairness using a tunable class of α-loss functions (which includes both log-loss and 0-1 loss) that capture a range of adversarial actions. For the setting with a known classification task, we have also shown that our FUR framework can be modified to approximate either DemP or EO fairness measures.

Finally, we have also validated the performance of the FUR framework on both synthetic and publicly available real datasets, including Gaussian mixture models, images, and datasets involving a mix of categorical and continuous features. Our results have allowed us to visually highlight three key results: (a) the tradeoff between representation fidelity and censoring guarantees (via accuracy in adversarially learning sensitive features); (b) the tradeoff between the adversarial accuracy in learning the sensitive features and the accuracy of multiple downstream tasks learned from the censored representation for a variety of datasets; and (c) the tradeoff between accuracy for a downstream learning task and the DemP or EO guarantees achieved by a predictor that is learned using a DemP fair representation. Result (c) further suggests that, for some datasets, DemP FURs do not adversely affect the downstream EO guarantees despite lack of access to the task labels Y, where the latter has been highlighted as a limitation of DemP fair predictors [8]. However, more work is needed to understand the conditions under which DemP FURs suffice to achieve meaningful EO guarantees. Such an exploration can also clarify if the definition of Δ_{DemP} as the worst case over all task output values $y \in \mathcal{Y}$ and sensitive values $s \in \mathcal{S}$ is too strong, especially for settings where $\mathcal{Y}$ and/or $\mathcal{S}$ are large, as our age prediction results suggest for the UTKFace dataset.

Going beyond this work, there are several questions that can be addressed. An immediate one is to explore the robustness of using α-loss (for $\alpha \neq 1$) in ensuring censoring and fairness for highly imbalanced datasets. Yet another is to explore the usefulness of FURs for unsupervised tasks building on recent work in [42]. More broadly, it will be interesting to investigate the robustness and convergence guarantees of the generative encoder learned in a data-driven fashion.

Appendix

A. Alternate Minimax Algorithm

Algorithm 1 details the steps used to learn the FUR model in a data-driven manner. To incorporate the distortion constraint, we use the *penalty method* [21] to replace the constrained optimization problem by a series of unconstrained problems. The unconstrained optimization problem is formed by adding a penalty to the objective function as a product of a parameter ρ_t and an appropriate measure of violation of the constraint. We start with a large value of ρ_t to enforce distortion from the outset and decrease ρ_t in exponential steps with respect to the number of training epochs. Such a decrease allows enforcing a smaller penalty when the model is closer to convergence. Finally, we also vary the learning rate η_t over training epochs as follows: we pick a small value of η_t at the beginning and compare the relative values of the adversarial loss and the average distortion. We adjust the initial η_t so that the adversarial loss and the distortion penalty values are on a similar scale in the first few epochs during training. When the algorithm terminates, we check the average distortion and manually fine tune the initial η_t and the update rule to make sure the distortion is within bounds after termination.

Algorithm 1 Alternating Minimax FUR Algorithm

```
Input: dataset $\mathcal{D}$, distortion parameter $D$, # of decorrelator iterations $T$, # of adversary iterations $J$ for each round of decorrelator update, minibatch size $M$
Output: Optimal generative decorrelator parameter $\theta_p$
procedure ALERNATE MINIMAX($\mathcal{D}, D, T, J, M$)
    Initialize decorrelator parameter $\theta_p^1$, adversary parameter $\theta_a^1$, and step size $\eta_1$
    for $t = 1, \ldots, T$ do
        Random minibatch of $M$ datapoints $\{x_{(1)}, \ldots, x_{(M)}\}$ drawn from full dataset
        Generate $\{\hat{x}_{(1)}, \ldots, \hat{x}_{(M)}\}$ via $\hat{x}_{(i)} = g(x_{(i)}; \theta_p^t)$
        Apply update rule for step size $\eta_t$
        Set $\omega_a^1 = \theta_a^t$
        for $j = 1, \ldots, J$ do
            Update the adversary parameter $\theta_a^{t+1}$ by stochastic gradient ascent for epoch $j$
            $\omega_a^{j+1} = \omega_a^j + \eta_t \nabla_{\omega_a^j} \frac{1}{M} \sum_{i=1}^{M} -\ell(h(\hat{x}_{(i)}; \omega_a^j), s_{(i)}), \quad \eta_t > 0$
        Set $\theta_a^{t+1} = \omega_a^{J+1}$
        Compute the descent direction $\nabla_{\theta_p^t} L_m(\theta_p^t, \theta_a^{t+1})$, where $L_m(\theta_p^t, \theta_a^{t+1})$ is defined in (11) for $n = m$
        Perform line search along $\nabla_{\theta_p^t} L_m(\theta_p^t, \theta_a^{t+1})$ and, for $\ell(\theta_p^t, \theta_a^{t+1})$ set as the objective in (13) for $n = m$, update
            $\theta_p^{t+1} = \theta_p^t - \eta_t \nabla_{\theta_p^t} \ell(\theta_p^t, \theta_a^{t+1})$
    return $\theta_p^{T+1}$
```

