PRE- AND POST- FAIRNESS PROCESSING
FOR BLACK-BOX CLASSIFIERS

by

Xavier Gitiaux
A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
In Partial fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Computer Science

Committee:

Dr. Huzefa Rangwala, Dissertation Director

Dr. Sanmay Das, Committee Member

Dr. Dov Gordon, Committee Member

Dr. Brittany Johnson, Committee Member

Dr. Martin Slawski, Committee Member

Dr. David Rosenblum, Department Chair

Date: _____

Spring Semester 2022
George Mason University
Fairfax, VA

Pre- and Post- Fairness Processing for Black-Box Classifiers

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

By

Xavier Gitiaux
Master of Arts
University of Colorado, Boulder, 2011
Bachelor of Science
Ecole Polytechnique, France, 2005

Director: Dr. Huzefa Rangwala, Professor
Department of Computer Science

Spring Semester 2022
George Mason University
Fairfax, VA

Dedication

I dedicate this dissertation to Mariesa, Liam, Lena, Obie and Luciolle.

Acknowledgments

I would like to express my deepest and sincerest gratitude to my research advisor, Professor Huzefa Rangwala, for his support and guidance that made this PhD possible. His prescience that fairness would become an emerging and essential topic in machine learning is the reason I wrote an entire thesis in the field. His drive toward real-world applications pushed me to orient my research toward practical implementations and to think of high-level motivations. There is not a single experiment in this thesis that has not benefited from his experience, nor a single contribution that does not reflect his intent to be useful not only to the field of computer science and but also to the communities in which we live. I am also grateful for his mentorship, availability, kindness and positive energy that carried me through my graduate studies.

I would be remiss not to mention each of my dissertation committee members for their feedback. I would like to thank Professor Sanmay Das for trusting me to collaborate on many research projects. I would like to recognize that both Professors Dov Gordon and Martin Slawski encouraged rigor in my work: many proofs in this thesis are inspired by their analytical mindset and teaching. I would also like to highlight that Professor Brittany Johnson-Matthews is my role model when it comes to turning complex ideas into powerful, but simple, messages accessible to a general audience. I would also like to thank all the staff of the computer science department, with a particular mention to Michele Pieper and Cecelia Kimes for how responsive, cheerful and dedicated they have been whenever I needed their help.

I would also like to recognize the multiple collaborations that have honed my research skills. I believe that conversations with graduate students at the GMU Data Mining Lab, including my co-authors Jonathan Vasquez and Tasfia Mashiat, have been instrumental to my progress. I was lucky to participate in interdisciplinary research at the GMU Quantum Science and Engineering Center. I would particularly like to thank Professors Mingzhen Tian from the GMU Department of Physics and Maria Emelianenko from the GMU Department of Mathematics for inviting me to engage with their research efforts in quantum machine learning. I have also learned invaluable skills from my research fellows at the Frontier Development Lab. Andrés Muñoz-Jaramillo from Southwest Research Institute, Anna Jungbluth from the University of Oxford and Paul J. Wright from Stanford University deserve a special mention not only for our scientific collaboration, but also for their kindness and their unwavering trust in me that, I believe, makes me a better person.

Most importantly, I am so grateful to my wife, Mariesa; my son, Liam; and my daughter, Lena. Their support, ideas, and emotions are the cornernstone of this work. They nudged me to apply to the graduate program and they never stopped believing in me. There are no words to describe how proud of them I am. I could not conclude without mentioning my bunnies, Obie who monitored my daily progress from under my desk, and Luciolle, who, as a firefly, lights even the darkest days.

Table of Contents

List of Tables

List of Figures

Abstract

PRE- AND POST- FAIRNESS PROCESSING FOR BLACK-BOX CLASSIFIERS

Xavier Gitiaux, PhD

George Mason University, 2022

Dissertation Director: Dr. Huzefa Rangwala

Machine learning algorithms increasingly support decision-making systems in contexts where outcomes have long-term implications on the subject's well-being. At issue is whether an algorithm's outcomes are unfair and depend on demographic characteristics – race, age, gender, religious or political beliefs – that are irrelevant to the task. Empirical evidence indicate that a wide range of applications do not deliver the same experience depending on demographic characteristics of the client.

This study focuses on two types of approaches to mitigate potentially unfair outcomes of a classifier while making no assumption on the classifier itself: (i) pre-processing to remove encoded biases in the data; and, (ii) post-processing to audit whether a classifier's outcomes meet a given fairness criteria.

In fair pre-processing, we focus on methods in unsupervised fair representation learning that extract from a data its underlying latent factors, while removing dependencies between latent variables and sensitive attributes. We make four contributions to the fair representation research. First, we recast fair representation learning as a rate-distortion problem and show that an encoder that filters out information redundancies would also remove dependencies between sensitive attributes and representations.

This insight motivates **FBC**, **F**airness **B**y **C**ompression, a compression-based approach to unsupervised fair representation learning that achieves state-of-the-art performance in terms of fairness-information trade-off.

Second, we implement a single shot fair representation learning method, **SoFaiR**, that allows the user to explore the entire unfairness-distortion curve at test time with one single trained model. SoFaiR adjusts the fairness/information properties of a representation at test time by masking bits in the tail of the bitstream. This reduces computational costs compared to existing methods in fair representation learning that require the user to re-train a model to explore different points on the fairness-information plane.

Third, we posit that for image data, sensitive attributes like gender or race are likely to be abstract concepts. At the same time, a high quality reconstruction of images requires to encode high resolution details. Therefore, a rate-distortion approach to fair representation learning needs to model low and high resolution latent variables. To test this hypothesis, we encode images into a hierarchy of quantized latent variables. Empirically, we find that only deep hierarchies, independently of model capacity, can generate representations orthogonal to the sensitive attributes, while maintaining low and high resolution information about the images.

Fourth, we derive necessary and sufficient conditions for a representation learned from a finite sample to offer fairness guarantees that generalize to any downstream user and to the infinite sample regime. The condition requires that for any distribution over the feature space, the encoder induces a distribution over the representation space such that the χ^2 mutual information between features and representation is finite.

Lastly, for both fairness pre-processing and auditing, it is reasonable to assume that classifiers that use the data are black-boxes that neither auditors nor data controllers can access to. In this context, we develop an auditing approach, **mdfa** (**M**ulti-**D**ifferential **F**airness **a**uditor), that verifies whether a classifier is nearly mean-independent of sensitive attributes within any subset of the feature space that can be computationally identifiable from a finite sample.

Chapter 1: Introduction

1.1 Rationale for Thesis

Machine learning algorithms are increasingly used to support decisions that could impose adverse consequences on an individual's life: for example, the criminal justice system uses machine learning algorithms to assess whether a criminal offender is likely to recommit crimes; or banks to determine whether a potential borrower is at risk of defaulting. At issue is whether these algorithms are fair. Although fairness can mean different and sometimes contradictory things, there is a growing social consensus that demographic characteristics like race or gender are *exogenous irrelevant characteristics* [4] to most machine learning tasks and should not affect outcomes.

Unfortunately, abundant examples show that for a wide range of applications, clients' experience varies with their demographic characteristics. A growing body of evidence has raised fairness concerns across a wide range of applications, including judicial decisions [5], face recognition [6], degree completion [7], medical treatment [8] or crime predictions [9].

Root causes of unfairness in machine learning include biases encoded in the data generating/collection process and direct or indirect use of sensitive attributes within the machine learning pipeline. This thesis addresses both causes, while making no assumption on the classifier itself. In the data science pipeline, we focus on two types of fairness interventions that are tailored toward classification tasks: (i) pre-processing to remove encoded biases in the data before it is used by the classifier; and, (ii) post-processing to audit whether the classifier's outcomes meet a given fairness criteria.

On one hand, fair pre-processing is an attractive solution to organizations dealing with data, since they are increasingly held accountable for the collection, use and disposal of the data. The European Union General Data Protection Regulation designates organizations

that collect data as data controllers [1] and makes explicit their responsibility to mitigate the discriminatory use of the data on the basis of sensitive attributes, including racial or ethnic origin, sexual orientation or political beliefs[2].

On the other hand, auditing aims at establishing contestability if a downstream application or data processor[3] generates outcomes that depend on sensitive attributes. Tools that provide evidence of disparate treatment are all the more valuable as a precedent in United States case law places the burden on the plaintiff to demonstrate disparate treatment – to establish that characteristics irrelevant to the task affect the algorithm's outcomes (*Loomis vs. the State of Wisconsin* [10]). Identifying the definitive characteristics of a classifier's discrimination empowers the victims of such discrimination. Moreover, a classifier's user needs warnings for individual instances in which severe profiling/discrimination has been detected.

For both fairness pre-processing and auditing, it is reasonable to assume that classifiers that use the data are black-boxes that neither auditors nor data controllers can access to. First, many assessment tools are proprietary and usually not transparent. Second, data controllers cannot always anticipate and control how downstream applications will process the data. Arms-length contracts between data controllers and third parties accessing the data are likely to be incomplete and leave out details related to the structure of the machine algorithm used by the data processor.

1.2 Research Questions

This thesis answers two research questions related to the development of methods to mitigate potential unfairness in machine learning algorithms without access to the algorithms themselves.

1. *Pre-processing:* How to pre-process data so that any future use of the data would not

[1] GDPR, Article 4
[2] GDPR, Recital 71
[3] GDPR, Article 4

discriminate against some demographic groups?

2. *Auditing:* How to audit black box classifiers to verify whether their outcomes meet a pre-defined fairness criteria?

1.3 Challenges and Solutions

1.3.1 Pre-processing: Fair Representation Learning

Organizations that collect and sell data, henceafter data controllers, are increasingly liable if future downstream uses of the data are biased against protected demographic groups. One of their challenges is to anticipate and control how the data will be processed by downstream users.

Unsupervised fair representation learning approaches ([11–13]) offers a flexible fairness solution to this challenge. A typical architecture (see Figure 1.1) in fair representation learning includes an encoder that maps the data into a representation and a decoder that reconstructs the data from its representation. The objective of the architecture is to extract from a data X the underlying latent factors Z that correlate with unobserved and potentially diverse task labels, while remaining independent of sensitive factors S. The idea has gained traction in the machine learning community since it is flexible in terms of the data science pipeline: it is independent of the modeling algorithm and can be integrated with data releases and publishing mechanisms.

Fair Information Bottleneck

Chapter 3 asks whether an encoder that filters out information redundancies would also remove dependencies between sensitive attributes and representations. Intuitively, if sensitive attributes S are direct inputs to the decoder, an encoder that aims for conciseness would not waste code length to encode information related to S in the latent factors Z. We show that in an information bottleneck framework [14], this intuition is theoretically founded:

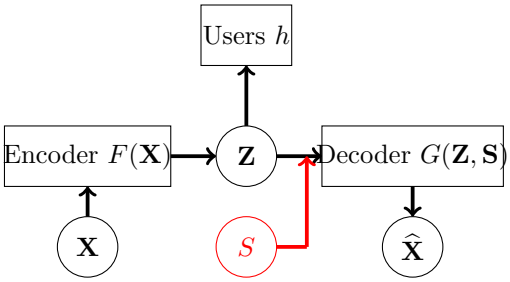


Figure 1.1: Unsupervised fair representation learning. Variables are: data $\mathbf{X}$; sensitive attribute $\mathbf{S}$; representation $\mathbf{Z}$; reconstructed data $\widehat{\mathbf{X}}$. The standard fair representation protocol includes an encoder F that maps X to its representation $\mathbf{Z}$; a decoder G that reconstructs $\mathbf{X}$ from $\mathbf{Z}$ and $\mathbf{S}$.

constraining the information flowing from the data X to the representation Z forces the encoder to control the dependencies between sensitive attributes S and representations Z. *It is sufficient to constraint the mutual information $I(Z, X)$ between Z and X in order to minimize the mutual information $I(Z, S)$ between Z and S.* This result contrasts with existing methods in fair representation learning that devote most of their effort to constraining the mutual information $I(Z, S)$ between representations Z and sensitive attributes S either via penalties measuring the statistical distance between the distributions of Z across sensitive attributes(e.g. [15]); or via an adversarial auditor that predicts sensitive attributes from Z (e.g [16, 17]).

Therefore, instead of directly penalizing $I(Z, S)$, we recast fair representation learning as a rate distortion problem that controls explicitly the bit rate $I(Z, X)$ encoded in the latent factors Z. We model the representation Z as a binary bit stream, which allows us to monitor the bit rate more effectively than floating point representations that may maintain redundant bit patterns. We estimate the entropy of the code Z with an auxiliary autoregressive network that predicts each bit in the latent code Z conditional on previous bits in the code. One advantage of the method is that the auxiliary network collaborates with the encoder to minimize the cross-entropy of the code.

Empirically, we demonstrate that the resulting method, Fairness by Binary Compression (henceforth, **FBC**) is competitive with state-of-the art methods in fair representation learning. Our contributions are as follows:

1. We show that controlling for the mutual information $I(Z, X)$ is an effective way to remove dependencies between sensitive attributes and latent factors Z, while preserving in Z, the information useful for downstream tasks.

2. We find that compressing the data into a binary code as in **FBC** generates a better accuracy-fairness trade-off than limiting the information channel capacity by adding noise (as in variants of β-**VAE**, [18]).

3. We show that increasing the value of the coefficient on the bit rate constraint $I(Z, X)$ in our information bottleneck framework allows to move smoothly along both rate-distortion and rate-fairness curves.

Single Shot Fair Representation Learning

All methods in fair representation learning generate a fairness-information trade-off (e.g [16]). A likely reason for this trade-off is that the unobserved mixing mechanism between Z and S may hide some confounding variable such that Z is not independent of S. In many social contexts, it is reasonable to assume complex mixing mechanisms where sensitive attributes cannot be factored out in the representation space. Moreover, even if the true generating process factorizes sensitive attributes out, such factorization may be difficult to obtain in a finite sample regime and the dimension of the representations space may be constrained to be lower than the dimension of the data manifold [19].

Current approaches in fair representation learning are flexible with respect to downstream tasks [11, 12, 16, 20]; partially to sensitive attributes [21]; but are inflexible with respect to their fairness-accuracy trade-off. A single learned representation can provide fairness for many downstream tasks, but the fairness-information trade-off is set at training time. This lack of flexibility with respect to the fairness-information trade-off is a limitation

to deployment of fair representation learning approaches. A data controller would like to adjust how much information about sensitive attributes it leaks depending on its client. For example, in some medical applications, information related to age, race or gender is a necessary diagnosis feature; in some applications, a partial release of information related to age, race or gender is sufficient; in others, it is completely inappropriate. With existing methods in fair representation learning, a data owner would have to re-train a fair encoder-decoder to meet each request. At issue are the computational cost and the lack of consistency between released representations. A data controller would not be able to explain easily the relation between each data product it releases, since they are generated by different models and the mapping between them is not easily explainable.

Chapter 4 introduces SoFaiR, **S**ingle Sh**o**t **Fai**r **R**epresentation, a method to generate a unfairness-distortion curve with *one single trained model*. We first expand our results from Chapter 3 and show that we can derive unfairness-distortion curves from rate-distortion curves. We can control for the mutual information $I(Z, S)$ between representation and sensitive attribute by encoding X into a bitstream and by controlling for its entropy. We then construct a gated architecture that masks partially the bitstream conditional on the value of the Lagrangian multiplier in the rate-distortion optimization problem. The mask adapts to the fairness-information trade-off targeted by the user who can explore at test time the entire unfairness-distortion curve by increasingly umasking bits.

Besides saving on computational costs, SoFaiR allows users to interpret what type of information is affected by movement along unfairness-distortion curves. Moving upward along unfairness-distortion curves unmasks bits in the tail of the bitstream and thus, increases the resolution of the representation encoded in a binary basis. By correlating these unmasked bits with data features, the practitioner has at hand a simple method to explore what information related to the features is added to the representation as its fairness properties degrade.

Empirically, we demonstrate on three datasets that at a cost constant with the number of points on the curve, SoFaiR constructs unfairness-distortion curves that are comparable

to the ones produced by existing multi-shot approaches whose cost increases linearly with the number of points. On the benchmark Adults dataset, we find that increasingly removing information related to gender degrades first how the representation encodes working hours; then, relationship status and type of professional occupations; finally, marital status.

The contributions of Chapter 4 are as follows:

1. We formalize fairness-information trade-offs in unsupervised fair representation learning with unfairness-distortion curves and show a tractable connection with rate-distortion curves.

2. We propose a single shot fair representation learning method to control fairness-information trade-off at test time, while training a single model.

3. Moreover, we offer a method to interpret how improving or degrading the fairness properties of the resulting representation affects the type of information it encodes.

Hierarchical Fair Representation Learning

In Chapter 3 and 4, our framework is unsupervised since the data controller does not access any task label that future downstream data processors will predict. Therefore, data controllers can only monitor the information content of a representation by measuring the distortion incurred when reconstructing a data point from its representation. In Chapter 5, we explore how to generate fair representation of images while producing high quality reconstruction of the original data.

Images are challenging to fair representation learning because sensitive attributes (e.g. race, gender) are likely to be abstract and global concepts while high quality reconstruction requires capturing localized details. In Chapter 5, we propose to encode the data into a hierarchy of features, where low-resolution latent variables capture global features and higher resolution latent variables are conditionally dependent on lower resolution ones. In this hierarchy, global features correlated with the sensitive attribute are redundant information if sensitive attributes are provided directly to the decoder. This intuition, consistent with

our theoretical and empirical results in chapters 3 and 4, allows us to solve unsupervised fair representation learning problem via hierarchical quantization.

We leverage recent contributions in deep variational auto-encoders (e.g. [22, 23]) and multi-resolution image compression [24,25] to reconstruct high quality images while filtering out sensitive attributes from a hierarchical representation of the data. The rationale is that if we were compressing images of faces to few bits, we would encode information related to abstract concepts like identity, gender, race. At that level of compression, the encoder would learn that some of these abstracts concepts related to sensitive attributes are redundant since sensitive attributes are directly provided to the decoder. However, the reconstruction of the images from these few bits will miss important properties of the image like hair color, pose, etc. Additional bits to encode the image would capture higher resolution details, but these details are likely to be entangled with sensitive attributes, which makes it challenging for the encoder to filter out sensitive attributes themselves. We propose to solve this paradox by learning a hierarchy of latent variables.

The contributions of chapter 5 are as follows:

- We verify empirically that depth independent of model capacity is critical to solve fair representation learning problems for images data.
- We find that depth is only beneficial to compression-based methods in fair representation learning and does not improve performances for adversary-based techniques.

Statistical Robustness of Fair Representations

Unsupervised fair representation learning approaches like **FBC** or **SoFaiR** use a finite sample to obtain representations whose fairness properties need to generalize to unseen data and unknown downstream data processors. However, to date, there is no study on what characteristics the representation must have to statistically guarantee this generalization. Chapter 6 explores conditions on the encoder to generate representation distributions with fairness guarantees that hold *for any data processor*.

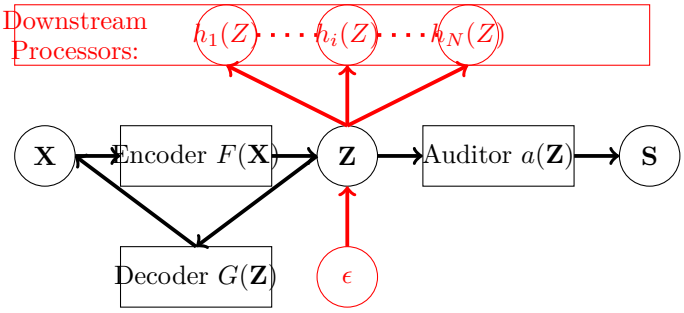


Figure 1.2: Robust fair representation learning. In addition to the encoder-decoder structure of Figure 1.1, an auditor a evaluates the statistical dependence between Z and S. An additive Gaussian white noise (AGWN) channel ϵ ensures that finite sample fairness guarantees can be established for all downstream data processors h_1,...., h_N using Z.

We show that for fairness guarantees derived from finite samples to generalize to all downstream data processors, it is necessary that a measure of information – the χ^2 mutual information – between feature and representation is finite. Moreover, we prove that a finite χ^2 mutual information between feature and representation is a sufficient condition on representation mappings to guarantee a $O(n^{-1/2})$ approximate rate of empirical certificates.

In practice, it is challenging to control whether the χ^2 mutual information is finite for unknown distributions over $\mathcal{X}$. However, we show that an additive Gaussian white noise (AGWN) channel placed after any representation mapping will bound the χ^2 mutual information once the representations have passed through the channel. The channel smoothes the representation distribution by transforming it into a mixture of Gaussian distributions that can be estimated by Monte Carlo integration ([26]). Therefore, a plug-in fairness auditor that relies on estimating the class conditional density functions over the representation space achieves a convergence rate of $O(n^{-1/2})$ and thus delivers meaningful empirical certificates of fairness.

We empirically find on various synthetic and fair learning benchmark datasets that an AGWN channel in fair representation learning is sufficient for empirical certificates to upper bound the demographic parity of multiple downstream users that attempts to predict sensitive attributes from samples of the representation distribution. An AGWN channel

improves upon existing approaches in adversarial fair representation learning whose fairness guarantees do not extend beyond a set of specific downstream users. Moreover, we did not find strong evidence that obtaining good approximation rates for empirical certificates comes at the cost of significantly degrading the accuracy-fairness trade-off of downstream predictive tasks.

The contribution of Chapter 6 are as follows:

1. We prove a necessary and sufficient condition for fair representations to be stamped with finite sample demographic parity certificates that generalize to the infinite sample regime.

2. We prove that adding a noisy channel allows to obtain an estimator of demographic parity certificates that converges at a $O(n^{-1/2})$ rate to the true demographic parity certificate of the data representation.

1.3.2 Auditing

While fair pre-processing is concerned with controlling the input to black box classifiers, auditing for fairness is concerned with investigating their output. Of particular interest is to establish whether a classifier's outcomes would change depending on demographic characteristics of the individual. This auditing task is a first and necessary task to identify disparate treatment, which characterizes classification for which sensitive attributes affect the algorithm's outcomes. Auditing for disparate treatment of a black box classifier faces at least two challenges:

(i) Average lack of disparate treatment does not necessarily imply that some individuals would not experience different outcomes, had their demographics characteristics been different;

(ii) The auditor does not observe the counterfactual outcome if their demographics characteristics of individual had been different.

Challenge (i) relates to limitations of aggregate definition of fairness, since it can only offers guarantees for an average representative of each demographic group, but not for a specific individual nor for a structured sub-group. [27] provide anecdotal evidence for what they coin subset targeting. [28] and [29] provide further empirical evidence that aggregate notion of disparate treatment do not protect in practice subgroups defined by complex intersections of many sensitive attributes. The influence of sensitive attributes on a classifier's outcomes could be complex, non-linear and could affect only a subset of individuals.

Challenge (ii) is akin to the classical problem of potential outcomes [30], [31], where each individual in the dataset possesses two outcomes: an observed outcome that corresponds to the classifier's prediction; a counterfactual outcome that would be the classifier's prediction, had the individual's sensitive attributes been different. At issue is that the auditor does not observe the counterfactual outcome. Moreover, in many real-world situations, the data is unbalanced across demographic groups: the distributions conditioned on sensitive attributes differ, which makes the auditing problem even harder.

Computationally Identifiable Disparate Treatment

In Chapter 6, we address challenge (i) by introducing the notion of multi-differential fairness to guarantee that the classifier is nearly mean-independent of sensitive attributes within any subset of the feature space.