B. Proof of Theorem 5

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 $\|\beta\|^2 + tr(\Sigma_p) \leq D$. Let us consider $\hat{X} = X + Z + \beta$, where $\beta \in \mathbb{R}^m$ and Σ_p is a diagonal covariance whose diagonal entries are given by $\{\sigma_{p_1}^2, \ldots, \sigma_{p_m}^2\}$. Given the MAP adversary's optimal inference accuracy in (14), the objective of the decorrelator is

$$\min_{\beta, \Sigma_p} P_{\text{d}}^{(\text{G})}, \text{ s.t. } \|\beta\|^2 + tr(\Sigma_p) \leq D. \tag{18}$$

Define $\frac{1}{\gamma} \ln \frac{1-q}{q} = \eta$. After some algebra, the gradient of $P_{\text{d}}^{(\text{G})}$ *w.r.t.* γ is given by

$$\frac{\partial P_{\text{d}}^{(\text{G})}}{\partial \gamma} = \frac{1}{2\sqrt{2\pi}} \left(q e^{-\frac{\left(\eta - \frac{\gamma}{2}\right)^2}{2}} + (1-q) e^{-\frac{\left(\eta + \frac{\gamma}{2}\right)^2}{2}} \right),$$

Authorized licensed use limited to: ASU Library. Downloaded on June 10,2022 at 04:26:01 UTC from IEEE Xplore. Restrictions apply.

which is always positive. Thus, $P_{\mathrm{d}}^{(\mathrm{G})}$ is monotonically increasing in γ. As a result, (18) is equivalent to

$$\begin{aligned} \min_{\beta,\sigma_{p_1}^2,\ldots,\sigma_{p_m}^2} \quad & \sum_{i=1}^{m} \frac{\mu_i^2}{\sigma_i^2+\sigma_{p_i}^2}, \\ \text{s.t. } & \|\beta\|^2 + tr(\Sigma_p) \le D \\ & \sigma_{p_i}^2 \ge 0 \quad \forall i \in \{1, 2, \ldots m\}. \end{aligned} \tag{19}$$

The optimization in (19) is analogous to the well-studied rate-distortion problem for independent Gaussian sources and the optimal solution given by reverse water-filling [43, Chap. 10.3.3]. Using similar techniques, we obtain $\sigma_{p_i}^{*\,2} = \max\{|\mu_i|/\sqrt{\lambda_0^*} - \sigma_i^2, 0\} = \left(|\mu_i|/\sqrt{\lambda_0^*} - \sigma_i^2\right)^{+}$ with $\sum_{i=1}^{m} \sigma_{p_i}^{*\,2} = D$ leading to the results in the theorem.

Acknowledgment

The authors would like to thank Prof. Ram Rajagopal and Xiao Chen at Stanford for their help with estimating mutual information for the GENKI and HAR datasets. They would also like to thank Mit Patel and Harshit Laddha at ASU for their help with the experiments on the UTKFace dataset.

References

[1] H. F. Ladd, "Evidence on discrimination in mortgage lending," *J. Econ. Perspect.*, vol. 12, no. 2, pp. 41–62, May 1998.

[2] D. Pedreshi, S. Ruggieri, and F. Turini, "Discrimination-aware data mining," in *Proc. 14th ACM SIGKDD*, 2008, pp. 560–568.

[3] C. Song and V. Shmatikov, "Overlearning reveals sensitive attributes," 2019, *arXiv:1905.11742*.

[4] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *Proc. 21st ACM SIGKDD*, Aug. 2015, pp. 259–268.

[5] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proc. 3rd Innov. Theor. Comput. Sci. Conf.*, 2012, pp. 214–226.

[6] F. D. P. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney, "Data pre-processing for discrimination prevention: Information-theoretic optimization and analysis," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 5, pp. 1106–1119, Oct. 2018.

[7] M. Kearns, S. Neel, A. Roth, and Z. S. Wu, "Preventing fairness gerrymandering: Auditing and learning for subgroup fairness," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2018, pp. 2564–2572.

[8] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Proc. NeurIPS*, 2016, pp. 3315–3323.

[9] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big Data*, vol. 5, no. 2, pp. 153–163, Jun. 2017.

[10] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proc. IEEE/ACM AIES*, Dec. 2018, pp. 335–340.

[11] D. Madras, E. Creager, T. Pitassi, and R. Zemel, "Learning adversarially fair and transferable representations," in *Proc. ICML*, 2018, pp. 3384–3393.

[12] H. Edwards and A. J. Storkey, "Censoring representations with an adversary," in *Proc. ICLR*, 2016, pp. 1–14.

[13] F. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney, "Optimized pre-processing for discrimination prevention," in *Proc. NeurIPS*, 2017, pp. 3992–4001.

[14] S. Hajian, J. Domingo-Ferrer, A. Monreale, D. Pedreschi, and F. Giannotti, "Discrimination- and privacy-aware patterns," *Data Mining Knowl. Discovery*, vol. 29, no. 6, pp. 1733–1782, Nov. 2015.