We represent subset or sub-population as collection of membership indicators. The richer the collection of indicators, the stronger the fairness guarantees. However, the granularity of our fairness guarantees is bounded below by the fact any auditing tool can only identify a sub-population via a finite sample drawn from the features distribution. We reduce searching for violations of multi-differential fairness to agnostic learning of the collection of membership indicators. We show that violations of differential fairness is a problem of finding correlations between sensitive attributes and classifier's outcomes. Searching for instances of violation of multi-differential fairness is akin to predicting where in the features

space the binary values of sensitive attributes and classifier's outcomes coincide. Therefore, multi-differential fairness controls for disparate treatment among all sub-population that can be computationally identifiable.

We propose an auditing tool, **mdfa**, to search for the worst-case violations of multi-differential fairness on re-balanced data. Applied to a case study of recidivism risk assessment in Broward County, Florida, **mdfa** identifies a sub-population of African-American defendants who are three times more likely to be considered at high risk of violent recidivism than similar individuals of other races. Moreover, when applied to three additional datasets related to crime, income and credit predictions, **mdfa** finds heterogeneous treatment among sub-group of individuals, even after adjusting the outcome of a black box classifier to meet group-level fairness criteria.

1.4 Outline

This proposal is organized in eight chapters.

- Chapter 1 provides an introduction to fairness in machine learning and presents a rationale for studying pre-processing and post-auditing methods.
- Chapter 2 introduces notations, different notions of fairness in machine learning and reviews the literature on fair representation learning and post-auditing.
- Chapter 3 demonstrates that data quantization is an effective method to filter out sensitive attributes. The method and results are published in [20].
- Chapter 4 demonstrates that we can compute unfairness-distortion functions from rate-distortion functions and uses this insight to propose a single-shot fair representation learning method. This work leads to a submission at IJCAI 2022.
- Chapter 5 proposes hierarchical quantization as a method to learn fair representations for images dataset.

- Chapter 6 proves necessary and sufficient conditions for the fairness guarantees of a representation obtained from a finite sample to generalize to the infinite sample regime. Theoretical and empirical results from this chapter are published in [13].

- Chapter 7 introduces mdfa, an auditing tool that searches for subsets of the features space where the classifier's outcomes differ significantly across demographic groups. This work is published at [32].

- Chapter 8 discuss the ethical implications and limitations of this thesis and proposes avenues for future work. Moreover, it briefly introduces additional contributions that complement this thesis, including auditing tools for education data mining published at [33]; and, for allocation of scarce resources submitted at FAccT 2022 [34].

The code related to each chapter can be find here.

Chapter 2: Notations and Literature Review

2.1 Preliminaries and Notations

2.1.1 General Setting

Consider a population of individuals represented by features $X \in \mathcal{X} \subset [0,1]^{d_x}$, sensitive attributes in $S \in \mathcal{S} \subset \{0,1\}^{d_s}$ and outcomes $Y \in \mathcal{Y} \subset \{0,1\}^{d_y}$.

A representation Z of the features X is a code or latent factors that encodes the data in a lower dimension space $\mathcal{Z} \subset [0,1]^{d_z}$, with $d_z > 0$.

A classification task consists of a mapping f that predicts outcomes Y given features X and possibly sensitive attributes S. We will also encounter in this thesis classification tasks that predict Y from a representation Z of the data instead of the data itself.

2.1.2 Mutual Information and Entropy

To measure the dependencies between two variables X_1 and X_2, we will extensively use the notion of mutual information $I(X_1, X_2)$[1] defined as

$$I(X_1, X_2) = \sum_{x_1} \sum_{x_2} p(x_1, x_2) \log \left(\frac{p(x_1, p(x_2)}{p(x_1)p(x_2)} \right). \tag{2.1}$$

By definition, the mutual information is equal to the Kullblack Leibler divergence between the joint distribution $p(X_1, X_2)$ and the product of the marginal distributions $p(X_1)$ and $p(X_2)$. $I(X_1, X_2)$ is non-negative and measures the price of encoding X_1 and X_2 as independent variables if they are not. The smaller is $I(X_1, X_2)$ the more independent are X_1

[1] In Chapter 6, we will call it Shannot mutual information to distinguish it from the χ^2- mutual information.

and X_2. Of particular interest in this thesis is the mutual information $I(Z, S)$ between a representation Z of the data and the sensitive attribute S.

The entropy $H(X)$ of a random variable X measures the amount of uncertainty about its possible outcomes. Formally, it is defined as

$$H(X) = \sum_x -p(x) \log p(x). \tag{2.2}$$

The mutual information between X_1 and X_2 compares the entropy of X_1 before and after observing X_2 or vice-versa:

$$I(X_1, X_2) = H(X_1) - H(X_1|X_2) = H(X_2) - H(X_2|X_1). \tag{2.3}$$

Table 2.1: Notations common across the thesis

Symbol	Meaning
X	Features
S	Sensitive Attribute
Y	Ground truth label
Z	Representation
$\mathcal{X}$	Feature space
$\mathcal{S}$	Sensitive attribute space
$\mathcal{Z}$	Representation space
d_x	Dimension of the feature space
d_s	Dimension of the sensitive attribute space
$f : \mathcal{X} \to \mathcal{Y}$	Classifier using data X
$F : \mathcal{X} \to \mathcal{Z}$	Encoder/Representation mapping
$G : \mathcal{X} \times \mathcal{S} \to \mathcal{X}$	Decoder
$I(X, Z)$	Mutual information between X and representation Z
$I(S, Z)$	Mutual information between sensitive attribute S and representation Z
$H(X\|Z, S)$	Entropy of X conditional on Z and S

2.2 What is Fairness in Machine Learning?

2.2.1 Unfairness Encoded in the Data

A classifier f can generate unfair outcomes for at least three reasons. First, f may use sensitive attributes directly in its predictions, i.e. f is a mapping from $\mathcal{X} \times \mathcal{S}$ to $\mathcal{Y}$. Therefore, the classifier f may predict different outcomes for two individuals with the same features X but different sensitive attributes S. Differential treatment, i.e. the direct use of sensitive attributes in a prediction task, is, at least in the United States, considered a source of illegal discriminatory behavior. However, protection against machine learning differential treatment may be difficult to enforce because (i) there are legal precedents that place the burden on the plaintiff to establish that sensitive attributes affect a classifier's predictions (*Loomis vs. State of Wisconsin*); and, (ii) many assessment tools are proprietary and challenging to audit. However, preventing the collection of sensitive attributes could alleviate the issue.

A more subtle cause of unfair classification is that the classifier f may not remove existing social biases that are encoded in the data generating process. For example, data features X might result from subtle and intricate channels through which past racial segregation and gender discrimination affect current socio-economic status. The distribution over outcomes $p(Y|X)$ might then depend on sensitive attributes. Therefore, a classifier trained for accuracy tries to learn the distribution $p(Y|X)$ and thus, generates outcomes dependent on sensitive attributes. For example, if f predicts the recidivism risk of an inmate, it will skew higher risks toward individuals self-identified as African American, since the prediction relies on who is arrested for a crime, not on who committed a crime and historically, arrest rates have been higher for African American [35]. Encoded biases exist also in the distribution $p(X)$ of features. Unobserved characteristics that affect both sensitive attributes S and features X generate statistical dependence between X and S: encoding are redundant, i.e. $p(X|S) \neq p(X)$ [27,36]. Therefore, obfuscating S would not remove its effect on f. For example, the absence of variables related to race in a loan application would not preclude its influence on mortgage costs if the loan application includes zipcode – a strong proxy for

race in segregated societies and a necessary input in mortgage application.

Encoded biases can be exacerbated by a data collection process that dynamically relies on past predictions to collect additional samples. Feedback loops can exacerbate existing biases. If a model f is used to predict crime rates and to decide which neighborhood police should patrol, those neighborhoods will be over-represented in future samples collected to train the model [37].

2.2.2 Definition of Fairness

Even if the root causes of fairness in machine learning were well understood, it would not be clear what fairness for a classifier f means. The literature has focused on two types of notions, *statistical* or *individual*, both of them with advantages and limitations.

Statistical Fairness.

Statistical definitions of fairness guarantee that a pre-specified statistics does not vary across groups that differ by a small set of sensitive attributes $\mathcal{S} = \{s_1, ..., s_K\}$. Popular statistics are *demographic parity*, *equality of odds* and *equality of opportunities*. Demographic parity [27] defines fairness as parity of positive classification rates across groups:

Definition 2.2.1. Demographic parity *Consider a distribution μ over $\mathcal{X} \times \{s_1, ..., s_K\}$. A classifier $f : \mathcal{X} \to \{0, 1\}$ satisfies $\delta-$ Demographic Parity on μ if and only if for $k, k^{'} = 1, .., K$,*

$$\Delta^{DP}(f, k, k^{'}) \triangleq |E_{x\sim\mu}[f(x)|S = s_k] - E_{x\sim\mu}[f(x)|S = s_{k'}]| \leq \delta \tag{2.4}$$

Demographic parity is a strong definition of fairness since it requires statistical independence of $f(X)$ and S, unconditional on the outcomes Y. Equality of odds and opportunities relax demographic parity by only requiring statistical independence conditional on outcomes Y [38].

Definition 2.2.2. Equality of Opportunities *Consider a distribution μ over $\mathcal{X} \times \{s_1, ..., s_K\}$. A classifier $f : \mathcal{X} \to \{0, 1\}$ satisfies $\delta-$ equality of opportunity on μ if and only if for*

$k, k^{'} = 1, .., K,$

$$\Delta^{EO}(f, k, k^{'}) \triangleq |E_{x\sim\mu}[f(x)|S = s_k, Y = 1] - E_{x\sim\mu}[f(x)|S = s_{k'}|Y = 1]| \leq \delta \tag{2.5}$$

Definition 2.2.3. Equality of Odds *Consider a distribution μ over $\mathcal{X} \times \{s_1, ..., s_K\}$. A classifier $f : \mathcal{X} \to \{0, 1\}$ satisfies $\delta-$ equality of odds on μ if and only if for $k, k^{'} = 1, .., K$ and $y \in \{0, 1\}$*

$$\Delta^{EOD}(f, k, k^{'}) \triangleq |E_{x\sim\mu}[f(x)|S = s_k, Y = y] - E_{x\sim}[f(x)|S = s_{k'}|Y = y]| \leq \delta \tag{2.6}$$

The appeal of statistical notions of fairness is that for a fixed classifier f, these criteria can be estimated from a sample $\mathcal{D}_n$ at a convergence rate of $O(1/n)$ and that Δ_n converges in distribution toward a normal distribution. For a classifier f and a sample $\mathcal{D}_n$, we can obtain confidence intervals of the population $\Delta^{DP}, \Delta^{EO}, \Delta^{EOD}$ [39].

The main limitation of statistical notions of fairness is that it only provides guarantees for an average representative of each group but not for a specific individual nor for a structured sub-group. Suppose for example, that S is a binary variable and that both demographic groups have the same size: $p(S = 1) = p(S = 0)$. Suppose further that X is also a binary variable $X \in \{0, 1\}$ and that $p(X = x|S = 0) = p(X = x|S = 1) = 1/2$ for $x \in \{0, 1\}$. A classifier f that predicts Y according to X in the first group ($f(X) = X$ if $S = 0$) and Y according $1 - X$ in the second group ($f(X) = 1 - X$ if $S = 1$) will satisfy demographic parity, but is clearly problematic from a fairness point of view. [27] provide further anecdotal evidence of what they call subset targeting. Moreover, [29], [28] and [32] show further empirical evidence that statistical notions of fairness do not protect subgroups defined by a complex intersection of many sensitive attributes [28,29] or a structured slicing of the features space [32].

Individual Fairness.

Individual fairness addresses the limitations of statistical notions of fairness and constrains that for each pair of individuals, the classifier treats them similarly if they have similar features [27]. Features similarity is measured by a metric distance $d : \mathcal{X} \times \mathcal{X} \to \mathrm{R}$:

Definition 2.2.4. Individual Fairness *A classifier* $f : \mathcal{X} \to \{0, 1\}$ *satisfies* (d, δ) *individual fairness if and only if for all* $x, x^{'} \in \mathcal{X}$,

$$|f(x) - f(x^{'})| \leq \delta d(x, x^{'}). \tag{2.7}$$

The definition of individual fairness in (2.7) is intuitive and protects individuals against many unfairness evils, including redlining and subset targeting (see [27]). However, unlike with statistical definitions of fairness, there is no guarantee that the individual fairness constraint generalizes to out-of-sample individual $x \notin \mathcal{D}_n$. Moreover, a definition of individual fairness as in (2.7) requires to agree upon a similarity metric d, which raises non-trivial fairness issues. In fact, advocates of affirmative actions would argue that fairness can be achieved only by differentiating the meaning of features between demographic groups. For example, it is debatable whether SAT scores should be interpreted similarly across demographic groups, given that historically, some minority populations have had systematically lower access to educational opportunities.

Not all at Once.

A natural question is how statistical and individual notions of fairness compose with each other. [40] show an impossibility result: unless the classification task is trivial, false negative rate $P(f(X) = 0|Y = 1)$, false positive rate $P(f(X) = 1|Y = 0)$ and positive classification rate $P(f(X) = 1)$ cannot be simultaneously equalized across demographic groups.

On the relation between demographic parity and individual fairness, [27] show that individual fairness implies that if $\mathcal{S} = \{0, 1\}$, Δ^{DP} is bounded above by the Wasserstein

distance between the distributions $p(X|S=0)$ and $p(X|S=1)$. However, as shown earlier in this chapter, there is no converse result: demographic parity has no implication on individual fairness as defined in (2.7).

Recent contributions explore the possibility to give a statistical meaning to definitions of individual fairness by comparing a classifier's outcomes between structured subgroups of individuals instead of pairs of individuals. [41], [29], [42] or [32] define notions of statistical fairness across an infinite number of subgroups, whose membership is controlled by a class of functions with bounded complexity. It generates a more granular notion of statistical fairness without the necessity to define a similarity metric. On one hand, it avoids some of the statistical pitfalls of individual-based definitions of fairness by defining fairness on average. On the other hand, it is a stronger notion of fairness than aggregate definition, since it imposes average fairness constraints to hold not just over coarsely defined demographic groups, but also over very finely defined sub-populations.

However, sub-group definitions of fairness raise issues of their own. It is not clear how to choose the class of functions that define the sub-population membership and how to choose which features to input in these functions [43]. For example, shall we include SAT scores or high school GPA as inputs of sub-population membership since they correlate with race and gender?

On the Choice of a 'Right' Definition of Fairness?

Our review of definitions of fairness in machine learning shows that (i) the literature proposes many, at time incompatible, fairness criteria; (ii) there is little guidance of which notions should be preferred. The latter conclusion is not necessarily a limitation of fairness in machine learning, but is more a reflection of social, ethical and philosophical debates that run throughout our societies. For example, questions related to the choice of a similarity metric [27] or features used to condition parity [44] are reminiscent of discussions related to affirmative action: *how do we choose whether an individual is 'qualified' for a given position; what are irrelevant characteristics for a given task; how do we compensate for complex social*

and historical mechanisms that systematically disfavor some demographic groups and make them look 'less qualified'?

One of the contributions of fairness in machine learning is (i) to bring questions at the heart of theories of distributive justice [45] to algorithmic design; (ii) to offer tools for outcomes to meet a pre-specified fairness criteria [4]; and, (iii) to verify ex-post whether this criteria has been met.

2.3 Fair Representation Learning

Existing pre-processing methods to remove biases encoded in the data include sampling and reweighting (e.g. [46, 47]), optimization procedures to learn a data transformation that both preserve utility and limit discrimination (e.g. [48]), and representation learning (e.g. [11]). Representation learning seeks to encode the data while removing correlations between features and sensitive attributes. The idea has gained traction in the machine learning community since it is flexible in terms of the data science pipeline: it is independent of the modeling algorithms and can be integrated in the knowledge discovery and data mining framework. Moreover, fair representation learning benefits from recent advances in representation learning [49] that map with little or no supervision high-dimensional observations to low dimension latent space such that the original observations can be approximately decoded from their lower-dimensional representations.

Many contributions use a supervised setting where the downstream task label is known while training the encoder-decoder architecture (e.g [16, 17, 50–52]). However, [11], [12] and [53] argue that in practice, an organization that collects data cannot anticipate what the downstream use of the data will be. Therefore, in this thesis, we explore theoretical properties and practical implementations of unsupervised fair representation learning: data controllers do not access downstream task labels when encoding the data into a fair representation.

Table 2.2: Methods in unsupervised fair representation learning organized by whether the fairness properties of the learned representations is obtained by minimizing the mutual information between sensitive attributes S and representations Z; or by minimizing the mutual information between data X and representations Z; and whether Z is modelled as a binary bit stream or is convolved with Gaussian noise.

Methods	Fairness by controlling:		Examples	Unsupervised
	$I(Z,S)$	$I(Z,X)$	✗	
Adversarial	Minimizing auditor's cross-entropy		[16], [17], [21]	✗
MMD	Mimizing maximum mean discrepancy	✗	[55], [15]	✗
$\beta-$ **VAE**	✗	Noisy Z	[18], Chapter 3, 5	✓
FBC, **SoFair**	✗	Binary Z	Chapter 3,4	✓

2.3.1 Removing Dependencies between Sensitive Attributes and Representations

In this unsupervised setting, data controllers can control for the mutual information between the representation and the sensitive attribute. Since in most setting the mutual information is intractable, we resort to proxies for the mutual information between representations and sensitive attributes to control for the fairness of the learned representations. Proxies can be either (i) a maximum mean discrepancy penalty [54] that extends a deterministic [55] or variational [15] auto-encoder; or, (ii) the cross-entropy of an adversarial auditor that predicts sensitive attributes from the representations [16,17,56,57] (see Table 2.2). Note that in the current literature, all these proxies for $I(Z,S)$ are used in the context of supervised setting (see Table 2.2, but can be readily adapted to the unsupervised setting.

Adversarial fair representation learning has benefited from recent developments in adversarial learning for generative modeling (see [58] for a survey) or domain adaptation (e.g. [59]). A data encoder generates a representation of the data and fools a neural network that attempts to predict sensitive attributes from samples of the representation distribution.

Our approach in chapter 3 to 5 – **FBC**, **SoFaiR** – contrasts with existing work since it does not control directly for the leakage between sensitive attributes and representations.

FBC obtains fair representations only by controlling its bit rate. In a supervised setting, [52] show that nuisance factors can be removed from a representation by over-compressing it. We extend their insights to unsupervised settings and show the superiority of bit stream representations over noisy ones to remove nuisance factors. Our insights could offer an effective alternative to methods that learn representations invariant to nuisance factors (e.g. [60–62]).

In chapter 4, we demonstrate theoretically how to derive unfairness-distortion curves from rate-distortion curves. On one hand, rate distortion functions [14, 63] characterize the minimum average number of bits $R(D)$ used to represent X by a code Z while the expected distortion incurred to reconstruct X from the code is less than D. On the other hand, unfairness-distortion functions measure the minimum mutual information between code Z and sensitive attribute S that a data controller has to tolerate for the distortion to be less than D. We show that both functions differ only by the entropy $H(X|S)$ of the data conditional on the sensitive attribute, provided that distortion is measured by $H(X|Z,S)$, i.e. the entropy of the reconstructed data conditional on the code Z and the sensitive attribute S.

By re-casting unsupervised fair representation learning as a rate-distortion problem, we can borrow techniques from end-to-end learning of compressible representation, including quantization and variable rate compression. We use soft-quantization techniques when backpropagating through the model [64] and hard quantization techniques during the forward pass [65]. In chapter 3, we estimate the entropy of the code as in [65] by computing the distribution $P(Z)$ of Z as an auto-regressive product of conditional distributions, and by modeling the auto-regressive structure with a PixelCNN architecture [66, 67].

In chapter 4, we apply approaches in rate-distortion that learn adaptive encoder and vary the compression rate at test time (e.g. [68,69]). Our adaptive mask relates to the gain function in [70] that selects channels depending on the targeted bit rate. We rely on successive refinement methods from information theory (e.g [71]) that use a common encoder for all points on the unfairness-distortion curve and add new information by appending bits

to a initially coarse representation. To our knowledge, chapter 4 is the first contribution to implement a deep learning multi-resolution quantization and apply it to the problem of fair representation learning.

2.3.2 Robustness to Unknown Downstream Users

State-of-the-art techniques in fair representation learning offer only fairness guarantees for some downstream users, but not for all of them [43]. Because of the Pinsker's inequality, methods minimizing a proxy for the mutual information between representations and sensitive attributes minimize an upper bound of the demographic disparity of the learned representation. However, the fairness guarantees apply only to classifiers that have a bounded norm in a reproducing kernel Hilbert space [15]; or that belong to the same class as the auditing adversary [16].

[72], [73] show empirically that those fairness guarantees do not generalize well to new downstream classifiers that use fresh samples from the representation distribution. It is a limitation since the main motivation for the fair representation learning agenda is to be an upstream step in the data mining pipeline that protects data controllers against all future discriminatory uses of the data.

In chapter 6, we explore conditions so that the learned representation offers fairness guarantees against adversaries that do not necessarily belong to the same class as the auditor used during the training of the encoder. [16] and [74] explore empirically whether representations that achieve demographic parity for a specific downstream task generalize to new tasks in terms of accuracy and fairness. We extend their work by showing theoretically and empirically that introducing an AGWN channel in fair representation learning offers generalization guarantees to all future tasks. Moreover, introducing an AGWN channel avoids the need for an adversarial auditor, since it allows approximating the empirical fairness certificate with a differentiable loss that can be computed by Monte Carlo sampling.

Similar to our approach, the differential privacy literature relies on noise injection to guarantee that two neighboring datasets are indistinguishable [75]. However, in the context

of differential privacy, indistinguishability is only obtained by adding Gaussian/Laplacian noise. In our fairness context, for a finite sample, statistical hiding comes from learning representations subject to a demographic parity constraint; the injection of Gaussian noise is only a means to generalize the statistical hiding property to the infinite sample regime.

2.3.3 Flexible Fair Representations

One advantage of fair representation learning as a pre-processing method is that it is flexible with respect to downstream tasks. A single learned representation tailored a specific task can transfer at test time to new downstream tasks [11, 16]. Moreover, recent contributions [21, 52] allow to adapt a single representation to multiple conjunctions of sensitive attributes, by randomly flipping the bits that correspond to the demographic groups that need protection for a given downstream task. Chapter 4 extends these efforts by allowing the data owner at test time to adapt not only which demographic groups need to be protected, but also how much information about these protected groups and about the data itself is released.

An application of flexible fair representation learning is its deployment to many sources of data, where downstream tasks occur in massively distributed networks. Most work related to fairness in a federated framework [55] focus on constraining directly downstream applications by adding local or global constraints [76, 77]. In our unsupervised setting where no downstream tasks is known at training time, [78] proposes a method to learn fair representations locally on edge devices. However, their approach does not allow a user to adjust the fairness-information trade-off at test time without re-training local copies of a central model. In Chapter 4, distributed data owners can vary at test time the length of the code that represents their data depending on the downstream application they share their data with.

2.4 Auditing Black Box Classifiers

Chapter 7, as in [41,42,79] provides a definition of fairness that protects group of individuals as small as computationally possible. This is a step in the right direction since empirical observations in [27, 79] show that aggregate level fairness cannot protect sub-populations against severe discrimination.

Prior contributions on algorithmic disparate treatment (e.g. [80]) have focused on whether sensitive attributes are used directly to train a classifier. This is a limitation when dealing with classifiers with unknown inputs. Multi-differential fairness addresses this limitation at the cost of a relaxation of disparate treatment compared to [81]. Our definition marginalizes the classifier's outcome over sub-population characteristics instead of marginalizing over individual characteristics. Although leading to a non causal notion of fairness, we show that our relaxation is necessary to efficiently find whether a black-box classifier violates differential fairness.

Multi-differential fairness borrows from the literature in differential privacy [75]. Reinterpretations of fairness as a privacy problem can be found in [82,83], but those contributions marginalize only over sensitive attributes.

The relaxation of differential fairness to sub-population requires rebalancing the distribution of features across sensitive attributes. Our kernel matching technique borrows from domain adaptation (see e.g. [84]) and counterfactual analysis (e.g. [85]). Because **mdfa** balances the features across sensitive classes, its outcome cannot be explained by disparate impact (as in [36]). In fact, empirical results in section 4 show that even after removing disparate impacts, **mdfa** still detects violations of multi-differential fairness.

Chapter 3: Fairness by Compression

In this chapter, we demonstrate that in an unsupervised framework, a simple encoder that filters out information redundancies removes also dependencies between representations Z and sensitive attributes S, provided that the sensitive attributes are direct inputs to the decoder. We show that in an information bottleneck, learning concise encoding of the data forces the encoder to control the information flow between Z and S. Therefore, for learning fair representation compression of the data is sufficient.

This simple result offers an alternative, fairness-by-compression **FBC**, to existing fair representation learning approaches that rely on additional penalties – maximum mean discrepancy, cross-entropy of adversarial auditors – to disentangle sensitive attributes from representations. Empirically, we find across four datasets that **FBC** maps data into representations that are state-of-the-art in terms of the fairness-accuracy trade-off they generate.

The work presented in this chapter have been published in [20].

3.1 Fair Information Bottleneck

Consider a population of individuals represented by features $X \in \mathcal{X} \subset [0,1]^{d_x}$ and sensitive attributes in $S \in \mathcal{S} \subset \{0,1\}^{d_s}$, where d_x is the dimension of the feature space and d_s is the dimension of the sensitive attributes space. In this chapter, we do not restrict ourselves to binary sensitive attributes and we allow $d_s > 1$. The objective of fair representation learning is to map the features space $\mathcal{X}$ into a $m-$dimensional representation space $\mathcal{Z} \subset [0,1]^m$, such that (i) Z maximizes the information related to X, but (ii) minimizes the information related to sensitive attributes S. We can express this as

$$\max_{Z} I(X, \{Z, S\}) - \gamma I(Z, S) \tag{3.1}$$

where $I(X, S)$ and $I(X, \{Z, S\})$ denote the mutual information between Z and S and between X and (Z, S), respectively; and $\gamma \geq 0$ controls the fairness penalty $I(Z, S)$.

Existing methods focus on solving directly the problem (3.1) by approximating the mutual information $I(Z, S)$ between Z and S via the cross-entropy of an adversarial auditor that predicts S from Z ([16], [17], [12]) or via the maximum mean discrepancy between Z and S ([15]).

In this chapter, we instead reduce the fair representation learning program (3.1) to an information bottleneck problem that consists of encoding X into a parsimonious code Z, while ensuring that this code Z along with a side channel S allows a good reconstruction of X. The mutual information between X and S can be written as

$$\begin{aligned} I(Z, S) &\overset{(a)}{=} I(Z, \{X, S\}) - I(Z, X|S) \\ &\overset{(b)}{=} I(Z, X) + I(Z, S|X) - I(Z, X|S) \\ &\overset{(c)}{=} I(Z, X) - I(Z, X|S) \\ &\overset{(d)}{=} I(Z, X) - I(X, \{Z, S\}) + I(X, S). \end{aligned}$$

where (a), (b) and (d) use the chain rule for mutual information; and, (c) uses the fact that Z is only encoded from X, so $I(Z, S|X) = 0$. Since the mutual information between X and S does not depend on the code Z, the fair representation learning (3.1) is equivalent to the following fair information bottleneck:

$$\max_{Z} (1 + \gamma) I(X, \{Z, S\}) - \gamma I(Z, X). \tag{3.2}$$

Intuitively, compressing information about X forces the code Z to avoid information redundancy, particularly redundancy related to the sensitive attribute S, since the decoder has direct access to S. Note that there is no explicit constraint in (3.2) to impose independence

between Z and S.

If the representation Z is obtained by a deterministic function of the data X, the mutual information $I(Z, X)$ is equal to the entropy $H(Z)$ of the representation Z. Since the entropy of the data X does not depend on the representation Z, we can replace $I(X, \{Z, S\})$ by $E_{z,s,x} \log(P(x|z, s)$ in the information bottleneck (3.2):

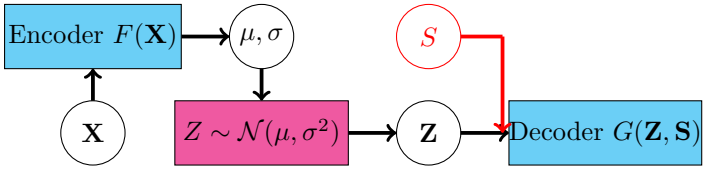
$$\min_{Z} E_{x,z,s}[-\log(P(X|Z, S)] + \beta H(Z), \tag{3.3}$$

where $\beta = \gamma/(\gamma + 1)$. Therefore, the fair representation problem, in its information bottleneck interpretation, can be recast as a rate-distortion trade-off. A lossy compression of the data into a representation Z forces the independence between sensitive attribute and representation but increases the distortion cost measured by the negative log-likelihood of the reconstructed data $E_{x,z,s}[-\log(P(X|Z, S)]$. The parameter β in equation (3.3) controls the competitive objectives of low distortion and fairness-by-compression: the larger β, the fewer the dependencies between Z and S.

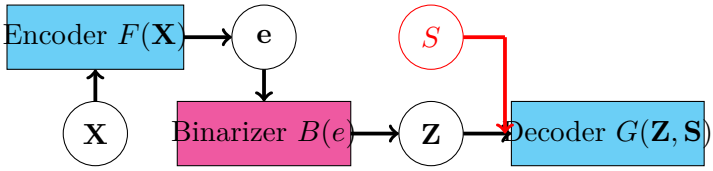
3.2 Proposed Method

There are two avenues to control for $I(Z, X)$ in the information bottleneck (3.2) (see Figure 3.1): (i) adding noise to Z to control the capacity of the information channel between X and Z; or, (ii) storing Z as a bit stream whose entropy is explicitly controlled.

The noisy avenue (i) is a variant of variational autoencoders, so called $\beta-$**VAE** [18], that models the posterior distribution $P(Z|X)$ of Z as Gaussian distributions (see Figure 3.1a). The channel capacity and thus the mutual information between X and Z is constrained by minimizing the Kullback divergence between these posterior distributions and an isotropic Gaussian prior ([86]). Formally, the standard $\beta-$**VAE** ([18]) assumes that the distribution $Q(z|x)$ is Gaussian with mean $\mu(x)$ and standard deviation $\sigma(x)$, and solves for the following minimization problem:



(a) $\beta-$**VAE**



(b) **FBC**

Figure 3.1: Unsupervised methods to obtain fair representations z by compression. Variables are: features $\mathbf{X}$; sensitive attribute $\mathbf{S}$; representation $\mathbf{Z}$. $\beta-$**VAE** generates noisy representations with mean μ and variance σ^2. **FBC** generates binary representations.

$$\min_q E_{x,z\sim q(z|x)}[-\log(P(x|z)] + \beta KL(Q(z|x)||P(z)), \tag{3.4}$$

where $P(z)$ is a isotropic Gaussian prior and $KL(Q(z|x)||P(z))$ is the Kullback-Keibler divergence between $Q(z|x)$ and the prior $P(z)$. [18] show that increasing the value of coefficient β leads to factorized representation Z. In the context of fair representation learning, [15] and [21] use variants of $\beta-$**VAE**, but do not focus on how limiting the channel capacity $I(Z, X)$ could lead to fair representations. Instead, they add further constraints on $I(Z, S)$.

We implement the binary avenue with a method –**FBC** (see Figure 3.1b) – that consists of an encoder $F : \mathcal{X} \to \mathbb{R}^m$, a binarizer $B : \mathbb{R}^m \to \{0, 1\}^m$ and a decoder $G : \{0, 1\}^m \times \mathcal{S} \to \mathcal{X}$. The encoder F maps each data point x into a latent variable $e = G(x)$. The binarizer B binarizes the latent variable e into a bit stream z of length m. The decoder G reconstructs a data point $\hat{x} = G(z, s)$ from the bitstream z and the sensitive attribute s. We model encoder and decoder as neural networks whose architecture varies with the type of data at hand.

The binarization layer controls explicitly the bit allowance of the learned representation and thus forces the encoder to strip redundancies – including sensitive attributes. Binarization is a two step process: (i) mapping the latent variable e into $[0,1]^m$; (ii) converting real values into 0-1 bit. We achieve the first step by applying a neural network layer with an activation function $\overline{z} = (\tanh(e) + 1)/2$. We achieve the second step by rounding $\overline{z}$ to the closest integer 0 or 1. One issue with this approach is that the resulting binarizer B is not differentiable with respect to $\overline{z}$. To sidestep the issue, we follows [65] or [68] and rely on soft binarization during backward passes through the neural network. Formally, during a backward pass we replace z by a soft-binary variable $\dot{z}$:

$$\dot{z} = \frac{exp(-\sigma||\overline{z} - 1||_2^2)}{exp(-\sigma||\overline{z} - 1||_2^2) + exp(-\sigma||\overline{z}||_2^2)},$$

where σ is an hyperparameter that controls the soft-binarization. During the forward pass, we use the binary variable z instead of its soft-binary counterpart $\dot{z}$ to control the bitrate of the binary representation Z [1].

To estimate the entropy $H(z)$, we factorize the distribution $P(z)$ over $\{0,1\}^m$ by writing $z = (z_1, ..., z_m)$ ([65]) and by computing $P(z)$ as the product of conditional distributions:

$$P(z) = \prod_{i=1}^{m} p(z_i|z_{i-1}, z_{i-2}, ..., z_1) \triangleq \prod_{i=1}^{m} p(z_i|z_{.<i}), \tag{3.5}$$

where $z_{.<i} = (z_1, z_2, ..., z_{i-1})$. The order of the bits z_1, ..., z_m is arbitrary, but consistent across all data points. We model P with a neural network Q that predicts the value of each bit z_i given the previous values $z_{i-1}, z_{i-2}, ..., z_1$. With the factorization (3.5), the entropy

[1] In Pytorch, the binarizer returns $(z - \dot{z}).detach() + \dot{z}$.

$H(z)$ is given by

$$\begin{aligned} H(z) &= E_z \left[\sum_{i=1}^{m} -\log(Q(z_i|z_{.<i})) \right] - KL(P||Q) \\ &\leq CE(P,Q), \end{aligned} \tag{3.6}$$

where $CE(P,Q)$ is the cross entropy between P and Q. Therefore, minimizing the cross-entropy loss of the neural network Q minimizes an upper bound of the entropy of the code z. The encoder F and the entropy estimator Q cooperate. The lower the cross-entropy of Q is, the lower is the estimate of the bit rate $H(z)$. Therefore, the encoder has incentives to make the bit stream easy to predict for the neural network Q. Designing a powerful predictor for the bit stream z does not necessary complicate the loss landscape, unlike what could happen with adversarial methods ([87]).

Since the prediction of Q for the i^{th} bit depends on the values of the previous bits z_{i-1}, ..., z_1, the factorization of $P(z)$ imposes a causality relation, where the $(i+1)^{th}$, ..., m^{th} bits should not influence the prediction for z_i. We could enforce this causality constraint by using an iterative method that would first compute $P(z_2|z_1)$, then $P(z_3|z_1, z_2)$,..., and lastly, $P(z_m|z_1, ..., z_{m-1})$. However, it will require $O(m)$ operations that cannot be parallelized. Instead, we follow [65] and enforce the causality constraint by using an architecture for Q similar to PixelCNN ([67], [66]). We model z as a $2D$ $\sqrt{m} \times \sqrt{m}$ matrix and convolve it with one-zero masks, which are equal to one only from their leftmost/top position to the center of the filter. Intuitively, the i^{th} output from this convolution depends only on the bits located to the left and above the bit z_i. The advantage of using a PixelCNN structure, as noted in [65], is to enforce the causality constraint and compute $P(z_i|z_{.<i})$ for all bits z_i in parallel, instead of computing $P(z_i|z_{.<i})$ sequentially from $i = 1$ to $i = m$.

3.3 Experiments

3.3.1 Comparative Methods

The objective of this experimental section is to demonstrate that Fairness by Binary Compression – **FBC** – can achieve state-of-the art performance compared to four benchmarks in fair representations learning: β**-VAE**, **Adv**, **MMD** and **VFAE**.

(i) β**-VAE** ([18]) solves the information bottleneck by variational inference and generates fair representations by adding Gaussian noise which upper-bounds the mutual information between Z and X;

(ii) **MMD** ([55]) uses a deterministic auto-encoder and enforces fairness by minimizing the maximum mean discrepancy ([54]) between the distribution of latent factors Z conditioned on sensitive attributes S;

(iii) **VFAE** ([15]) extends β**-VAE** by adding a maximum mean discrepancy penalty;

(iv) **Adv** ([17]) uses a deterministic auto-encoder as for **MMD**, but enforces the fairness constraint by maximizing the cross-entropy of an adversarial auditor that predicts sensitive attributes S from representations Z.

Although **FBC** shares the deterministic nature of **Adv** and **MMD**, it is more closely related to $\beta-$**VAE**, since $\beta-$**VAE** obtains fairness without explicit constraint on the mutual information of $I(Z, S)$. The main difference between our approach **FBC** and $\beta-$**VAE** is that **FBC** controls the entropy of a binary coding of the data, while $\beta-$**VAE** generates noisy representations and approximates the mutual information $I(Z, X)$ with the Kullback divergence between $Q(z|x)$ and a Gaussian prior $P(z)$. Note that the use of a vanilla $\beta-$**VAE** in a fairness context is novel: only its cousin **VFAE** with an additional MMD penalty has been proposed as a fair representation method.

3.3.2 Experimental Protocol

The overall experimental procedure consists of:

(i) Training an encoder-decoder architecture (F, B, G) along with an estimator of the code entropy Q;

(ii) Freezing its parameters;

(iii) Training an auditing network $Aud : \mathcal{Z} \to \mathcal{S}$ that predicts sensitive attributes from Z.

(iv) Training a task network $T : \mathcal{Z} \to \mathcal{Y}$ that predicts a task label Y from Z.

The encoder-decoder does not access the task labels during training: our representation learning approach is unsupervised with respect to downstream task labels. Datasets are split into a training set used to trained the encoder-decoder architecture; two test sets, one to train both task and auditing networks on samples not seen by the encoder-decoder; one to evaluate their respective performances.

Pareto fronts.

To compare systematically performances across methods, we rely on Pareto fronts that estimates the maximum information that can be attained by a method for a given level of fairness. We approximate information content as the accuracy A_y of the task network T when predicting the downstream label Y. The larger A_y, the more useful is the learned representation for downstream task labels.

We measure how much a representation Z leaks information related to sensitive attributes S by the best accuracy A_s among a set of auditing classifiers $Aud : \mathcal{Z} \to \mathcal{S}$ that predict S from Z. The intuition is that if the distributions $p(Z|S = s)$ of Z conditioned on S do not depend on s, the accuracy of any classifier predicting S from Z would remain near chance level. In the binary case $\mathcal{S} = \{0, 1\}$, comparing A_s to chance level accuracy is a statistical test of independence with good theoretical properties ([88]). If the sensitive classes are furthermore balanced ($P(S = 0) = P(S = 1)$) and the task labels are binary ($\mathcal{Y} = \{0, 1\}$), A_s estimates the worst demographic disparity that can be obtained by a downstream task classifier T that uses Z as an input ([12]). In the general case $S = \{0, 1\}^{d_s}$, the lower A_s compared to chance level, the more independent Z and S are.

To sweep the plan (A_y, A_s) and generate the Pareto fronts, we vary the parameter that controls fairness in each of the competitive methods and for each parameter value, we repeat the experimental protocol 50 times. We then bin the resulting values of A_s and compute the 75%− quantile of A_y attained within each bin.

Rate distortion curves.

To demonstrate further our theoretical insights from section 3.1, we study both rate-distortion and rate-fairness curves of compressing methods **FBC** and $\beta-$**VAE**.

The rate-distortion function $RD(D)$ of an encoder-decoder is measured as the minimum bitrate (in nats) necessary for the distortion $E_{x,z,s}[-\log(p(X|Z,S)]$ to be less than D ([14]):

$$RD(D) = \min I(Z,X) \text{ s.t. } E_{x,z,s}[-\log(p(X|Z,S)] \leq D. \tag{3.7}$$

We introduce a new concept, rate-fairness function $RF(\Delta)$, and define it as the maximum bit rate allowed for the accuracy A_s of the auditing classifier to remain less than Δ

$$RF(\Delta) = \max I(Z,X) \text{ s.t. } A_s \leq \Delta. \tag{3.8}$$

The rate-fairness function captures the maximum information Z can contain while keeping A_s under a given threshold. To obtain both rate-distortion and rate-fairness curves for either our binary compression –**FBC** – or variational –β**-VAE** and **VFAE** – approaches , we vary the value of the parameter β controlling the rate-distortion trade-off and for each value of β, we train the model 50 times with different seeds. For our binary compression method, **FBC**, the bit rate is approximated by the cross-entropy of the entropy estimator Q in (3.6); for variational-based methods, the bit rate is approximated by the Kullback divergence between $Q(z|x)$ and a Gaussian prior. In both cases, the approximation is an upper bound to the true bit-rate (in nats) of Z. We estimate the distortion generated by the encoder-decoder procedure as the $l2$ loss between reconstructed data $\widehat{X} = G(B(F(X)))$ and observed data X.

Robustness to Fairness Metrics

The fair information bottleneck (3.1) aims at controlling the flow of information between Z and S without a prior knowledge of specific downstream task labels Y. Therefore, (3.1) is not designed to control for fairness criteria that rely on labels Y (e.g. equality of odds or opportunites, [38]) or on a specific classifier (e.g. individual fairness, [27]). However, in our experiments, we explore whether the representations generated by **FBC** is empirically robust to different fairness metrics, even though it is not optimized for these metrics to be met.

In particular, we explore whether **FBC** reduces differences $\Delta FP(T)$ in true positive rates of the downstream task network T across demographic groups, where

$$\Delta FP(T) = \sum_{s \in \mathcal{S}} |P(T(x) = 1 | Y = 0, S = s) - P(T(x) = 1 | Y = 0, S \neq s)|. \tag{3.9}$$

3.3.3 Datasets

First, we apply our experimental protocol to a synthetic dataset – DSprites Unfair, [89] – that contains 64 by 64 black and white images of various shapes (heart, square, circle). Images in the DSprites dataset are constructed from six independent factors of variation: color (black or white); shape (square, heart, ellipse), scales (6 values), orientation (40 angles in $[0, 2\pi]$); x- and y- positions (32 values each). The dataset results in $700K$ unique combinations of factor of variations. We modify the sampling to generate a source of potential unfairness. In our experiment, the sensitive attribute is quarternary and encodes which quadrant of the circle the orientation angle belongs to: $[0, \pi/2]$, $[\pi/2, \pi]$, $[\pi, 3/2\pi]$ and $[3/2\pi, 2\pi]$. All factors of variation but shapes are uniformly drawn. When sampling shapes, we assign to each possible combination of attributes a weight proportional to $1 + 10\left[\left(\frac{i_{orientation}}{40}\right)\right)^3 + \left(\frac{i_{shape}}{3}\right)^3\Big]$, where $i_{shape} \in \{0, 1, 2\}$ and $i_{orientation} = \{0, 1, ..., 39\}$. Since shapes and orientation are correlated, a downstream task predicting shapes could risk to

discriminate against some orientation.

Then, we extend our experimental protocol to three benchmark datasets in fair machine learning: **Adults**, **Compas** and **Heritage**. The Adults dataset [2] contains $49K$ individuals and includes information on 10 features related to professional occupation, education attainment, race, capital gains, hours worked and marital status. Sensitive attributes is made of 10 categories that intersect gender and race to which individuals self-identify to. The downstream task label Y correspond to whether an individual earns more than $50K$ per year.

The Compas data [3] contains $7K$ individuals with information related to their criminal history, misdemeanors, gender, age and race. Sensitive attributes intersect self-reported race and gender and result in four categories. The downstream task label Y assesses whether an individual presents a high risk of recidivism.

The Health Heritage dataset [4] contains $220K$ individuals with 66 features related to age, clinical diagnoses and procedure, lab results, drug prescriptions and claims payment aggregated over 3 years. Sensitive attributes are 18 categories that intersect the gender which individuals self-identify to and their reported age. The downstream task label Y relates to whether an individual has a positive Charlson comorbidity Index.

3.3.4 Architectures and Hyperparameters

Encoder-decoders.

For the DSprites dataset, the autoencoder architecture – taken directly from [21] – includes 4 convolutional layers and 4 deconvolutional layers and uses ReLU activations. For the three real world datasets, the encoder and decoder are made of fully connected layers with ReLU activations. Table 3.1 shows more architectural details for each dataset. For all dataset, the hyperparameter σ used for soft-quantization is set to 1. Other hyperparameter values are in Table 3.2.

[2] https://archive.ics.uci.edu/ml/datasets/adult
[3] https://github.com/propublica/compas-analysis/
[4] https://foreverdata.org/1015/index.html

Table 3.1: Architecture details. $Conv2d(i, o, k, s)$ represents a 2D-convolutional layer with input channels i, output channels o, kernel size k and stride s. $ConvT2d(i, o, k, s)$ represents a 2D-deconvolutional layer with input channels i, output channels o, kernel size k and stride s. $Linear(i, o)$ represents a fully connected layer with input dimension i and output dimension o.

Dataset	Encoder	Decoder	Activation
DSprites	Conv(1, 32, 4, 2) Conv(32, 32, 4, 2) Conv(32, 64, 4, 2) Conv(64, 64, 4, 2) Linear(1024, 128) Linear(1024, 128)	Linear(28, 128) Linear(128, 1024) ConvT2d(64, 64, 4, 2), ConvT2d(64, 32, 4, 2) ConvT2d(32, 32, 4, 2) ConvT2d(32, 61, 4, 2)	ReLU ReLU
Adults	Linear(9, 64), Linear(64, 10)	Linear(20, 10) Linear(10, 64), Linear(64,9)	ReLU
Compas	Linear(6, 16), Linear(16, 8)	Linear(12, 8) Linear(8, 16), Linear(16,6)	ReLU
Heritage	Linear(65, 128), Linear(128, 24)	Linear(42, 24) Linear(24, 128), Linear(128,65)	ReLU

Auditor and task classifiers.

Downstream classifiers and fairness auditors are multi-layer perceptrons with varying width (64 to 256 neurons) and depth (2 to 3 hidden layers).

Entropy estimator

The objective of our entropy estimator Q is to predict the $i-th$ bit of the code Z given the values of the previous bits $z_1, ..., z_{i-1}$. We organize the representation into a 2D vector, where the order is arbitrary but consistent across data points. Only the bits before the i^{th} bit (shaded in Figure 3.2) are to be used to estimate $P(z_i|z_{.<i})$. To do so, we convolve the representations with $c \times c$ filters as in Figure 3.2: top filter for the first convolution; bottom filter for the subsequent convolutions. By repeating these convolutions, we are guaranteed that the resulting feature maps (i) satisfy the causality constraint (z_i, ..., z_m do not affect the i^{th} entry in the feature maps); and (ii) all previous bits $z_1, ..., z_{i-1}$ influence the i^{th} entry

Table 3.2: Hyperparameter values for FBC.