[15] D. Wei, K. N. Ramamurthy, and F. Calmon, "Optimized score transformation for fair classification," in *Proc. AISTATS*, 2020, pp. 1673–1683.

[16] J. Hamm, "Minimax filter: Learning to preserve privacy from inference attacks," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 4704–4734, 2017.

[17] C. Huang, P. Kairouz, X. Chen, L. Sankar, and R. Rajagopal, "Context-aware generative adversarial privacy," *Entropy*, vol. 19, no. 12, p. 656, Dec. 2017.

[18] A. Tripathy, Y. Wang, and P. Ishwar, "Privacy-preserving adversarial networks," 2017, *arXiv:1712.07008*.

[19] M. Bertran *et al.*, "Adversarially learned representations for information obfuscation and inference," in *Proc. ICML*, 2019, pp. 614–623.

[20] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. NeurIPS*, 2014, pp. 2672–2680.

[21] W. E. Lillo, M. H. Loh, S. Hui, and S. H. Zak, "On solving constrained optimization problems with neural networks: A penalty method approach," *IEEE Trans. Neural Netw.*, vol. 4, no. 6, pp. 931–940, Nov. 1993.

[22] *Google*. Accessed: Jun. 2021. [Online]. Available: https://github.com/google-research/tensorflow_constrained_optimization/blob/master/README.md

[23] R. Kohavi, "Scaling up the accuracy of Naive–Bayes classifiers: A decision-tree hybrid," in *Proc. KDD*, vol. 96, 1996, pp. 202–207.

[24] Z. Zhang, Y. Song, and H. Qi, "Age progression/regression by conditional adversarial autoencoder," in *Proc. CVPR*, 2017, pp. 5810–5818.

[25] J. Whitehill and J. Movellan, "Discriminately decreasing discriminability with learned image filters," in *Proc. CVPR*, Jun. 2012, pp. 2488–2495.

[26] D. Anguita *et al.*, "A public domain dataset for human activity recognition using smartphones," in *Proc. ESANN*, 2013, pp. 437–442.

[27] E. Creager *et al.*, "Flexibly fair representation learning by disentanglement," in *Proc. ICML*, vol. 97, 2019, pp. 1436–1445.

[28] *Generating Fair Universal Representations Using Adversarial Models*. Accessed: Mar. 2022. [Online]. Available: https://github.com/SankarLab/Fair-Universal-Representations-Code-Papers

[29] J. Liao, O. Kosut, L. Sankar, and F. D. P. Calmon, "Tunable measures for information leakage and applications to privacy-utility tradeoffs," *IEEE Trans. Inf. Theory*, vol. 65, no. 12, pp. 8043–8066, Dec. 2019.

[30] T. Sypherd, M. Diaz, J. K. Cava, G. Dasarathy, P. Kairouz, and L. Sankar, "A tunable loss function for robust classification: Calibration, landscape, and generalization," 2019, *arXiv:1906.02314*.

[31] S. Asoodeh, M. Diaz, F. Alajaji, and T. Linder, "Estimation efficiency under privacy constraints," *IEEE Trans. Inf. Theory*, vol. 65, no. 3, pp. 1512–1534, Mar. 2019.

[32] J. Song, P. Kalluri, A. Grover, S. Zhao, and S. Ermon, "Learning controllable fair representations," in *Proc. AISTATS*, 2019, pp. 2164–2173.

[33] J. Eckstein and W. Yao, "Augmented Lagrangian and alternating direction methods for convex optimization: A tutorial and some illustrative computational results," RUTCOR, Piscataway, NJ, USA, Res. Rep. 3, 2012, vol. 32.

[34] R. G. Gallager, *Stochastic Processes: Theory for Applications*. Cambridge, U.K.: Cambridge Univ. Press, 2013.

[35] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning* (Springer Series in Statistics). New York, NY, USA: Springer, 2001.

[36] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," 2016, *arXiv:1603.04467*.

[37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[38] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 69, no. 6, 2004, Art. no. 066138.

[39] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine," in *Proc. Int. Workshop Ambient Assist. Living*. Berlin, Germany: Springer, 2012, pp. 216–223.

[40] C. Dwork, "Differential privacy: A survey of results," in *Theory and Applications of Models of Computation* (Lecture Notes in Computer Science). New York, NY, USA: Springer, Apr. 2008.

[41] P. Kairouz, S. Oh, and P. Viswanath, "Extremal mechanisms for local differential privacy," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 492–542, Jan. 2016.

[42] P. Lahoti, K. P. Gummadi, and G. Weikum, "IFair: Learning individually fair data representations for algorithmic decision making," in *Proc. IEEE 35th Int. Conf. Data Eng. (ICDE)*, Apr. 2019, pp. 1334–1345.

[43] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 2012.

Authorized licensed use limited to: ASU Library. Downloaded on June 10,2022 at 04:26:01 UTC from IEEE Xplore. Restrictions apply.