Dataset	Number of iterations	Learning rate
DSprites	270K	10^{-4}
Adults	55K	10^{-3}
Compas	22K	10^{-3}
Heritage	55K	0.5×10^{-4}

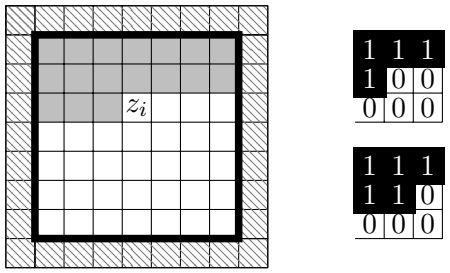


Figure 3.2: Masks for PixelCNN entropy estimator. Left: binary representations organized into a $\sqrt{m} \times \sqrt{m}-$ 2D structure. Strided cells indicate potential padding with zeros. Top right: filter used in the first layer of our PixelCNN entropy estimator. Bottom right: filter used in the subsequent layers of our PixelCNN entropy estimator.

of the feature maps.

In practice, we stack four convolutions with $0-1$ filters as in Figure 3.2. Each convolution layer is followed by batch normalization and ReLU activation. The resulting feature maps are then passed through a traditional convolution layer with a learnable filter and a sigmoid activation function.

3.4 Results and Discussion

3.4.1 Pareto fronts

Figure 3.3 shows the Pareto fronts across five comparative methods for the DSprites and real-world datasets, respectively. Across all dataset, the higher and more leftward the Pareto

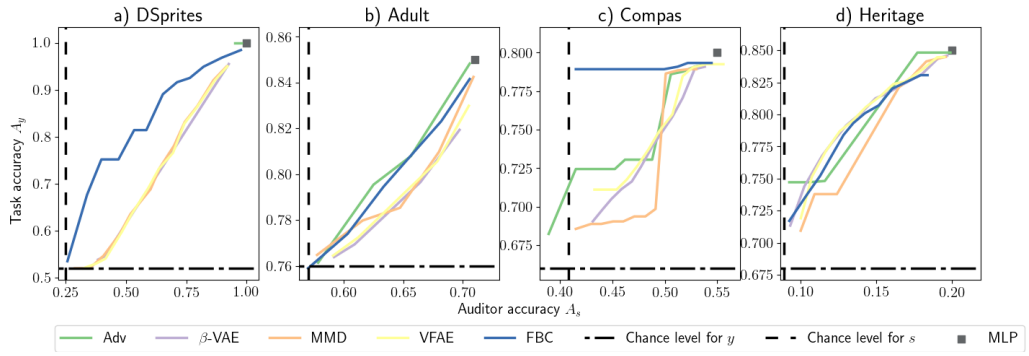


Figure 3.3: Pareto Front for fair representation learning approaches for DSprites and three benchmark datasets. This shows an accuracy-fairness trade-off by comparing the accuracy A_s of auditors that predict sensitive attributes S from representations Z to the accuracy of predicting a task label Y from Z. The dashed horizontal line represents the chancel level of predicting Y. The dashed vertical line represents the chance level of predicting S. Ranges of $x-$ and $y-$ axes varies across datasets.

front, the higher is the task accuracy A_y for a given auditor accuracy A_s and the better is the accuracy-fairness trade-off. From these Pareto fronts, we can draw three conclusions.

First, on all datasets, controlling for the mutual information between Z and X – as in **FBC** and $\beta-$**VAE** – is sufficient to reduce the accuracy A_s of the auditor Aud. This result is consistent with our theoretical observation that minimizing proxies for the information rate $I(Z, X)$ is sufficient to minimize $I(Z, S)$, provided that a side-channel provides the sensitive attributes S to the decoder.

Second, an explicit control of the bit stream encoded in Z achieves a better accuracy-fairness trade-off than floating point approaches. In the $(A_s, A_y)-$ plan, our method, **FBC** achieves either similar (Adults, Heritage) or better (DSprites, Compas) accuracy-fairness trade-off than the variational method $\beta-$**VAE** that controls $I(Z, X)$ by adding noise to the information channel between X and Z. Across all experiments, the Pareto fronts obtained from **FBC** are at least as upward and leftward as for $\beta-$**VAE**.

Third, **FBC** offers a better accuracy-fairness for Compas and DSprites than **MMD**, **VFAE** and **Adv** and is competitive for Adults and Heritage. This is true although **Adv**, **VFAE** and **MMD** control directly the mutual information between Z and S, while **FBC**

controls only $I(Z, X)$. For the Compas dataset, **FBC** shows vast improvement over existing methods since it obtains a representation for which the auditor's accuracy A_s is limited to chance level, while the accuracy A_y of the downstream task is near its level when no fairness constraint is imposed to the encoder-decoder. The adversarial methods do not manage to generate representations with low A_s for the DSprites dataset, possibly because in this higher dimensional problem, the optimization gets stuck in local minima where the adversary has no predictive power, regardless of the encoded representation.

3.4.2 Rate-distortion and rate-fairness

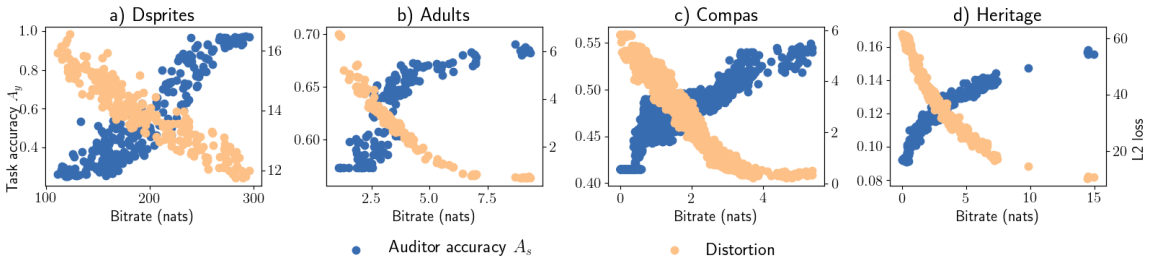


Figure 3.4: Rate distortion/fairness curves. Each dot corresponds to one simulation of **FBC**. Distortion is measured as the $l2$ loss between reconstructed and observed data.

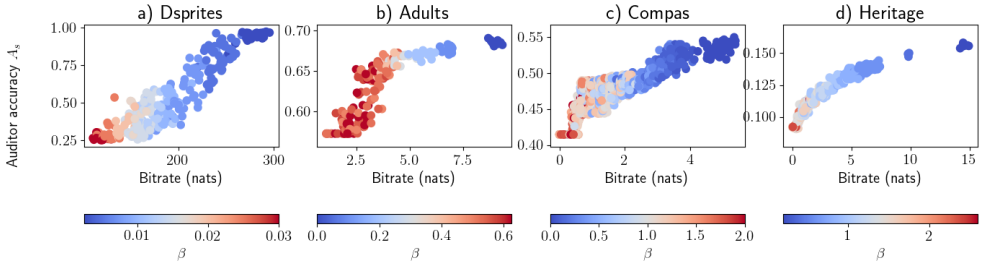


Figure 3.5: Effect of β. This shows the effect of increasing the coefficient β for the code entropy in (3.3) on the bit rate and the auditor's accuracy A_s of representations generated by **FBC**. Changes in β allows to move smoothly along the rate-fairness curve.

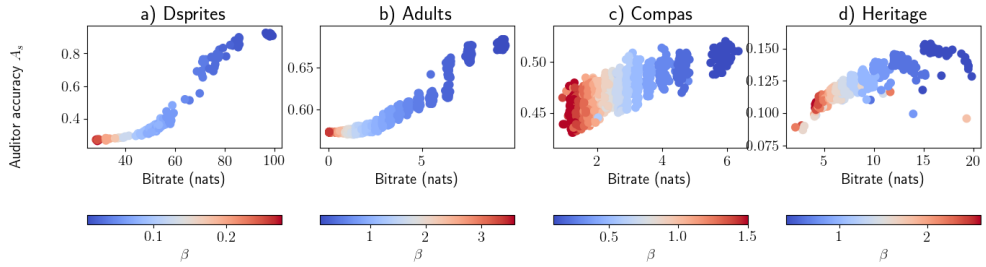


Figure 3.6: Effect of β on the fairness of representation generated by $\beta-$**VAE**. Increasing the coefficient β for the Kullback Leibler divergence in (3.4) reduces the bit rate and the auditor's accuracy A_s of representations generated by $\beta-$**VAE**. Changes in β allows to move smoothly along the rate-fairness curve.

Figure 3.4 confirms that for **FBC**, a lower bit rate estimated by the cross entropy $CE(p, q)$ corresponds to a lower accuracy for the auditing classifier *Aud*. Both rate-distortion (R, D) and rate-fairness (R, Δ) curves show the same monotonic behavior: as distortion moves up along the rate-distortion curves, lack of fairness as measured by A_s moves down. However, for real-word datasets, particularly for Adults and Compas, we observe more variance in the auditor accuracy's A_s given a representation bit rate. We attribute this higher variance to a smaller sample size – 617 for Compas and $3,256$ for Adult on the test set.

Figure 3.5 shows that controlling for the level of compression by increasing the value of β in (3.3) allows moving smoothly along the rate-fairness curve. This is true whether the mutual information $I(Z, X)$ between data and representation is controlled by the bitstream entropy as in **FBC** (Figure 3.5) or by adding a noisy channel as in $\beta-$**VAE** (Figure 3.6). However, binary compression allows a tighter control of the fairness of the representation Z than variational-based methods since in Figure 3.3, for a given auditor's accuracy A_s, **FBC** allows the downstream classifier to achieve a higher accuracy A_y while predicting Y from Z.

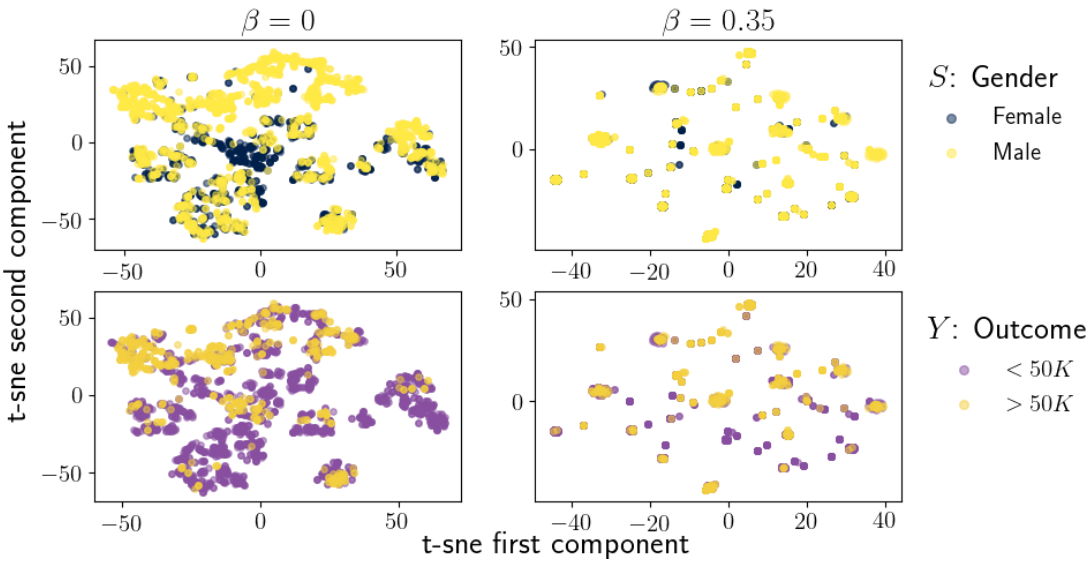


Figure 3.7: Adults – t-SNE visualizations colored with gender (S) and income level (Y) of the representations obtained by **FBC** for different values of the parameter β controlling the compression rate of **FBC**.

3.4.3 Representation Embeddings

Figure 3.7 shows the $t-SNE$ visualizations ([90]) of the representations generated by **FBC** for different values of the parameter β that controls the rate-distortion trade-of in (3.3) for the Adults dataset. Without control of the representation bit rate – $\beta = 0$ – the $t - SNE$ plot shows a cluster of Females that are isolated from males and thus, are easily detected by an auditor that predicts S from Z.

With enough compression – $\beta = 0.35$ – the representation not only looks more parsimonious, but also does not separate Females from Males as much as without compression ($\beta = 0$). In the embeddings space, Females plots are either within clusters of Males or on the edges of these clusters. Moreover, the $t - SNE$ visualizations separate individuals by income level regardless of the compression level, which confirms that the representations generated by **FBC** are useful for classification tasks that predict income level from Z. $t - SNE$ plots for Compas and Heritage are in the technical appendix.

To quantitatively assess the local homogeneity of the sensitive attribute in the embedding

space (Figure 3.7, top), we compute the average distance of females to their top-10 male neighbors and normalize it by the average distance between all individuals. We find that our homogeneity measure decreases by 30% when compressing the data (from left to right plot). But, a similar measure of homogeneity for outcomes (bottom row) decreases only by 8%. This result confirms the visual perception that compression decreases the local homogeneity of sensitive attributes more than the homogeneity of downstream task labels.

Figure 3.8 and 3.9 show $t-SNE$ visualizations for Compas and Heritage, respectively. For both datasets, as β increases, representations not only become more concise, but also hide better the protected group which is made of individuals self-identified as African American for Compas (Figure 3.8, right) or individuals of age 60 and older (Figure 3.9, right) for Heritage.

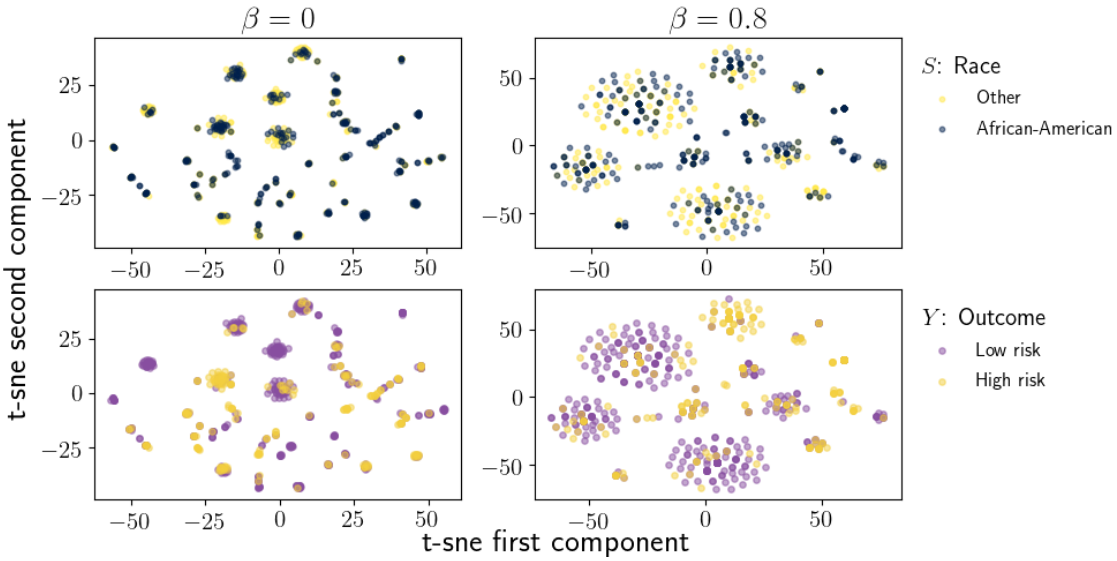


Figure 3.8: Compas – t-SNE visualizations labelled race (S, top) and recidivism risk (Y, bottom) of the representations obtained by **FBC** for different values of the parameter β controlling the compression rate of **FBC**.

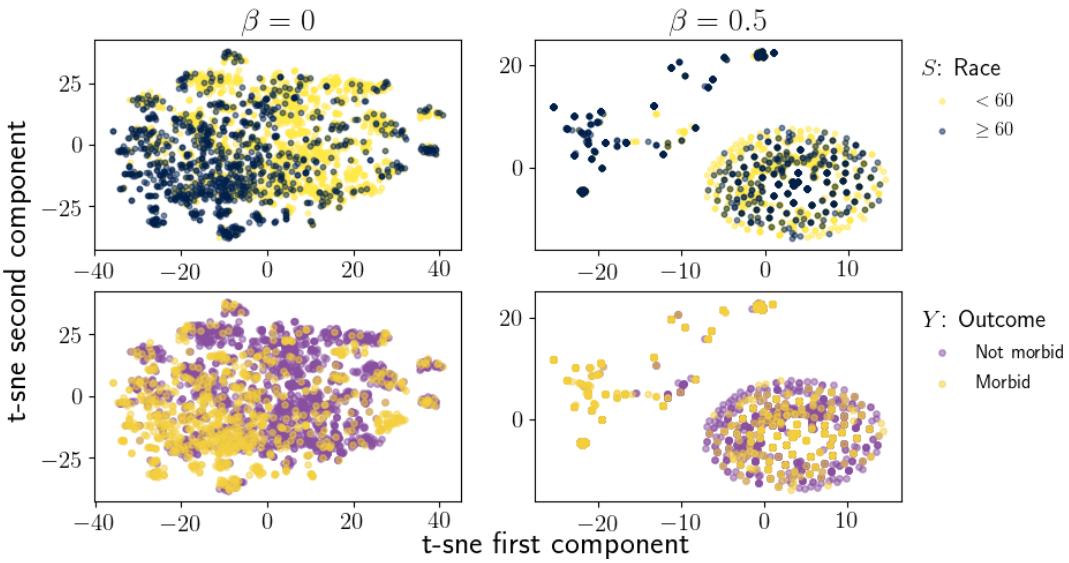


Figure 3.9: Heritage – t-SNE visualizations labelled by age category (S, top) and comorbidity index (Y, bottom) of the representations obtained by **FBC** for different values of the parameter β controlling the compression rate of **FBC**.

3.4.4 Differences in False Positive Rates

Figure 3.10 extends the pareto fronts of Figure 3.3 to additional fairness criteria. Instead of estimating the mutual information between Z and S via A_s, we use differences in false positive rates $\Delta FP(T)$ of the downstream task network T as a fairness criteria. Figure 3.10 plots the median accuracy of T for a given value of $\Delta FP(T)$.

First, all the methods tested – **Adv**, $\beta-$**VAE** and **FBC** – reduce differences of false positive rates across demographic groups on Adult and Compas datasets. This result illustrates, at least on these examples, the ability of fair representation learning to transfer its fairness properties to fairness criteria that models have not been specifically trained to meet. This transfer is all the more remarkable for $\Delta FP(T)$ since this fairness criteria relies on downstream task labels Y that are not observed by fair auto-encoder during its training.

Second, for a given value of $\Delta FP(T)$, **FBC** reaches higher task accuracy A_y than adversarial (**Adv**) and variational ($\beta-$**VAE**) methods. That is, **FBC** appears to generate representations that offer to the task network a better trade-off between similar false rates

across demographic groups (low $\Delta FP(T)$) and accuracy.

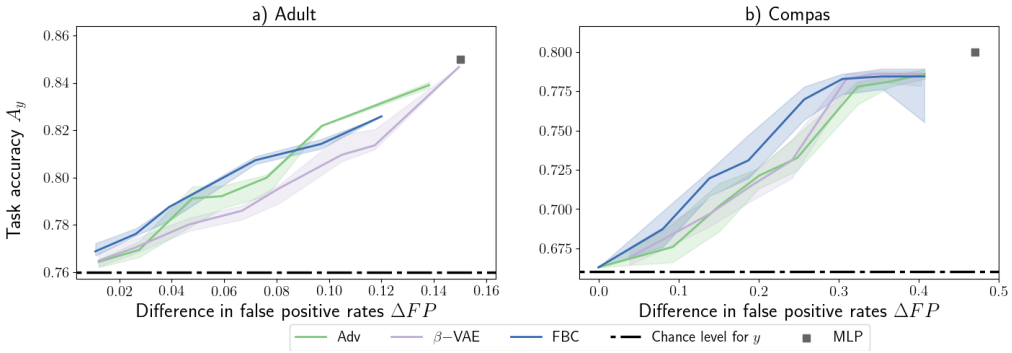


Figure 3.10: Pareto front for representation learning approaches when using difference in false positive rates as a fairness criteria. This plots the median of the accuracy A_y of downstream task networks T for a given value of $\Delta FP(T)$. The dashed horizontal line represents the chancel level of predicting Y. MLP represents the accuracy and $\Delta FP(T)$ obtained by a mutli-layer perceptron trained on the data X instead of its representation Z. Shaded areas represent the range between the 25^{th} and 75^{th} quantiles of accuracy attained by various task networks T for a given value of $\Delta FP(T)$. Ranges of $x-$ and $y-$ axes vary across datasets.

3.5 Conclusion

This chapter introduces a new method – Fairness by Binary Compression (**FBC**) – to map data into a latent space, while guaranteeing that the latent variables are independent of sensitive attributes. Our method is motivated by the observation that in an information bottleneck framework, controlling for the mutual information between representation and data is sufficient to remove unwanted factors, provided that these unwanted factors are direct inputs to the decoder.

Our empirical findings confirm our theoretical intuition: **FBC** offers a state-of-the-art accuracy-fairness trade-off across four benchmark datasets. Moreover, we observe that encoding the representation into a binary stream allows a tighter control of the fairness-accuracy trade-off than limiting the information channel capacity by adding noise. Our results suggest further research into encoder-decoder whose architecture allows a tighter

control of the representation's bit rate and thus, of its fairness.

Chapter 4: Single Shot Fair Representation Learning

In chapter 3, FBC generates a fairness-information trade-off that can only be discovered by training many models. To achieve different points on the fairness-information plane, one must train different models. In this chapter, we first demonstrate that fairness-information trade-offs are fully characterized by rate-distortion trade-offs. Then, we use this key result and propose SoFaiR, a single shot fair representation learning method that generates with one trained model many points on the fairness-information plane.

Our approach relies on a conditional gated mechanism that masks/unmasks representation features at test time depending on the desired fairness / distortion properties. For example, in medical applications, at no additional computational cost, the practitioner can mask/unmask bits depending on whether age or gender are appropriate features for the downstream task at play.

Besides its computational saving, our single-shot approach is, to the extent of our knowledge, the first fair representation learning method that explains what information is affected by changes in the fairness / distortion properties of the representation. Empirically, we find on three datasets that SoFaiR achieves similar fairness-information trade-offs as its multi-shot counterparts.

4.1 Problem Statement

4.1.1 Preliminaries.

Consider a population of individuals represented by features $X \in \mathcal{X} \subset \mathbb{R}^{d_x}$ and sensitive attributes in $S \in \mathcal{S} \subset \{0,1\}^{d_s}$, where d_x is the dimension of the feature space and $d_s \geq 1$ is the dimension of the sensitive attributes space.

The objective of unsupervised fair representation learning is to map features $X \in \mathcal{X}$ into a $d-$dimensional representation $Z \in \mathcal{Z}$ such that (i) Z maximizes the information related to X, but (ii) minimizes the information related to sensitive attributes S. We control for the fairness properties of the representation Z via its mutual information $I(Z,S)$ with S. $I(Z,S)$ is an upper bound to the demographic disparity of any classifier using Z as input [91]. We control for the information contained in Z by constraining a distortion $d(X,\{Z,S\})$ that measures how much information is lost when using a data reconstructed from Z and S instead of the original X. Therefore, fair representation learning is equivalent to solving the following unfairness-distortion problem

$$I(D) = \min_{F} I(Z,S) \text{ s.t. } D(X,\{Z,S\}) \leq D \tag{4.1}$$

where $F : \mathcal{X} \to \mathcal{Z}$ is an encoder, either stochastic or deterministic. The unfairness-distortion function $I(D)$ defines the minimum mutual information between Z and S a user can expect when encoding the data with a distortion less or equal to D. The unfairness-distortion problem (4.1) implies a fairness-information trade-off: lower values of the distortion constraint D degrade the fairness properties of Z by increasing $I(D)$.

Definition 4.1.1. *The unfairness distortion function $I(D)$ is the infimum of mutual information between the representation Z and sensitive attribute S such that the distortion is less than D.*

The objective of this chapter is given a data X to obtain the unfairness-distortion function $I(D)$ with a single encoder-decoder architecture.

4.1.2 Unfairness Distortion Curves.

Rate distortion theory characterizes the minimum average number of bits $R(D)$ used to represent X by a code Z while restricting the expected distortion incurred when reconstructed X from the code to be less than D.We show how to derive unfairness-distortion functions $I(D)$ from rate distortion functions $R(D)$.

Theorem 4.1.1. *Suppose that the distortion is given by* $D(X, \{Z, S\}) = E[-\log(p(x|z, s)]$. *Then, the unfairness distortion function* $I(D)$ *is equal to* $R(D) + D - C$ *if* $\frac{\partial R}{\partial D} \leq -1$ *and* 0 *otherwise.* $C = H(X|S)$ *is a constant that does not depend on* D*, but only on the data* X*. Moreover,* $I(D)$ *is a non-increasing convex function.*

Phase Diagram. The proof of Theorem 4.1.1 is in the appendix. Figure 4.1 shows a graphical interpretation of Theorem 4.1.1 in a (D, R) plane. (D^*, R^*) denotes the point on the rate-distortion curve where $\frac{\partial R}{\partial D} = -1$. For $D \leq D^*$, we can show that the rate distortion curve is above the line defined by $R + D = H(X|S)$ and that the difference between both curves is equal to $I(Z, S)$. For $D > D^*$, the rate-distortion curve is exactly the line $R + D = H(X|S)$ and the unfairness-distortion curve is the horizontal axis. We call the regime $D^* \leq D \leq H(X|S)$ the fair-encoding limit where the distortion is less than its upper limit, but Z is independent of sensitive attribute S. The existence and size of a fair-encoding limit characterizes the possibility to perfectly decorrelate Z from S while representing some information present in the data ($R > 0$).

Information bottleneck. Theorem 4.1.1 implies that fairness-distortion trade-offs are fully characterized by rate-distortion trade-offs. A fundamental result in rate distortion theory ([14, 63]) shows that the rate-distortion function is the solution of the following information bottleneck

$$R(D) = \min_{F} I(X, Z) \text{ s.t } D(X, \{Z, S\}) \leq D. \tag{4.2}$$

By solving this information bottleneck with $D(X, \{Z, S\}) = H(X|Z, S)$ and then, invoking Theorem 4.1.1, we can recover the unfairness-distortion $I(D)$. [20] provide an intuition for this result. Controlling for the mutual information $I(Z, X)$ allows to control for $I(Z, S)$ because an encoder would not waste code length to represent information related to sensitive attributes, since sensitive attributes are provided directly as an input to the decoder. We

can write the information bottleneck in its Lagrangian form as

$$\min_{F} \beta I(Z, X) + E[-\log p(x|z, s)] \tag{4.3}$$

The coefficient β relates to the inverse of the slope of the rate-distortion curve: $\frac{\partial R}{\partial D} = -1/\beta$. Each value of β generates a different point along the rate-distortion curve and thus, by Theorem 4.1.1 a different point along the unfairness-distortion curve. Higher values of β lead to representations with lower bit rate and lower mutual information with S. To explore a unfairness-distortion curve, existing multi-shot strategies are prohibitively expensive as they learn a new encoder F for each value of β. This is computationally expensive as the practitioner needs to train a new model for each point on the unfairness-distortion curve, which limits its ability to explore the entire curve. Moreover, there is no straightforward explanation on how changes in β affect the representation generated by the encoder.

4.2 Method: Single-Shot Unfairness-Distortion Curves.

We propose a single-shot method, SoFaiR, to generate with one model as many points as desired on the unfairness-distortion curve.

An encoder $F : \mathcal{X} \to \{0, 1\}^{d \times r}$ common to all values of β encodes the data into a d dimensional latent variable $e \in [0, 1]^d$. We quantize each dimension e_j of the $d-$dimensional latent variable with a resolution $r_j(\beta)$: we transform e_j into a quantized representation $z_j(\beta) = [e * r(\beta)]/r(\beta)$, where [.] denotes the rounding-up operation and $r(.)$ is a decreasing function of β.

4.2.1 Interpretability

To maintain an interpretable relation between $z(\beta)$ and $z(\beta^{'})$ for $\beta^{'} < \beta$, we write $r_j(\beta) = 2^{a_j(\beta)}$, where $a_j(.)$ is a decreasing function of β for $j = 1, ..., d$. Each dimension $z_j(\beta)$ of the quantized representation is then encoded into $a_j(\beta)$ bits. Moreover, for $\beta^{'} < \beta$, each

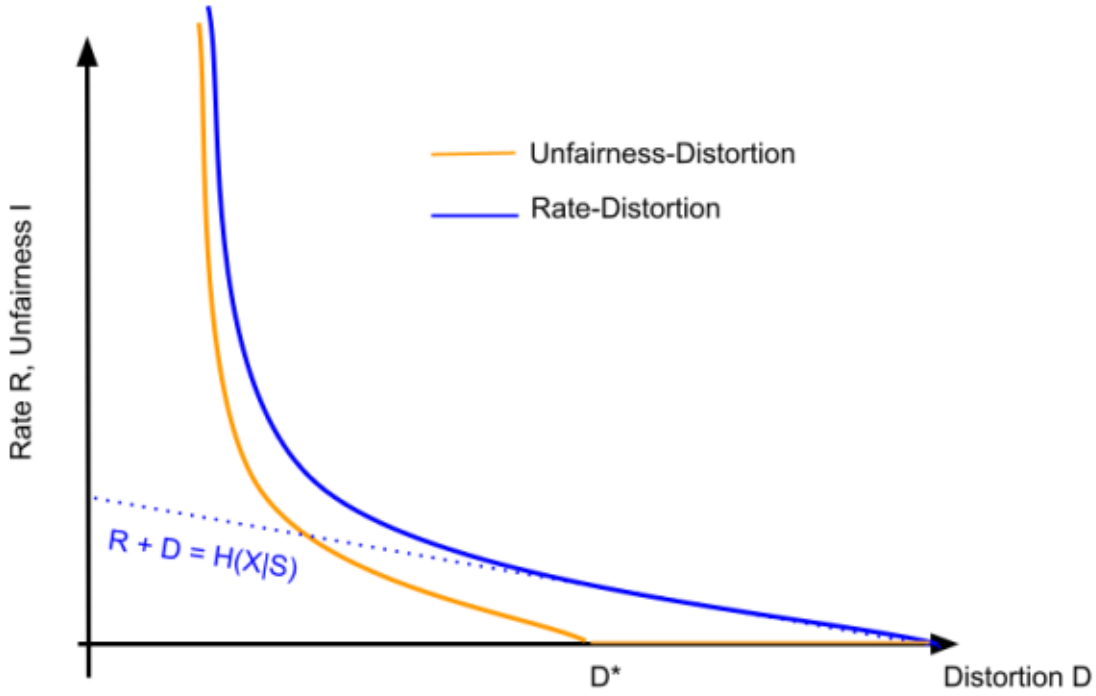


Figure 4.1: Unfairness-distortion curves $I(D)$ vs. rate-distortion curve $R(D)$. The unfairness distortion $I(D)$ can be deduced from the rate-distortion $R(D)$ curve by a downward shift equal to $D - H(X|S)$ if the distortion is less than D^*.

dimension j of the representation $z(\beta^{'})$ is made of the same $a_j(\beta)$ bits as $z_j(\beta)$, followed by $a_j(\beta^{'}) - a_j(\beta)$ additional bits. Each dimension $z_j(\beta)$ of the quantized representation is encoded into $a_j(\beta)$ bits $b_{j,1}, b_{j,2}, ..., b_{j,a(\beta)}$, where $b_{j,l} \in \{0, 1\}$ for $l = 1, ..., a_j(\beta)$. For $\beta^{'} < \beta$ and for $j = 1, ..., d$, we have

$$z_j(\beta^{'}) = z_j(\beta) + \sum_{l=a_j(\beta)}^{a_j(\beta^{'})} b_{j,l} 2^{-l}. \tag{4.4}$$

Therefore, we have a tractable and interpretable relation between $z_j(\beta^{'})$ and $z_j(\beta)$.

This construction allows relaxing fairness constraints and decreasing distortion by unmasking additional bits for each dimension of the representation. Figure 4.2 shows an example for a 2-dimensional representation. A user who has released $z1$ with high distortion and low mutual information $I(Z, S)$ reduces distortion at the cost of fairness by unmasking one bit for the first dimension and two bits for the second and by generating $z2$.

4.2.2 Quantization

We assign a maximum number of bits $A > 0$ to encode each dimension of the representation. We apply a function h_e to map the $d-$dimensional latent variable e into $[0, 1]^{d \times A}$ and then, apply a rounding-up operator $[h_e(e)]$ to generate a $d \times A$ matrix, each row encoding a dimension of the representation with A bits (see Figure 4.2 with $A = 3$). For each dimension j, we implement $a_j(.)$ by applying a function h_a to map e into a $d-$dimensional vector of $\mathbb{R}+^d$ and by computing

$$a_j(\beta) = A\,[1 - \tanh(h_a(e)_j \beta))]. \tag{4.5}$$

For each value of β and each row of the matrix $[h_e(e)]$, we mask all the entries in position $l > a_j(\beta)$: for each row j and each column l, we compute a soft mask

$$m_{j,l}(\beta) = \sigma\left(a_j(\beta) - l\right), \tag{4.6}$$

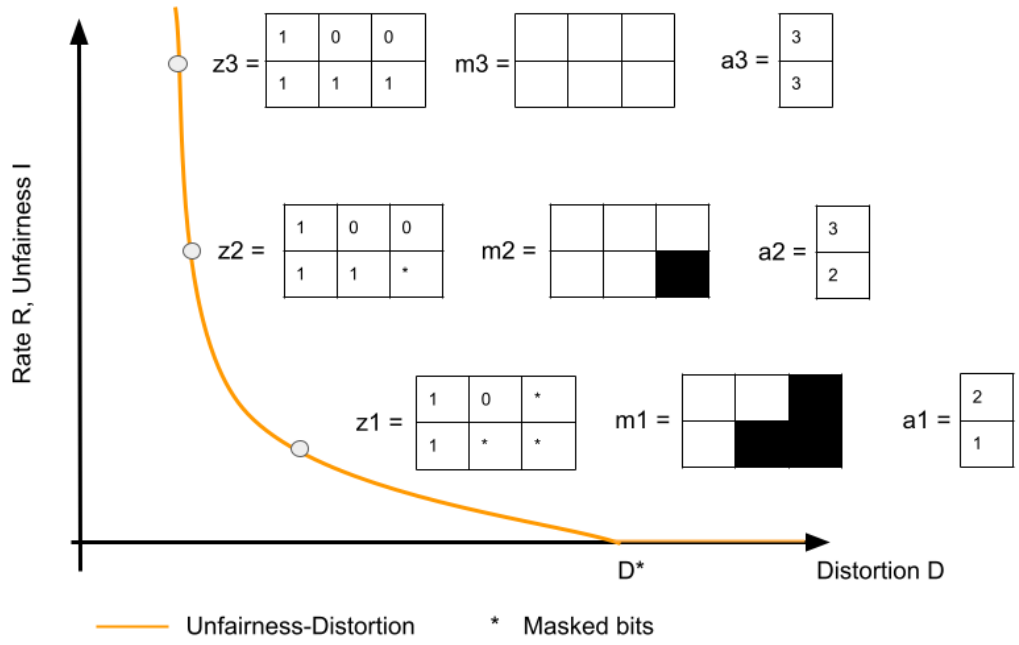


Figure 4.2: SoFaiR generates interpretable shifts along the unfairness-distortion curve. For a point $z1$, SoFair learns a mask $m1$ that hides bits on the tails of each dimension of the representation. By relaxing the mask to first $m2$ then $m3$, the number of bits used to represent the data increases from $a1$ to $a2$ and then $a3$; and, the representation moves to $z2$ then $z3$, which reduces the distortion at the expenses of degraded fairness properties. $z1$, $z2$ and $z3$ only differ by their masked bits (black squares).

where σ denotes a sigmoid activation; and then, we apply a rounding operator $[m_{j,l}(\beta)]$ to our soft mask.

The binarization caused by the rounding operation [.] is not differentiable. We follow [68] and use a gradient-through approach that replaces [.] by the identity during the backward pass of back-propagation, while keeping the rounding operation during the forward pass.

4.2.3 Entropy estimation.

To estimate the entropy of the representation Z, we use an auto-regressive factorization to write the discrete distribution $P(z|\beta)$ over the representation Z

$$P(z|\beta) = \prod_{j=1}^{d} P(z_j|z_{.<j}, \beta), \tag{4.7}$$

where the order of the dimension $j = 1, ..., d$ is arbitrary and $z_{.<j}$ denotes the dimension between 1 and $i-1$. This is similar to the entropy estimation in [20]. However, unlike [20] who model P as a discrete distribution, we follow a more standard approach in rate-distortion [68, 69] and approximate the discrete distribution $p(z_j|z_{.<j}, \beta)$ by a continuous distribution $q(z_j|z_{.<j}, \beta)$ such that the probability mass of q on the interval $[z_j - 1/2^{a_j(\beta)}, z_j + 1/2^{a_j(\beta)}]$ is equal to $p(z_j|z_{.<j}, \beta)$. Therefore,

$$\begin{aligned}
H(z|\beta) &= -\sum_{j=1}^{d} E\left[\log p(z_j|z_{.<j}, \beta)\right] \\
&= -\sum_{j=1}^{d} E\left[\log\left(\int_{-1/2^{a_j(\beta)}}^{1/2^{a_j(\beta)}} q(z_j + u|z_{.<j}, \beta)du\right)\right] \\
&+ KL\left(p||\int_{-1/2^{a_j(\beta)}}^{1/2^{a_j(\beta)}} q(z_j + u|z_{.<j}, \beta)du\right) \\
&\overset{(a)}{\leq} -\sum_{j=1}^{d} E\left[\log\left(\int_{-1/2^{a_j(\beta)}}^{1/2^{a_j(\beta)}} q(z_j + u|z_{.<j}, \beta)du\right)\right],
\end{aligned} \tag{4.8}$$

where (a) uses the non-negativity of the Kullback-Leibler divergence KL between the true distribution $p(z|\beta)$ and its approximation $q(z|\beta)$ once convolved with a uniform distribution over $[-1/2^{a_j(\beta)}, 1/2^{a_j(\beta)}]$.

We follow [92] and for each $j = 1, ..., d$ we model $q(.|z_{.<j}, \beta)$ as a mixture of K logistic distributions with means $\mu_{j,k}(\beta)$, scales $\gamma_{j,k}(\beta)$ and mixtures probability $\pi_{j,k}(\beta)$, which allows to compute exactly the integral term in (4.8). Specifically, we compute

$$\mu_{j,k} = \mu^0_{j,k}(\beta) + w^{\mu}_{j,k}(\beta)\Gamma_j \odot z_j, \tag{4.9}$$

and

$$\log(\gamma_{j,k}) = \gamma^0_{j,k}(\beta) + w^{\gamma}_{j,k}(\beta)\Gamma_j \odot z_j, \tag{4.10}$$

where $\mu^0_{j,k}(.), \gamma^0_{j,k}()$ are functions from $[0, 1]$ to $\mathbb{R}$; $w^{\mu}_{jk}()$ and $w^{\gamma}_{j,k}()$ are functions from $[0, 1]$ to $\mathbb{R}^d$; and, $\Gamma_j = (1, 1, .., 1, 0, ...0)$ is a $d-$ dimensional vector equal to one for entry before j and zero otherwise. Γ_j guarantees that the distribution $q(.|z_{.<j})$ is conditioned only on $z_{.<j}$ only and not on any $z_{j'}$ for $j^{'} \geq j$.

The use of logistic distribution allows to compute the upper bound in (4.8) as $H_q(z|\beta)$ where $H_q(z|\beta)$ is given by

$$\begin{aligned} -\sum_{j=1}^{d} E\Bigg[\log\Bigg(\sum_{k=1}^{K} \pi_{j,k}\sigma\left(\frac{z_j + \mu_{j,k}(\beta)}{\gamma_{j,k}(\beta)} + \frac{1}{2^{a_j(\beta)}}\right) \\ -\sigma\left(\frac{z_j + \mu_{j,k}(\beta)}{\gamma_{j,k}(\beta)} - \frac{1}{2^{a_j(\beta)}}\right)\Bigg)\Bigg]. \end{aligned} \tag{4.11}$$

The adaptive information bottleckneck can be written as:

$$\min_{G, h_e, h_a, \mu, \gamma, w, \pi} E[-\log p(x|G(z, s, \beta)) + \beta H_q(z|\beta)]. \tag{4.12}$$

4.3 Experiments

We design our experiments to answer the following research questions: (RQ1) Does SoFaiR generate in a single-shot unfairness-distortion curves comparable to the ones generated by multi-shot models? (RQ2) Do representations learned by SoFaiR offer to downstream tasks a fairness-accuracy trade-off on par with state-of-the-art multi-shots techniques in unsupervised fair representation learning? (RQ3) What information is present in the additional bits that are unmasked as we move up the unfairness-distortion curve?

4.3.1 Datasets

We validate our single-shot approach with three benchmark datasets in fair machine learning: **DSprite-Unfair**, **Adults** and **Heritage**.

DSprite Unfair is a variant of the DSprites data [1], a synthetic dataset that contains 64 by 64 black and white images of various shapes (heart, square, circle). Images in the DSprites dataset are constructed from six independent factors of variation: color (black or white); shape (square, heart, ellipse), scales (6 values), orientation (40 angles in $[0, 2\pi]$); x- and y- positions (32 values each). We modify the sampling to generate a source of potential unfairness and use as sensitive attribute a variable that encodes whether the shape has an orientation within $[0, \pi/2)$, $[\pi/2, \pi)$, $[\pi, 3/2\pi)$ or $[3/2\pi, 2\pi)$.

The **Adults** dataset [2] contains $49K$ individuals with information to professional occupation, education attainment, capital gains, hours worked, race and marital status. We consider two variants of the Adults dataset: **Adults-Gender** and **Adults-Gender-Race**. In Adults-Gender we define as sensitive attributes the gender to which each individual self-identifies to. In Adults-Gender-Race, we define as sensitive attribute an intersection of the gender and race an individual self-identifies to. For both Adults-Gender and Adults-Gender-Race, the downstream task label Y correspond to whether individual income exceeds $50K$ per year.

[1] `https://github.com/deepmind/dsprites-dataset/`
[2] https://archive.ics.uci.edu/ml/datasets/adult

The **Health Heritage** dataset [3] contains $95K$ individuals with 65 features related to clinical diagnoses and procedure, lab results, drug prescriptions and claims payment aggregated over 3 years (2011-2013). We define as sensitive attributes a $18-$ dimensional variable that intersects the gender which individuals self-identify to and their reported age. The downstream task label Y relates to whether an individual has a non-zero Charlson comorbidity Index, which is an indicator of a patient's 10-year survival rate.

4.3.2 Unfairness-distortion curves.

To plot unfairness-distortion curves, we estimate the distortion as the l_2- loss between reconstructed and observed data, which is equal to $E_{x,z,s}[-\log p(x|z,s)]$ (up to a constant) if the distribution of $p(X|Z,S)$ X is an isotropic Gaussian. Moreover, l_2- loss is a common measure of distortion in the rate-distortion literature.

We also approximate the mutual information $I(Z,S)$ with an adversarial lower bound. For any approximation $q(s|Z)$ of $p(s|Z)$, we have

$$\begin{aligned} I(Z,S) &= H(S) - H(S|Z) \\ &= H(S) - E_{s,z}[-\log q(s|z)] + KL(p(s|z)||p(s|z) \\ &\geq H(S) - E_{s,z}[-\log q(s|z)], \end{aligned} \tag{4.13}$$

where the inequality comes from the non-negativity of the Kullback-Leibler divergence $KL(p|q)$. Therefore, we lower bound $I(Z,S)$ with

$$H(S) - \min_q E_{s,z}[-\log q(s|z)], \tag{4.14}$$

where the minimum is taken over classifiers that predict S from Z. This lower bound leads to train adversary-based autoencoder in fair representation learning (e.g. [16,17,93]). In this

[3] https://foreverdata.org/1015/index.html

chapter, we only use the lower bound for a post mortem evaluation of the fairness-distortion trade-off generated by SoFaiR.

In practice, we train a set of 5 classifiers $c : \mathcal{Z} \to \mathcal{S}$ modeled as fully connected neural networks and use their average cross-entropy to estimate the right hand side of (4.13).

4.3.3 Area under unfairness-distortion curves.

We compare the performance of different fair representation learning methods in terms of unfairness-distortion curves. Besides visual inspection of unfairness-distortion curves, we propose to compute the area under unfair-distortion curve, AUFDC, to allow for quantitative comparison between fair representation methods. A model that achieves a lower AUFDC generates representations that achieve lower $I(Z, S)$ for a given level of distortion.

We compute the area under unfair-distortion curve, AUFDC, as follows:

- Since we only generate a finite number of points, empirical fairness-distortion curves do not have to exhibit a perfectly decreasing and smooth behavior. Therefore, to compute our AUFDC metric, we first filter out the points on the curve that have higher distortion than points with higher $I(Z, S)$. That is, for any point (D, I), we remove all points $(D^{'}, I^{'})$ for which $D^{'} > D$ and $I^{'} > I$.

- We estimate the largest obtainable mutual information I_{max} between Z and S as

$$H(S) - \min_{c} E_{s,x}[-\log c(s|x)], \tag{4.15}$$

  where $c : \mathcal{X} \to \mathcal{S}$ are classifiers that predict S from the data X. That is, we use an adversarial estimate of $I(X, S)$ and use this estimate as I_{max}, since by the data processing inequality [63], $I(Z, S) \leq I(X, S)$.

- For models that do not reach $I(Z, S) = 0$, we compute the distortion D_{max} obtained by generating random representations; and then, we add to the AUFDC score, the area $I_{min} \times (D_{max} - D_{min})$ where (D_{min}, I_{min}) is the point the furthest on the bottom

right of the unfairness-distiortion curve achieved by the model.

- To allow comparison across datasets, we normalize the value of AUFD by the area of the rectangle $[0, D_{max}] \times [0, I_{max}]$.

4.3.4 Comparative Methods.

Most methods in fair representation learning are supervised since they are tailored toward a specific downstream classification task. We follow [20] and re-purpose each comparative method to an unsupervised setting where we replace the cross-entropy of the downstream classification task by our measure of distortion $E[-\log p(x|Z, s)]$.

We compare SoFaiR with the following fair representation methods:

- **LATFR** ([16, 17]) controls the mutual information $I(Z, S)$ via the cross-entropy of an adversary that predicts S from Z. LATFR controls the fairness properties of the representation Z with a single parameter in $\{0.1, 0.2, 0.3, 0.5, 0.7, 1.0, 1.5, 2.0, 3.0, 4.\}$ as prescribed in the original paper.

- **MaxEnt-ARL** [94] is a variant of LATFR that replaces the cross-entropy of the adversary with the entropy of is prediction. The fairness properties of Z are controlled by a single parameter that we vary between 0 and 1 in steps of 0.1, between 1 and 10 in steps of 1 and then between 10 and 100 in steps of 10.

- $\beta - VAE$[18, 20] controls for $I(Z, S)$ by controlling the Kullback-Leibler divergence between $p(z)$ and an isotropic Gaussian prior. The fairness properties of the representations are controlled via the coefficient β on the Kullback-Leibler divergence term: larger values of β force Z be more noisy, reduce the capacity of the channel between the data and the representation and thus, the mutual information $I(Z, X)$. We vary the value of β between 0 and 1 in steps of 0.05.

- **CVIB** [95] controls for $I(Z, S)$ via both the Kullback-Leibler divergence between

$p(z)$ and an isotropic Gaussian prior (as in β-VAE) and an information-theory upper bound. The first term is controlled by a parameter β that takes values in $\{0.001, 0.01, 0.1\}$; the second is controlled by a parameter λ that vary between 0.01 to 0.1 in steps of 0.01 and 0.1 to 1.0 in steps of 0.1 [91].

- **MSFaiR** reproduces SoFaiR, but solves the rate-distortion problem (4.12) separately for different values of $\beta \in [0, 1]$.

4.3.5 Pareto Fronts

We construct Pareto fronts that compare the unfairness properties of the representation to the accuracy A_y of a downstream task classifier that predicts a downstream label Y from Z. Critically in our unsupervised setting, we do not provide the labels Y to encoder-decoders. To match existing benchmarks, we measure the unfairness properties of the representation with the average accuracy A_s of auditing classifiers that predict S from Z. The higher A_y for a given A_s, the better is the fair representation method.

To generate Pareto fronts, we implement the following protocol:

- Train an encoder-decoder architecture and freeze its parameters;
- Train an auditing classifier $c : \mathcal{Z} \to \mathcal{S}$ to predict S from Z;
- Train a downstream task classifier $T : \mathcal{Z} \to \mathcal{Y}$ to predict a task Y from Z.

The encoder-decoder does not access the task labels during training and our representation learning framework remains unsupervised with respect to downstream task labels. Critically in our unsupervised setting, we do not provide the labels Y to encoder-decoders. All comparative methods share the same encoder-decoder architecture and differ only by how they control the mutual information between Z and S.

Pareto fronts differ from unfairness-distortion curves by how tailored they are to a specific downstream task. In unsupervised fair representation learning, unfairness-distortion

curves is a general purpose method that allows the practitioner to estimate the fairness-information trade-off without any particular downstream task –either classification or regression – in mind.

4.3.6 Architectures

Table 4.1: Architecture details. $Conv2d(i, o, k, s)$ represents a 2D-convolutional layer with input channels i, output channels o, kernel size k and stride s. $ConvT2d(i, o, k, s)$ represents a 2D-deconvolutional layer with input channels i, output channels o, kernel size k and stride s. $Linear(i, o)$ represents a fully connected layer with input dimension i and output dimension o. Activations are not applied on the last layer of the decoder.

Dataset	Encoder	Decoder	Activation
DSprites	Conv(1, 32, 4, 2), Conv(32, 32, 4, 2) Conv(32, 64, 4, 2) Conv(64, 64, 4, 2) Linear(1024, 128)	Linear(28, 128), Linear(128, 1024) ConvT2d(64, 64, 4, 2), ConvT2d(64, 32, 4, 2) ConvT2d(32, 32, 4, 2), ConvT2d(32, 1, 4, 2)	ReLU ReLU ReLU ReLU ReLU
Adults	Linear(9, 128), Linear(128, 128), Linear(128, 8)	Linear(8, 128), Linear(128, 128), Linear(128, 9),	ReLU ReLU
Heritage	Linear(65, 256), Linear(256, 256), Linear(256, 12)	Linear(12, 256), Linear(256, 256), Linear(256, 65),	ReLU ReLU

Table 4.2: Hyperparameter values for SoFaiR / MSFaiR.

Dataset	Number of iterations	Learning rate
DSprites	546K	0.3×10^{-4}
Adults	27K	0.3×10^{-4}
Heritage	74K	0.3×10^{-4}

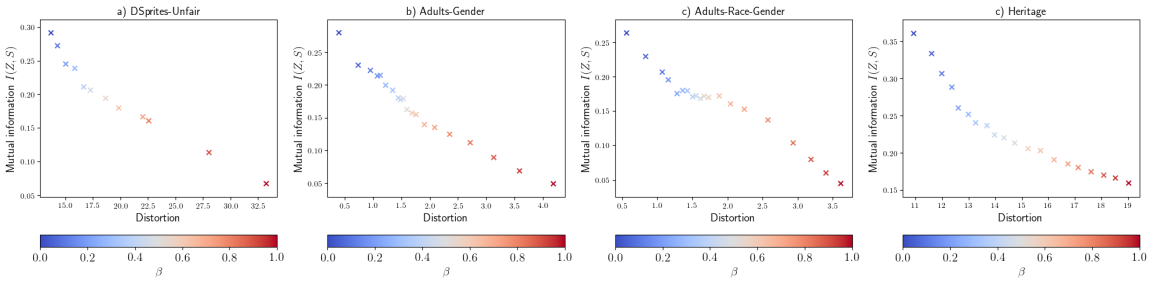


Figure 4.3: Unfairness-Distortion curves for a) DSprites, b) Adults-Gender, c) Adults-Race-Gender(left) and d) Heritage.

Encoder-decoders. For the DSprites dataset, the autoencoder architecture – taken directly from [21] – includes 4 convolutional layers and 4 deconvolutional layers and uses ReLU activations. For Adults and Heritage, the encoder and decoder are made of fully connected layers with ReLU activations. Table 4.1 shows more architectural details for each dataset. Moreover, means μ, scales γ and mixture probabilities π are modeled as fully connected linear layers with input dimension 1 and output dimension d, i.e. the dimension of the latent space. We choose $K = 5$ logistic distributions in the mixture. We also set the maximum number of bits per dimension, A, to be equal to 8. Other hyperparameter values are in Table 4.2.

Auditor and task classifiers. Downstream classifiers and fairness auditors are multilayer perceptrons with 2 hidden layers of 256 neurons each. Learning rates for both auditing and downstream tasks are set to 0.001

4.4 Results

4.4.1 Single Shot Fairness-Distortion Curves.

4.4.2 RQ1: Single Shot Fairness-Distortion Curves.

Figure 4.3 shows SoFaiR's unfairness-distortion curves for DSprites (left), Adults-Gender (middle left), Adults-Gender-Race (middle right) and Heritage (right). By increasing at

test time the value of β, the user can smoothly move down the unfairness-distortion curve: values of β close to zero lead to low distortion - high $I(Z, S)$ points; values of β close to one lead to higher distortion - low $I(Z, S)$ points. Figure 4.3 demonstrates that a solution to the adaptive bottleneck (4.12) allows one single model to capture different points on the unfairness-distortion curve. This result is consistent with Theorem 4.1.1 and illustrates that controlling for the bit rate of Z via its entropy $H(Z)$ is sufficient to control for $I(Z, S)$.

Table 4.3: Area under the unfairness-distortion curve of single-shot (SoFaiR) versus multi-shot (MSFaiR) fair representation learning methods. Lower (↓) is better. This shows that SoFaiR provides unfairness-distortion curves with similar AUFDC.

Dataset	Model	AUFDC (↓)
DSprites-UnfaiR	SoFaiR	0.21
	SoFaiR-NOS	0.25
	MSFaiR	**0.14**
Adults-Gender	SoFaiR	**0.32**
	SoFaiR-NOS	0.58
	MSFaiR	0.35
Adults-Gender-Race	SoFaiR	**0.30**
	SoFaiR-NOS	0.53
	MSFaiR	0.36
Heritage	SoFaiR	0.62
	SoFaiR-NOS	0.73
	MSFaiR	**0.56**

Ablation study. UFDC scores in Table 4.3 show that SoFaiR is competitive with its multi-shot counterpart: SoFaiR outperforms MSFaiR for Adults-Gender and Adults-Gender-Race (lower AUFDC), but is slightly outperformed for Heritage and DSprites-Unfair (higher AUFDC).

On the other hand, SoFaiR unambigously outperforms SoFaiR-NOS, a model similar to SoFaiR but with a decoder that does not use the sensitive attribute S as side-channel. The relation between unfairness-distortion and rate-distortion curves in Theorem 4.1.1 is tractable only if we use $E[-\log(p(x|z, s)]$ as a measure of distortion and does not hold if

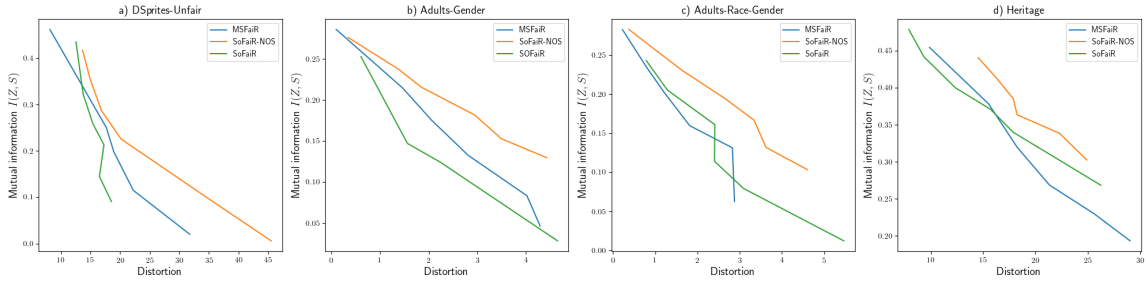


Figure 4.4: Ablation study for a) DSprites, b) Adults-Gender, c) Adults-Race-Gender(left) and d) Heritage. This compares unfairness-distortion curves generates by our single shot approach SoFaiR to the ones generated by its multi-shot counterpart MSFaiR; and, to the ones generated by SoFaiR-NOS, which is similar to SoFaiR but for the decoder that does not receive the sensitive attribute S as an input.

we use $E[-\log(p(x|z)]$ instead and the decoder does not receive S as side channel.

In Figure 4.4, we plot the unfairness-distortion curves that correspond to the AUFDC that we report in Table 4.3 of the main text. We report the median value of distortion for a given level of mutual information $I(Z,S)$, where the median is taken over 10 similar models trained with different seeds. The lower is the distortion for a given value of $I(Z,S)$, the better the fair representation learning method. Conclusions from Figure 4.4 are similar to the ones from Table 4.3. SoFaiR outperforms MSFaiR for Adults-Gender and Adults-Race-Gender at all values of $I(Z,S)$, while it is outperformed by MSFaiR for Heritage at low values of $I(Z,S)$. Moreover, for all datasets, SoFaiR outperforms SoFaiR-NOS, which confirms that rate-distortion solutions to fair representation learning need the decoder to receive the sensitive attribute as a side channel.

Computational costs. Table 4.4 compares the computational costs of SoFaiR and MSFaiR. We average the cpu and gpu times of a training step over 10 profiling cycles and the number of training epochs. We perform the experiment on a AMD Ryzen Threadripper 2950X 16-Core Processor CPU and a NVIDIA GV102 GPU. The average computing cost of a training step is similar for SoFaiR and MSFaiR since both methods rely on similar architecture. However, SoFaiR's computational costs remain constant as the number of

points on the unfairness-distortion curve increases, while MSFaiR's costs increase linearly. For example, 16 points for the DSprites-Unfair require about 137 hours of running time with MSFaiR and only 8 hours with SoFaiR.

Table 4.4: Area under the unfairness-distortion curve and computational costs of single-shot (SoFaiR) versus multi-shot (MSFaiR) fair representation learning methods. Lower (↓) is better. This shows that SoFaiR provides unfairness-distortion curves with similar AUFDC as MSFaiR, but at much lower computational costs.

Dataset	Model	Average per step CPU/GPU (ms)	Total time (10^6 ms): CPU/GPU (↓) 4 points	8 points	16 points
DSprites-UnfaiR	SoFaiR	79 ± 1.2 / 55 ± 0.2	**18.5/13.0**	**18.5/13.0**	**18.5/13.0**
	MSFaiR	76 ± 3.2 / 55 ± 0.3	71.4/52.1	142.9/104.2	285.8/208.0
Adults-Gender	SoFaiR	$91 \pm 3.3/6 \pm 0.0$	**2.3/0.1**	**2.3/0.1**	**2.3/0.1**
	MSFaiR	$92 \pm 1.0/6 \pm 0.0$	9.4/0.6	18.9/1.1	37.7/2.3
Adults-Gender-Race	SoFaiR	$92 \pm 4.3/6 \pm 0.0$	**2.4/0.1**	**2.4/0.1**	**2.4/0.1**
	MSFaiR	$90 \pm 4.0/6 \pm 0.0$	9.1/0.6	18.3/1.1	36.6/2.3
Heritage	SoFaiR	$125 \pm 3.0/8.6 \pm 1.6$	**3.7/0.3**	**3.7/0.3**	**3.7/0.3**
	MSFaiR	$123 \pm 3.1/10 \pm 0.8$	14.7/1.2	29.4/2.3	58.7/4.8

4.4.3 RQ2: Pareto Fronts

In this section, we investigate the fairness and accuracy of downstream classifiers that use representations generated by SoFaiR as inputs to predict a task label Y. IIn Figure 4.5, the larger the downstream classifier's accuracy A_y for a given value of the auditor's accuracy A_s, the better the Pareto front. First, SoFaiR and MSFaiR's Pareto fronts are either as good or better than the ones generated by $LATFR$, $CVIB$, $Maxent-ARL$ and $\beta-VAE$. Exceptions to this observations include Adults-Gender-Race for low values of A_s where $LATFR$ outperforms SoFaiR/MSFaiR. Rate distortion approaches are competitive, which confirms the tight connection between rate-distortion and unfairness-distortion as presented in Theorem 4.1.1. Both SoFaiR and MSFaiR offer more consistent performances than $LATFR$ or $Maxent-ARL$ whose representations keep leaking information related to S for Adults-Gender regardless of the constraints placed on the adversary. And, $\beta-VAE$ exhibits non-monotonic behavior for Adults-Gender. Second, Figure 4.5 shows that SoFaiR's Pareto

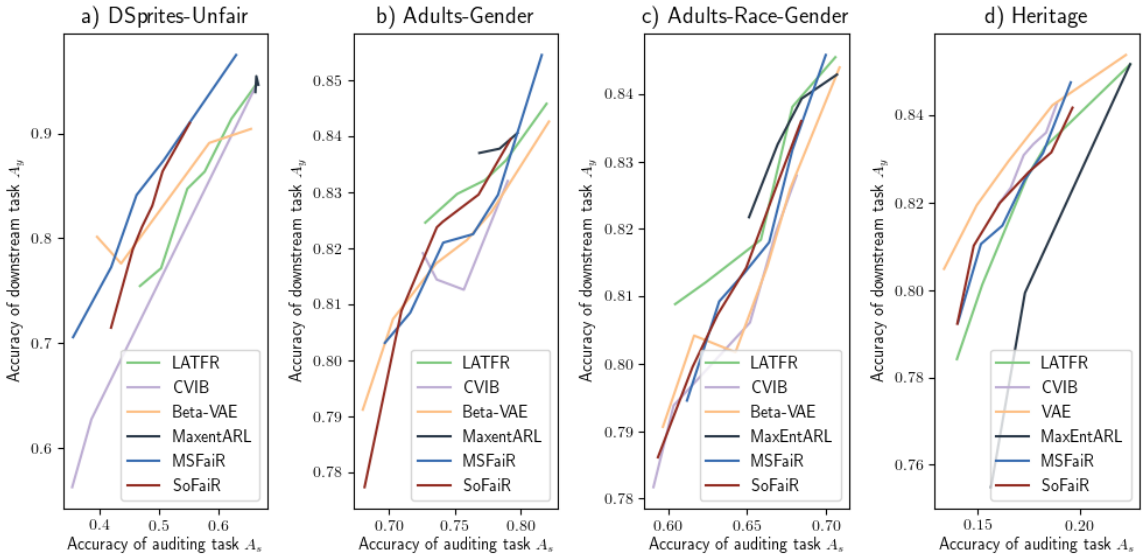


Figure 4.5: Pareto fronts for a) DSprites, b) Adults-Gender, c) Adults-Race-Gender(left) and d) Heritage. The downstream task label is whether income is larger than 50K for Adults/Adults-Race-Gender; whether a comorbidity index is positive for Heritage; which shape the image corresponds to for DSprites-Unfair.

fronts are similar to the ones offered by MSFaiR, its multi-shot counterpart. This result is consistent with AUFDC scores in Table 4.4.

4.4.4 RQ3: Interpretability

Our single-shot fair representation learning approach relies on multi-refinement of Z by adding bits to each of its dimension. This multi-refinement allows to degrade gracefully the reconstruction of the original data as fairness constraints tighten. Conversely, loosening fairness constraints is akin to send additional bits to the downstream applications that needs better reconstruction / additional information. The benefit of this multi-refinement approach is that it allows the practitioner to measure (i) how much disparity a given bit contributes to; (ii) what type of information is added to the representation as the fairness constraints loosen.

Bit Disparity. We measure the disparity of each bit b as

$$\Delta(b) = \max_{s \in \mathcal{S}} |P(b = 1|S = s) - P(b = 1|S \neq s)|. \tag{4.16}$$

Bit disparity is the demographic disparity of a classifier that returns 1 if $b = 1$ and 0 otherwise. Moreover, we show in the supplementary file that $\max_b \Delta(b)$ is a lower bound of $I(Z, S)$: a large value of $\Delta(b)$ means that the presence of bit b in the bitstream will significantly degrade the fairness properties of Z.

In Figure 4.6, loosening the fairness constraint at test time – decreasing β – unmasks more bits, while keeping the leftmost bits identical to ones obtained with higher values of β. SoFaiR degrades gracefully the fairness properties of the representation by increasing its resolution.

Figure 4.6 also shows that for Adults-Gender-Race, bits with higher disparity Δ are less likely to be unmasked with stringent fairness constraints – high β – and are only active when more leakages related to sensitive attribute are tolerated – low β. Therefore, by forcing SoFaiR to generate many points on the unfairness-distortion curve, we obtain an information ordering that pushes to the tail of the bitstreams the bits the most correlated with S. In Figure 4.7, we observe a similar pattern with Adults-Gender-Race.

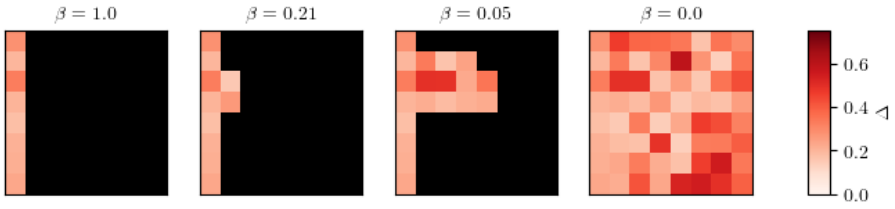


Figure 4.6: Unmasked bits for different values of the fairness coefficient β for the Adults-Gender-Race dataset. Each row is a dimension of Z. Each colored square is an unmasked bit. Black squares represent masked bits. Darker bits exhibit higher bit demographic disparity $\Delta(b)$. As β decreases, SoFaiR unmasks more bits for each dimension of Z. And, bits with higher disparity are more likely to be the last unmasked.

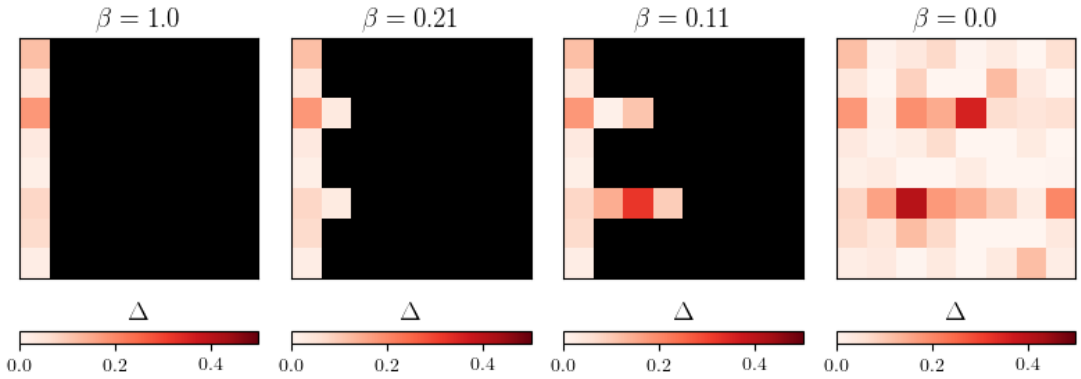


Figure 4.7: Same as Figure 4.6 but with Adults-Gender.

Fairness and information loss. Unlike alternative methods in fair representation learning, SoFaiR offers a simple tool to interpret at test time what information is lost as the fairness constraint tightens. In Figure 4.8, we plot for Adults-Gender and Adults-Gender-Race how additional bits unmasked as β decreases correlate with data features. As we move up the unfairness-distortion curve for Adults-Gender, additional information first relates to marital status; then, occupation type, relationship status and hours-per-week. It means that for downstream tasks that predict marital status, a representation on the bottom right of the unfairness-distortion curve (high distortion, low $I(Z, S)$) is sufficient to achieve good accuracy. But, downstream tasks that need hours-per-week would find more difficult to obtain good accuracy without moving up the unfairness-distortion curves, i.e leaking additional information related to sensitive attribute S.

4.5 Conclusion

In this chapter, we present SoFaiR, a single-shot fair representation learning method that allows with one trained model to explore at test time the fairness-information trade-offs of a representation of the data. Our implementation relies on a tight connection between rate-distortion and unfairness-distortion curves. SoFaiR is a step toward practical implementation of unsupervised fair representation learning approach, all the more as users can

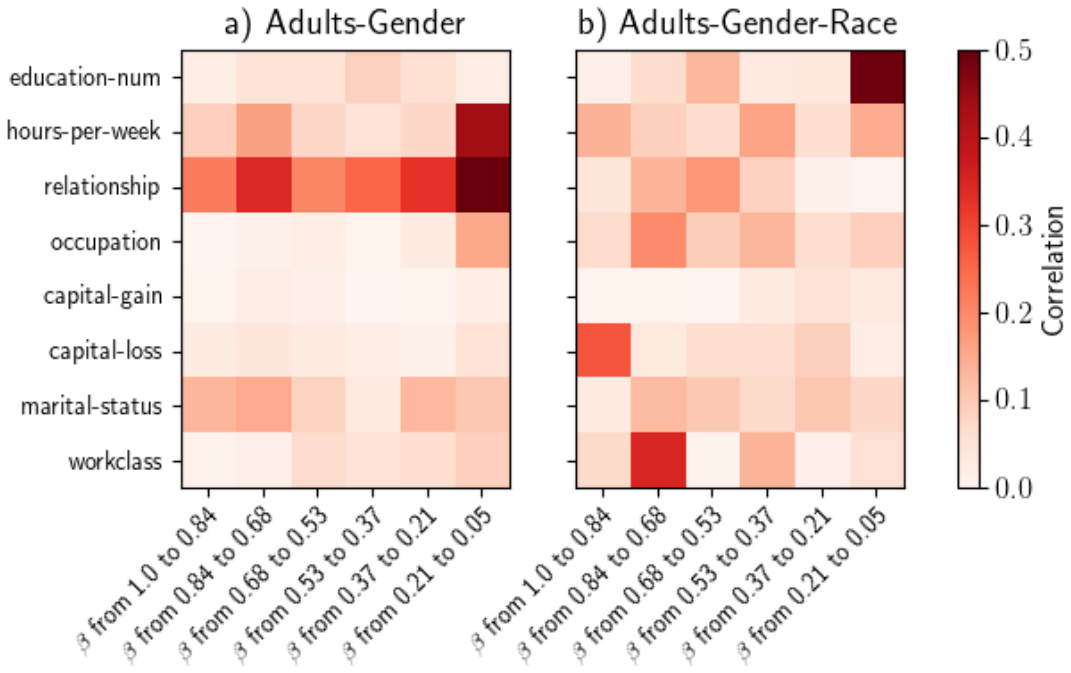


Figure 4.8: Additional information provided by refining the representation for Adults-Gender (left) and Adults-Gender-Race (right) dataset. This shows the correlation between data features and additional bits that SoFaiR unmasks when loosening the fairness constraint. Correlations are computed between the data features and the first principal component of newly unmasked bits. Each column corresponds to a decrease of β as labeled on the horizontal axis.

now explain what information is lost as the fairness properties of the representation improve.

Chapter 5: Hierarchical Fair Representation Learning

In Chapter 3 and 4, we show that in an unsupervised setting, unfairness-distortion problems can be solved as rate-distortion problems. A sufficient condition is that the decoder uses a side channel that provides with direct access to the sensitive attribute. In this chapter, we explore architectural choices to solve fair representation learning problems via rate-distortion approaches in the context of images.

A challenge to generate fair representation of images is that the encoder needs to act on high level concepts (e.g. sensitive attributes), while maintaining high resolution details to reconstruct the input image. In this chapter, we propose a hierarchical quantization approach, **HQ-FR** –**H**ierarchical **Q**uantization **F**air **R**epresentation – where low-resolution variables capture global features and condition higher resolution variables.

We verify empirically that for face images, depth, independent of model capacity, is necessary for rate-distortion based approaches to solve unsupervised fair representation learning problems. We also find that regardless of depth, alternative approaches – e.g. adversary-based methods – generate representations that still leak information related to sensitive attributes.

5.1 Method

5.1.1 Preliminary

Consider a collection of images X of dimension $C \times H \times W$, where C is the number of channels, H the height and W the width. To each image corresponds a sensitive attribute $S \in \mathcal{S} \subset \{0,1\}^{d_s}$, where $d_s \geq 1$ is the dimension of the sensitive attributes space.

The objective of unsupervised fair representation learning is to map images $X \in \mathcal{X}$ into a $d-$dimensional representation $Z \in \mathcal{Z}$ such that (i) Z maximizes the information related

to X, but (ii) minimizes the information related to sensitive attributes S. We control for the fairness properties of the representation Z via its mutual information $I(Z, S)$ with S. In an unsupervised setting, there is no pre-specified downstream task to which the representation needs to be tailored to. Therefore, we control for the information contained in Z by constraining a distortion $d(X, \{Z, S\})$ that measures how much information is lost when using a data reconstructed from Z and S instead of the original X. Fair representation learning is then equivalent to solving the following unfairness-distortion problem

$$\min_F D(X, \{Z, S\}) - \lambda I(Z, S), \tag{5.1}$$

where $F : \mathcal{X} \to \mathcal{Z}$ is an encoder; and $\lambda \geq 0$ controls the trade-off between distortion and mutual information between Z and S.

A key insight in this thesis (chapter 3 and 4) is that if we choose $D(X, \{Z, S\}) = H(X|Z, S)$, solving the unfairness-distortion trade-off (5.1) is equivalent to solving the following information bottleneck:

$$\min_F (1 + \lambda) H(X, \{Z, S\}) - \lambda I(Z, X). \tag{5.2}$$

By re-scaling the optimization (5.2) by $1/(1 + \lambda)$ and taking the limit $\lambda \to \infty$, we obtain the following fair representation learning problem:

$$\min_F H(X, \{Z, S\}) - I(Z, X). \tag{5.3}$$

The fair representation learning we propose to solve in this chapter corresponds to the fair-encoding limit whose existence we demonstrated in chapter 4: with $\lambda \to \infty$, we aim to obtain a representation for which $I(Z, S) \approx 0$, while minimizing $H(X, \{Z, S\})$. Assuming a deterministic encoder $F : \mathcal{X} \to \mathcal{Z}$, the information bottleneck in its fair-encoding limit can

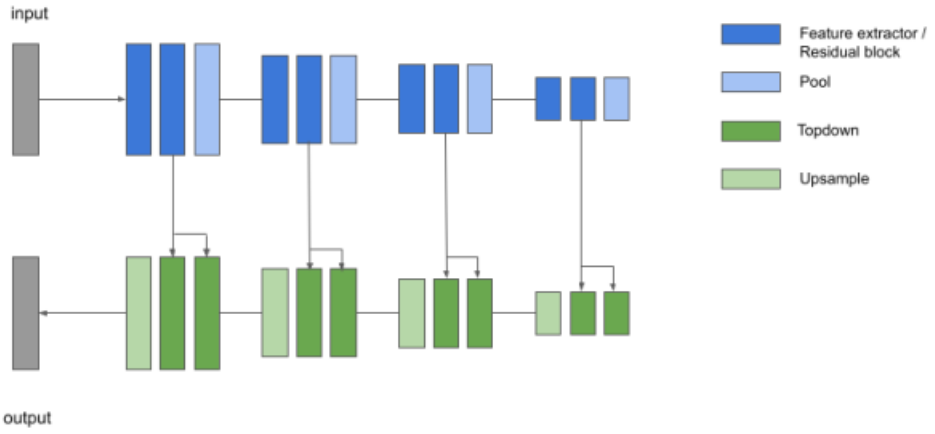


Figure 5.1: Diagram of hierarchical quantization approach. Details on topdown and residual blocks are in Figure 5.2. Pooling layers are average $2D-$pooling with a stride and a kernel of 2. Upsample uses a nearest-neighbor approach.

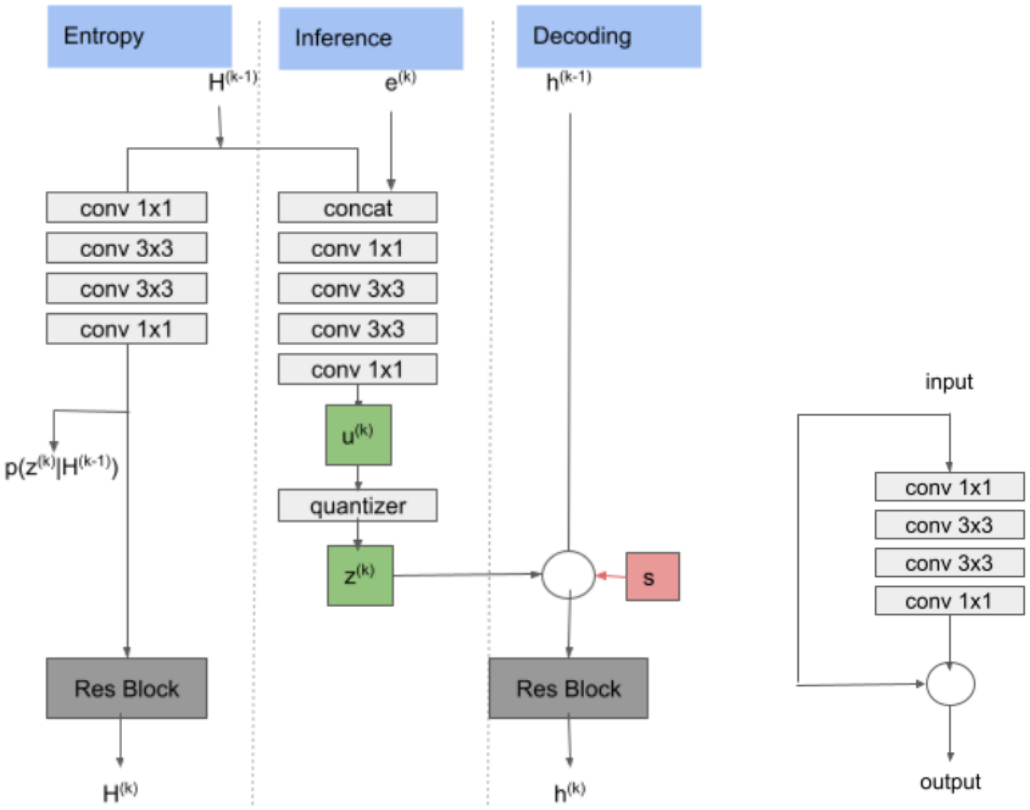
be written as

$$\min_F E_{x,s}[-\log p(x|F(x), s)] + H(F(x)) \tag{5.4}$$

5.1.2 Hierarchical Quantization

We assume that images results from a combination of high level features that capture the global characteristics of the image; and lower level features that capture local details. We also assume that sensitive attributes (e.g. gender or race) are abstract concepts that are better represented by high level variables. A simple method to model a low-to-high resolution ladder of information is a hierarchical structure of group of latent variables $z^{(0)}$, $z^{(1)}$..., $z^{(K)}$. $z^{(0)}$ captures low resolution information at the top of the hierarchy and $z^{(K)}$ high resolution details at the bottom of the hierarchy. Each group is conditionally dependent on the other ones.

We adapt the top-down structure of hierarchical VAE [1] to fair image quantization: the



(a) Topdown hierarchy

(b) Residual block

Figure 5.2: Topdown quantization. The topdown architecture (left) is similar to the one hierarchical VAE [1], but with the addition of the sensitive attribute to the decoder. Residual blocks are as in [2] with GeLU non-linearity [3].

encoder and the prior generates latent variables in parallel (see Figure 5.1):

$$p(z) = p(z^{(0)}) \prod_{k=1}^{K} p(z^{(k)}|z^{(0)}, z^{(1)}, ..., z^{(k-1)}) \tag{5.5}$$

and

$$z^{(k)} = F_k(x; z^{(0)}, z^{(1)}, ..., z^{(k-1)}). \tag{5.6}$$

A stack of feature extractors E_0, E_1, ..., E_K map x to a stack of embedding $e^{(0)}$, $e^{(1)}$, ...,

$e^{(K)}$ and the code $z^{(k)}$ is obtained as $z^{(k)} = Q \odot R_k(e_k; z^{(0)}, z^{(1)})$, where Q is a differentiable quantizer (see section 5.3.1) and R_k combines embedding from group k with quantized representation from groups higher in the hierarchy.

5.1.3 Implementation

Quantization

Using our top-down architecture, we obtain a continuous latent variable $u^{(k)} = R_k(e_k; z^{(0)}, z^{(1)})$. We use a scalar quantization as in [24,64] to model the quantizer Q. Given integer between 0 and $2^L - 1$, we quantize $u^{(k)}$ to its nearest neighbor in $\{0, ..., 2^L - 1\}$ to obtain $z^{(k)}$:

$$z^{(k)} = Q(u^{(k)}) = argmin_{i=0,...,2^L-1} ||u^{(k)} - i||. \tag{5.7}$$

Since the nearest-neighbor assignment is not differential, we use the same soft-quantization Q_{soft} as in chapter 3 and 4:

$$Q_{soft}(u^{(k)}) = \sum_{i=0}^{2^L-1} \frac{\exp(-\sigma||u^{(k)} - i||)}{\sum_{j=1}^{2^L-1} \exp(-\sigma||u^{(k)} - j||)} i, \tag{5.8}$$

where $\sigma > 0$ is a hyperparamater that controls for how hard the soft-quantization is.

Entropy Estimation

To estimate the entropy of the representation Z, we model $p(z)$ as a discrete distribution since z takes only discrete values. At each resolution k, $z^{(k)}$ is a feature map of quantized values with C_k channels. We generate a hierarchy of factors $h^{(0)}, h^{(1)}, ..., h^{(K)}$ such that

$$p(z^{(k)}|z^{(0)}, z^{(1)}, ..., z^{(k-1)}) = p(z^{(k)}|h^{(k-1)}) \tag{5.9}$$

and $h^{(k)} = G_k(h^{(k-1)}, z^k)$ (see Figure 5.2). The factor $h^{(k)}$ summarizes the information in the latent variable $z^0, ..., z^k$ and is used as input the inference leg and the entropy estimation leg of the top-down ladder in Figure 5.2 . The discrete probability distribution $p(z^{(k)}|h^{(k-1)})$ is model as a mixture of M logistic distribution [92] with mean $\mu_{k,m}$, standard deviation $\gamma_{k,m}$ and mixture weights $\pi_{k,m}$ for $k = 0, 1, ..., K$ and $m = 1, ..., M$. With this parametrization, the entropy of $p(z^{(k)}|h^{(k-1)})$ is tractable:

$$- \log p(z^{(k)}|h^{(k-1)}) = \\ - \log \left(\sum_{m=1}^{M} \pi_{k,m} \left[sigmoid \left(\frac{z^k + \mu_{k,m}}{\gamma_{k,m}} + \frac{1}{2} \right) - sigmoid \left(\frac{z^k + \mu_{k,m}}{\gamma_{k,m}} - \frac{1}{2} \right) \right] \right). \tag{5.10}$$

Means $\mu_{k,m}$, standard deviations $\gamma_{k,m}$ and mixture weights $\pi_{k,m}$ have the same number of channels as $z^{(k)}$ and obtained from $h^{(k-1)}$ via a stack of convolutional layers (see Figure 5.2).

Differences with Hierarchical VAE

Our top-down quantization has the same overall structure as state-of-the-art deep hierarchical variational autocencoder models [23] (see Figure 5.3) with three noticeable differences. First, we introduce the sensitive attribute S into the decoder as a side channel, so that it will be redundant to capture information related to the sensitive attribute captured in the quantized representation Z.

Second, since we are only interested to control for $I(Z, X)$ via the code entropy $H(z)$, while [22, 23] learn a hierarchical prior to sample new images.

Third, state-of-the-art hierarchical VAE methods [1, 22, 23] collapse the decoding factor $H^{(k)}$ and the entropy factor $h^{(k)}$. [1] show that the sharing of information and parameters between inference and generative models leads to better generative performances. In our implementation, we keep these factors separate to avoid passing information to the entropy

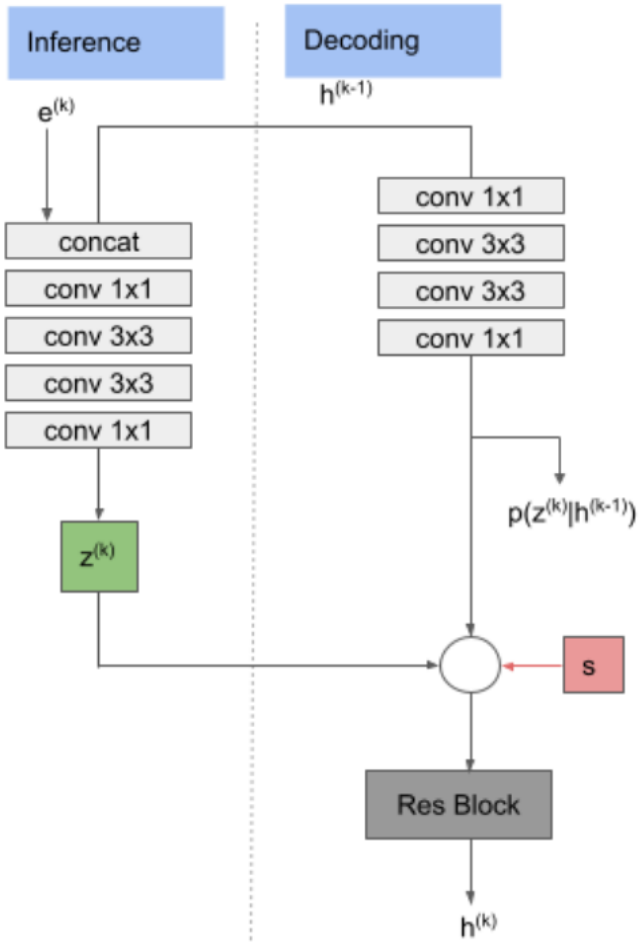


Figure 5.3: Topdown architecture in VD-VAE. Compared to Figure 5.2, VD-VAE collapses decoding and entropy legs into one leg.

factor that is related to S and is encoded in the decoder factor $H^{(k-1)}$. However, we share the information $h^{(k)}$ when predicting $H^{(k)}$. We also share the entropy factor $h^{(k)}$ between entropy and inference legs.

Differences with Multi-Resolution Deep Compression.

Recent contributions use architectures similar to the one in Figure 5.2 (e.g [24]) to solve rate-distortion problems with hierarchy of features extractors and decoders. However, these architectures have only two to three levels in the hierarchy and thus are not deep enough (see section 5) to filter out global features that correlate with the sensitive attribute S. To manage to flow information deeper in the hierarchy, we follow recent contributions in hierarchical VAE [1] and impose information sharing between feature extraction and entropy estimation legs of the encoder-decoder architecture.

Loss

With our choice of parametrization, images x are reconstructed from the sensitive attribute S and the decoding factor $H^{(K)}$ that summarizes all information in $z^{(0)}$, $z^{(1)}$, ..., $z^{(K)}$. We follow [92] and model $p(X|H^{(K)}, S)$ as a mixture of $M^{'}$ logistic distributions with means μ_m, standard deviations γ_m and mixture weights π_m. Means μ_m, standard deviations γ_m and mixture weights π_m have the same dimension as the input image with three RGB channels. We assume an auto-regression over the RGB channels [24]. That is, for an input image x the final means $\tilde{\mu}_m$ of each mixture m is given by

$$\tilde{\mu}[1]_m = \mu[1]_m,\ \tilde{\mu}[2]_m = \mu[2]_m + \nu[2]_m x[0],\ \text{and}\ \tilde{\mu}[3]_m = \mu[3]_m + \nu[3]_m x[2], \tag{5.11}$$

where [.] indexes channels; and, ν_m is a set of learnable coefficients that control the auto-regression over the RGB channels. Finally, we can write the rate-distortion problem 5.4

as

$$\min_{F,G,E} E_{x,s}\left[-\log(x|H^{(K)},s)+\sum_{k=0}^{K}-\log(p(z^{(k)}|h^{(k)}))\right]. \tag{5.12}$$

5.2 Experiments

5.2.1 Dataset

To test our hierarchical fair representation learning method, we use a standard image dataset, CelebA 64×64 and CelebA HQ.

CelebA face dataset [96] contains ten thousands of identities, each of them with twenty images. There is a total of about $200K$ images, each of them annotated with 40 labels, including information related to gender, pose or face attributes. CelebA 64×64 uses the complete set of $200K$ images cropped at the center and downsampled to a 64×64 resolution. CelebA HQ [97] is derived from CelebA by selecting the $30K$ highest quality images at 1024×1024 resolution after a series of pre-processing steps. We downsample images in CelebA HQ to a 256×256 resolution. For both CelebA 64×64 and CelebA HQ 256×256, we use gender as a sensitive attribute.

5.2.2 Evaluation of the Effect of Stochastic Depth

We want to show that depth, i.e the number of levels K in the hierarchy, improves performances in fair representation learning independently of model capacity. We consider a encoder-decoder architecture with $K = 32$ levels of latent variables. To vary the depth while keeping the model capacity unchanged, we follow the same protocol as [23]. We group the K levels into N groups where all variables within a group are emitted in parallel and are independent of each other conditionally on the variables in the previous group. Therefore, the stochastic depth of an architecture with N groups is exactly N.

5.2.3 Comparative Methods

To the extent of our knowledge, there is no method in unsupervised fair representation learning that has tackled the problem of de-biaising images. To benchmark our method, we compare it to two potential alternatives: (i) **Adversarial**, an extension of adversarial learning [16, 17]; (ii) **VD-VAE**, an extension of variational autoencoder to a hierarchy of latent variables [23].

Adversarial Learning

Adversarial fair representation learning (e.g. [16, 17]) controls for the mutual information between sensitive attribute S and representation Z via the cross-entropy of an auditor that predicts S from Z. Auditor and encoder-decoder are trained simultanously and solve a min-max optimization problem. [16,17]) tailor the learned representation to a pre-specified downstream task. Instead, we control for the information contained in the representation with $H(X|Z,S)$ as in our fair information bottleneck (5.4). To benchmark adversarial techniques to our method, we use a similar stack of feature extractors as in Figure 5.1 and encode each feature map into a representation upon which the auditor attempts to predict the sensitive attribute.

Hierarchical VAE

VD-VAE [23] is a state-of-the-art hierarchical variational autoencoder that learns very deep hierarchy of latent variables. In Chapter 3 and 4, we argue that if we recast fair representation learning as a rate-distortion, we can model the encoder as an approximate posterior and control the mutual information between X and Z via the Kullback-Leibler divergence between the approximate posterior and the prior distribution over the code Z [98, 99]. In Chapter 3 and 4, we model the approximate posterior as factorized Gaussian distributions. In this chapter, we explore whether we can turn **VD-VAE** as fair encoder-decoder by providing the sensitive attribute S as a side channel to the decoder (see Figure 5.3).

5.2.4 Fairness-Information Trade-off

To evaluate the performance of a unsupervised fair representation learning approach, we train an encoder-decoder architecture; freeze its parameter; and, compare the accuracy A_s of an auditing classifier that predicts S from Z to the accuracy A_y of a downstream classifier that predicts a task label Y from Z. The higher A_y and the lower A_s the better is the fair representation method.

Since all comparative methods we study in this chapter generate representation as a hierarchy of feature maps, we first upsample them to a common dimension $H^{'} \times W^{'}$ and stack them into a cube of dimension $C^{'} \times H^{'} \times W^{'}$, where $C^{'} = \sum_k C_k$ is the total number of channels. We then use a AlexNET type architecture [100] adapted to the dimension of the inputs.

5.3 Results and Discussion

5.3.1 Statistical Depth Improves Unfairness-Distortion Trade-off

Table 5.1: Accuracy of predicting sensitive attribute S and downstream task labels Y from representation Z with different configurations of stochastic layers on Celeba64. Convolutional networks with equal number of layers but increasing stochastic depth lead to lower distortion (↓ is better); lower accuracy of auditing networks that predict sensitive attribute from representation (↓ is better), while maintaining the same accuracy for downstream tasks (↑ is better).

Depth	Parameters	Distortion (↓)	Accuracy			
			Gender (↓)	Smile (↑)	Black Hair (↑)	Cheek bone (↑)
4	4.5M	2.32	0.93	0.90	0.91	0.86
8	4.5M	2.11	0.92	0.91	0.90	0.85
16	4.5M	1.94	0.77	0.91	0.90	0.87
32	4.5M	1.91	0.57	0.89	0.92	0.85

We test on CelebaA 64×64 whether stochastic depth improves the fairness-information properties of the learned representation. In Table 5.1, we train networks with hierarchical

quantization, similar capacity but various stochastic depth (4, 8, 16, 32). First, stochastic depth improves the model performance in terms of distortion, which is consistent with results in [23]. Second, only deeper models ($K = 16, 32$) reduce the ability of the auditing classifier to predict the sensitive attribute with high accuracy. At $K = 32$, the auditing classifier achieves an accuracy of 0.57, which is equal to the accuracy of a classifier that randomly predicts the gender associated with each image. Third, regardless of stochastic depth, all methods achieve similar performance with respect to downstream tasks. Therefore, results in Table 5.1 are evidence that on one hand, with enough depth, a compression approach can filter out sensitive attributes. On the other hand, compression models that are too shallow to separate global concepts from local details fail to generate fair representations of face images.

5.3.2 Comparative Methods

Table 5.2: Same as in Table 5.1 but with differents method in fair representation learning.

Method	Distortion (↓)	Accuracy			
		Gender (↓)	Smile (↑)	Black Hair (↑)	Cheek bone (↑)
Adversarial	0.9	0.98	0.90	0.92	0.88
VD-VAE[23]	1.88	0.92	0.91	0.90	0.85
HQ-FR (Ours)	1.91	0.57	0.89	0.92	0.85

We test whether we can obtain fair representations of images with adversarial or state-of-the-art hierarchical variational autoencoder. In Table 5.2, we find that **Adversarial** achieves lower distortion than **HQ-FR**, but does not provide any fairness guarantees. The solution of the min-max optimization is likely to converge to an encoder that fools the auditing adversary during training, but this hiding of sensitive attribute does not generalize to new auditors during a post-mortem evaluation of the fairness properties of the representation. This observation is consistent with findings in [13, 20] and will be expanded upon in Chapter 6.

Moreover, among rate-distortion approach to fair representation learning, we find that a state-of-the-art variational autoencoder **VD-VAE** does not filter out the sensitive attribute. By design, we pass the sensitive attribute S to each level of the decoding leg and this information is shared with the inference leg at lower level of the hierarchy, which leaks information about S into the latent variables.

5.4 Conclusion

This chapter demonstrates that with a carefully designed hierarchy of latent variables, rate-distortion approaches to unsupervised fair representation learning can filter out sensitive attributes while maintaining useful information related to the data. We argue that deeper hierarchies should perform better for image data, propose a deep hierarchical quantization approach and show that its fairness-information trade-off outperforms adversarial techniques.

We posit that in fair representation learning of images, depth improves unfairness-distortion trade-off because (i) sensitive attributes like gender are likely to be abstract concepts; and, (ii) latent variables higher in the hierarchy are more likely to encode abstract concepts. We hope that our empirical results foster future research on how to measure at which level of abstraction sensitive attributes sit in a hierarchy of quantized latent variables and how to use this information to design architectures in fair representation learning.

Chapter 6: Learning Smooth and Fair Representations

This chapter explores conditions on auto-encoders in unsupervised fair representation learning so that the encoder generates distributions over the representation space with fairness guarantees that hold for any downstream task.

Our first result is to establish a necessary and sufficient condition – finite χ^2 mutual information between data and representation – for guarantees estimated from a finite sample to generalize to all downstream tasks and to the infinite sample regime.

We ask then how to guarantee that the χ^2 mutual information between data and representation is finite while not knowing ex-ante the distribution over the feature space. We show that introducing an additive Gaussian white noise channel – **AGWN** – in standard fair representation methods bounds the χ^2 mutual information once the representation has passed through the channel, regardless of the distribution of features.

We empirically find across four datasets that an AGWN channel in fair representation learning guarantees that empirical fairness certificates estimated on finite samples upper bound the demographic disparity of multiple and diverse downstream users. This is an improvement over existing methods in fair representation learning for which we find that fairness guarantees do not extend beyond a set of specific downstream users.

The work presented in this chapter have been published in [13].

6.1 Certifying Fair Representations

6.1.1 Background

Consider a data controller who wants to release samples from a distribution μ over $\mathcal{X}\times\mathcal{S}$ with features in $\mathcal{X} \subset [0,1]^D$ and sensitive attributes in $\mathcal{S}$. Although our setup can be extended

to richer spaces of sensitive attributes, we focus here on binary sensitive attributes and assume that $\mathcal{S} = \{0, 1\}$.

A transformation F that maps the features space $\mathcal{X}$ into a representations space $\mathcal{Z} \subset \mathbb{R}^d$ induces a distribution μ_F over $\mathcal{Z} \times \{0, 1\}$: $\mu_F(A) = \mu(\{x \in \mathcal{X} | F(x) \in A\})$ for any $A \subset \mathcal{Z}$.

The data controller's objective is to obtain a representation mapping F that minimizes the statistical dependence between representation Z and sensitive attribute S. Therefore, for any test $f : \mathcal{Z} \to \{0, 1\}$ that decides whether the class conditional distributions $\mu_F^0 = P(Z|S = 0)$ and $\mu_F^1 = P(Z|S = 1)$ are identical, the data controller would like to minimize the discrepancy

$$\Delta(f, F) \triangleq |E_{x \sim \mu_F^1}[f(x)] - E_{x \sim \mu_F^0}[f(x)]|, \tag{6.1}$$

where we make the dependence of Δ on representation mapping F explicit. In the context of fair machine learning, the test function f is either an auditor used by the data controller to estimate the statistical dependence between Z and S; or, a classifier used by a data processor (function h in Figure 1.1) and $\Delta(f, F)$) then measures the demographic parity of f (see [38]):

Definition 6.1.1. Demographic parity *Consider a representation distribution μ_F induced by a representation mapping $F : \mathcal{X} \to \mathcal{Z}$. A classifier $f : \mathcal{Z} \to \{0, 1\}$ used by a data processor satisfies $\delta-$ Demographic Parity on μ_F if and only if $\Delta(f, F)) \leq \delta$.*

Since the data controller does not know ex-ante which classifier data processors will use, she has to construct a mapping F such that all classifiers $f : \mathcal{Z} \to \{0, 1\}$ satisfy $\delta-$ demographic parity on μ_F for some pre-specified $\delta > 0$. A demographic parity certificate is therefore an upper bound on the demographic disparity of any classifiers that access samples from the representation distribution μ_F.

Definition 6.1.2. Demographic Parity Certificate *Let $\delta \geq 0$. A representation space*

$(\mathcal{Z}, \mu_F)$ can be certified with $\delta-$ demographic parity if and only if

$$\Delta^*(F) \triangleq \sup_{f:\mathcal{Z}\to\{0,1\}} \Delta(f, F)) = \delta. \tag{6.2}$$

To construct a representation mapping certified with $\Delta^*(F)-$ demographic parity, the data controller needs to evaluate the supremum over all test functions/auditors f_n that are constructed on the basis of a finite sample $\mathcal{D}_n = \{(x_i, s_i)\}_{i=1}^n$. Let $\mathcal{F}_n$ denote the set of all auditors $f_n : \mathcal{Z} \times (\mathcal{Z} \times \{0,1\})^n \to \{0,1\}$ constructed from a sample of size n.

Definition 6.1.3. Empirical Demographic Parity Certificate *Let $n \geq 1$ and $\delta \geq 0$. A representation space $(\mathcal{Z}, \mu_F)$ is certified with an empirical $\delta-$ demographic parity certificate if and only if*

$$\Delta_n(F) \triangleq \sup_{f_n:\in\mathcal{F}_n} \Delta(f_n, F) = \delta. \tag{6.3}$$

The question is how to choose a representation mapping $F : \mathcal{X} \to Z$ so that empirical certificates $\Delta_n(F)$ approximate well the true demographic parity certificate $\Delta^*(F)$. Approximation properties of empirical certificates are important for a data controller to upper bound the demographic disparity of any downstream processor that uses fresh samples obtained after F has been constructed.

Since the data controller cannot constrain the data distribution over $\mathcal{X} \times \{0,1\}$, we are looking for distribution-free approximation rates. In general, distribution-free rates do not exist ([101], ch. 7). But, in our setting, the data controller has some control over the representation distribution via F. In fact, the approximation $\Delta^*(F) - \Delta_n(F)$ depends on how much information in X is encoded by F in Z. If F randomly maps $\mathcal{X}$ to $\mathcal{Z}$, the data controller can certify μ_F with $0-$ demographic parity, but μ_F is useless to downstream data processors. The data controller trades-off representation demographic parity with information by learning an encoder $F : \mathcal{X} \to \mathcal{Z}$ and a decoder function $G : \mathcal{Z} \to \mathcal{X}$ that

solves the following fair empirical representation problem

$$\min_{F,G} \mathcal{L}_{rec}(g, t, \mathcal{D}_n) \text{ subject to } \Delta_n(F) \leq \delta, \tag{6.4}$$

where $\delta > 0$ is a pre-specified demographic parity threshold and $\mathcal{L}_{rec}$ is a reconstruction loss.

6.1.2 Necessary Condition

This section identifies a necessary condition on F for an empirical demographic parity certificate to approximate $\Delta^*(F)$ well. The necessary condition bounds the amount of information measured by the χ^2 mutual information between feature X and representation Z:

$$I_{\chi^2}(X, Z) \triangleq E_x E_z \left(\frac{\mu_F(z) - \mu_F(Z|X = x)}{\mu_F(z)} \right)^2. \tag{6.5}$$

The χ^2 mutual information relies on a statistical distance, the χ^2-divergence – $\chi^2(Z, Z|X) = \int_z \left(dP(Z|X)/dP(Z) - 1\right)^2 dP(Z)$ – to average the distance between Z and $Z|X = x$ for $x \in \mathcal{X}$. It has been used in information theory to estimate the information that flows through a neural network (see [26]). In the context of fair representation learning, we find that empirical demographic parity certificates cannot provide good approximations of the representation's true demographic parity if the χ^2 input-output mutual information is large:

Theorem 6.1.1. *Let $n \geq 1$. Consider a representation function $F : \mathcal{X} \to \mathcal{Z}$. Then, for all test function $f_n \in \mathcal{F}_n$*

$$\sup_{\mu} E_{\mathcal{D}_n} |\Delta^*(F) - \Delta(f_n, F)| \geq \sup_{\mu_x} \left(1 - \frac{1}{I_{\chi^2}(X, Z)} \right)^n, \tag{6.6}$$

where the suppremum on the left hand side is taken over all distributions μ over $\mathcal{X} \times \mathcal{S}$ and the suppremum on the right hand side is taken over all distribution μ_x over $\mathcal{X}$.

Encoding more information of X in Z exposes the representation distribution μ_F to mirroring distributions over $\mathcal{X}$ with heavy tails. Intuitively, μ_F is a (possibly infinite) mixture of conditional distributions $P(Z|X=x)$ for $x \in \mathcal{X}$ and $I_{\chi^2}(X,Z)$ measures an average distance between those conditional distributions. As $I_{\chi^2}(X,Z)$ increases, the conditional distributions $P(Z|X=x)$ become far apart for a growing mass of $x \in \mathcal{X}$. It generates a representation distribution too complex for a finite sample to represent it and for an auditor f_n to detect all the correlations between representation and sensitive attribute.

Theorem 6.1.1 implies a trade-off between the information passed from features to representations and the approximation rate of empirical demographic parity certificates:

Corollary 6.1.1. *With the notations from Theorem 6.1.1,*

- *If* $\inf_{f_n \in \mathcal{F}_n} \sup_{\mu} E_{\mathcal{D}_n} |\Delta^*(F) - \Delta(f_n, F)| \leq \epsilon_n$*, then for all distributions over the feature space* $\mathcal{X}$*,* $I_{\chi^2}(X,Z) \leq \frac{1}{1-\epsilon_n^{\frac{1}{n}}}$.

- *If there exists a distribution over* $\mathcal{X}$ *such that* $I_{\chi^2}(X,Z) = \infty$*,*

$$\inf_{f_n \in \mathcal{F}_n} \sup_{\mu} \Delta^*(F) - \Delta(f_n, F) \geq 1. \tag{6.7}$$

For the approximation rate of $\Delta^*(F) - \Delta(f_n, F)$ to be $O(n^{-s})$ for some $s > 0$, it is necessary for the χ^2 mutual information between feature and representation to be bounded above by $O(n/(s\ln(n))$ for *all* distributions over $\mathcal{X}$. On the other hand, representation functions F for which the χ^2 mutual information is infinite for some distribution over the features space, never guarantee a meaningful approximate rate between $\Delta^*(F)$ and $\Delta(f_n, F)$.

Examples: The results in corollary 6.1.1 imply that empirical certificates of representation distributions induced by many common encoders do not have meaningful approximation rates:

- Suppose that F is injective from $\mathbb{R}^D$ to $\mathbb{R}^d$. Then, there exists a distribution over $\mathcal{X} \times \{0,1\}$ such that $I_{\chi^2}(X,Z) = \infty$ and thus, $\Delta^*(F) = 1$, but $\Delta(f_n, F) = 0$ for all auditing functions f_n.

- Suppose that $|\{F(x)|x \in \mathcal{X}\}| \geq n/(\ln(n))^\alpha$, for some $\alpha < 1$. Then, the approximation rate of $\Delta(f_n, F)$ for all auditing functions f_n is $\omega(n^{-s})$ for any $s > 0$.

6.1.3 Sufficient Condition

This section shows that a finite χ^2 mutual information between feature and representation for all distributions over $\mathcal{X}$ is a sufficient condition for empirical demographic parity certificates to converge at a $O(n^{-1/2})$ rate.

Theorem 6.1.2. *Let $n \geq 1$. Consider a representation mapping $F : \mathcal{X} \to \mathcal{Z}$. Then, for all distribution μ over $\mathcal{X} \times \{0,1\}$ with where $n_s = |\{i|s_i = s\}|$ and for all $f_n \in \mathcal{F}_n$*

$$E_{\mathcal{D}_n}|\Delta^*(F) - \Delta(f_n, F)| \leq 2 \sum_{s=0,1} \sqrt{\frac{I_{\chi^2}(X, Z|S=s)}{n_s}}$$

A finite χ^2 mutual information between X and Z implies that $p(Z)$ and $p(Z|X)$ are close in the sense of the χ^2 divergence and thus by sampling representations from $P(Z|X)$, we have a non-zero probability to sample all the atoms that can form the representation distribution μ_F and thus to detect all the dependence between representations and sensitive attributes.

6.1.4 Chi - versus Classic Mutual Information

Our results in Theorems 6.1.1 and 6.1.2 highlight a connection between χ^2 mutual information and approximation rate of empirical certificates. A similar result cannot be obtained with the classic mutual information $I_{Sh}(X,Z)$ that is based on Shannon entropy.

To demonstrate this point, we construct the following distribution μ over $\mathcal{X} \times \{0,1\}$.

Features are uniformly distributed over $[0,1]$ and $F(x) = i$ for $x \in [1/i, 1/(i+1))$ and $i > 0$. For each $i > 0$, the sensitive attribute is constant over $[1/i, 1/(i+1))$ and equal to 1 with probability $1/2$. We show in the appendix that $I_{Sh}(X,Z) < \ln(2)/2+2$, but $I_{\chi^2}(X,Z) = \infty$. Since the sensitive attribute S is a deterministic function of the representation $Z = F(X)$, $\Delta^*(F) = 1$. But, for a finite sample of size n, $E_{\mathcal{D}_n}\Delta(f_n, F)$ is zero for all auditors f_n, despite $I_{Sh}(X,Z) < \infty$.

6.2 Smooth and Fair Representations

The previous section suggests restricting the fair representation problem (6.4) to encoder F for which the χ^2-mutual information between feature and representation is finite for all distributions over $\mathcal{X}$. Here, we meet this condition by adding an additive Gaussian white noise (AGWN) channel to the encoder. For any representation mapping $F : \mathcal{X} \to \mathcal{Z}$, we denote F_σ the convolution of F with a Gaussian noise $\mathcal{N}(0, \sigma^2 I_d)$: $F_\sigma(X) = F(X) + noise$, with $noise \sim \mathcal{N}(0, \sigma^2 I_d)$.

6.2.1 Convergence of Smoothed Empirical Certificate

The convolved representation $Z_\sigma = Z + noise$ generated by F_σ has a distribution denoted $\mu_{t*\sigma}$. The convolution smoothes the representation distribution by making $P(Z_\sigma|X)$ a Gaussian whose support covers the support of the representation distribution $P(Z_\sigma)$ and thus, guarantees that samples from different conditional distributions $P(Z_\sigma|X = x)$ are not too far away.

Theorem 6.2.1. *Let $\sigma > 0$ and $n \geq 1$. For all representation mapping $F : \mathcal{X} \to \mathcal{Z}$ and for any distribution over $\mathcal{X}$, if $||F||_\infty \triangleq \sup_{x\in\mathcal{X}} ||F(x)||_2$, then for $s \in \{0,1\}$*

$$I_{\chi^2}(X, Z|S = s) \leq \exp\left(\frac{||F||_\infty^2}{\sigma^2}\right) < \infty. \tag{6.8}$$

Therefore,

$$\begin{aligned} &\inf_{f_n \in \mathcal{F}_n} \sup_{\mu} E_{\mathcal{D}_n}[\Delta^*(F_\sigma) - \Delta(F_\sigma, f_n)] \\ &\leq 2 \exp\left(\frac{||F||_\infty^2}{2\sigma^2}\right)(n_0^{-1/2} + n_1^{-1/2}). \end{aligned} \tag{6.9}$$

The upper bound in Theorem 6.2.1 does not depend on the dimensions d of the representation space $\mathcal{Z}$, but only on $n^{-1/2}$ and on the ratio $||F||_\infty/\sigma$ that can be interpreted as a signal-to-noise ratio in the AGWN channel. Larger values of $||F||_\infty$ increase the variance of Z and thus require larger noise σ to keep the conditional distribution $P(Z_\sigma|X)$ close to the distribution $P(Z_\sigma)$. The bound is only meaningful if $||F||_\infty < \infty$, which holds, for example, if the features space is bounded and F is a continuous mapping.

Both Theorems 6.1.2 and 6.2.1 rely on a plug-in auditor that first estimates the class-conditional densities $\mu^0_{t*\sigma}$ and $\mu^1_{t*\sigma}$. From a sample $\mathcal{D}_n = \{(x_i, s_i)\}_{i=1}^n$, we construct an empirical estimate of $\mu_{t*\sigma}$ over $\mathcal{Z} \times \{0, 1\}$ as

$$\mu_{n,\sigma}(z, s) = \frac{1}{n} \sum_{i=1, s_i=s}^{n} P(z|X = x_i) \tag{6.10}$$

with $P(.|X = x_i) \sim \mathcal{N}(F_n(x_i), \sigma I_d)$. Our plug-in auditor f_n^{plug} compares $\mu_{n,\sigma}(z, 0)$ to $\mu_{n,\sigma}(z, 1)$:

$$f_n^{plug}(z) = \begin{cases} 0 & \text{if } \mu_{n,\sigma}(z, 0) \geq \mu_{n,\sigma}(z, 1) \\ 1 & \text{otherwise.} \end{cases} \tag{6.11}$$

Since we obtain the upper bounds in Theorems 6.1.2 and 6.2.1 with the plug-in auditor f_n^{plug}, we can guarantee that the representation demographic parity is within $O(n^{-1/2})$ of the empirical certificate signed by f_n^{plug}.

6.2.2 Learning Fair Representation

In practice, the representation mapping F and the decoder G are modelled by neural networks. An AGWN channel is added to F to learn a smoothed representation distribution $\mu_{t*\sigma}$. The data controller trades off minimizing a reconstruction loss $\mathcal{L}_{rec}(F,G) = E_x[l_{rec}(F,g,x)$ with minimizing demographic disparity $\mathcal{L}_{DP}(F) = \Delta^*(F_\sigma)$. With a sample $\mathcal{D}_n = \{(x_i,s_i)\}_{i=1}^n$, the data controller uses the plug-in auditor and solves the empirical minimization problem as

$$\min_{F,G} \frac{1}{n} \sum l_{rec}(F,G,x_i) + \lambda \Delta(f_n^{plug}, F_\sigma), \tag{6.12}$$

where λ controls for the strength of the fairness constraint imposed on the representation distribution. The minimization problem in (6.12) differs from previous work on fair representation learning because of the noise added to Z and thus, provides theoretical guarantees that $\Delta(f_n^{plug}, F_\sigma,)$ approximates $\Delta^*(F_\sigma)$ at a rate $O(n^{-1/2})$.

Moreover, the empirical demographic parity certificate can be computed without modelling the auditor by an additional adversarial neural network. This is because we can use our empirical estimates (6.10) of the class-conditional densities to estimate the posterior distribution $\eta(z,s) = P(S=s|Z=z)$ as $\eta_n(z,s) = \mu_{n,\sigma}(z|S=s)/\mu_{n,\sigma}(z)$, where $\mu_{n,\sigma}(z) = \mu_{n,\sigma}(z,1) + \mu_{n,\sigma}(z,0)$. Since $\Delta^*(F)$ relates to the balanced error rate of predicting the sensitive attributes (see proof of 6.1.2 or [36]), we can write $\Delta^*(F) = \mathcal{L}_{DP}(\mu_{t,\sigma})$, where $\mathcal{L}_{DP}(\mu_{t,\sigma}) = E_{z\sim\mu_{t,\sigma}}[|\eta(z,1) - \eta(z,0)|]$ (see [102]). Our approach relies on two results: (i) for any finite sample of size n, $\mathcal{L}_{DP}(\mu_{n,\sigma})$ approximates well $\mathcal{L}_{DP}(\mu_{t*\sigma})$; (ii) $\mathcal{L}_{DP}(\mu_{n,\sigma})$ can be estimated efficiently by Monte-Carlo estimation. The first observation uses the following result, which is a consequence of Theorem 6.2.1

Theorem 6.2.2. *Let $\sigma > 0$ and $n \geq 1$. For all representation mapping $F : \mathcal{X} \to \mathcal{Z}$*

$$\begin{aligned} &\sup_{\mu} E_{\mathcal{D}_n} |\mathcal{L}_{DP}(\mu_{t*\sigma}) - \mathcal{L}_{DP}(\mu_{n,\sigma})| \\ &\leq 2 \exp\left(\frac{||F||_\infty^2}{2\sigma^2}\right) (n_0^{-1/2} + n_1^{-1/2}). \end{aligned} \tag{6.13}$$

Therefore, we can use $\mathcal{L}_{DP}(\mu_{n,\sigma})$ as an approximation of $\mathcal{L}_{DP}(\mu_{t*\sigma})$. That is, in place of $\mu_{t,\sigma}$, we propose to use the distribution $\mu_{n,\sigma}$, for which η_n is the posteriori probability. Moreover, $\mathcal{L}_{DP}(\mu_{n,\sigma})$ can be efficiently approximated by Monte Carlo integration. For a sample $\mathcal{D}_n = \{(x_i, y_i)\}_{i=1}^n$, $\mu_{n,\sigma}^0$ and $\mu_{n,\sigma}^1$ are mixtures of d-dimensional Gaussians. Thereby, we approximate $\mathcal{L}_{DP}(\mu_{n,\sigma})$ with

$$\hat{\mathcal{L}}_{DP}(\mu_{n,\sigma}) = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} E_\epsilon[|\eta_n(z_{ij}, 1) - \eta_n(z_{ij}, 0)|, \tag{6.14}$$

where $z_{ij} = F(x_i) + noise_{ij}$, $\{noise_{ji}\}$ is a vector of $n \times m$ draws from a d-dimensional Gaussian $\mathcal{N}(0, \sigma I_d)$ and m is the number of draws per sample point. $\hat{\mathcal{L}}_{DP}(\mu_{n,\sigma})$ is an unbiased approximation of $\mathcal{L}_{DP}(\mu_{n,\sigma})$ and achieves a Mean-Squared-Error (MSE) of order $O(n^{-1}m^{-1})$ (see proof of Theorem 4 in appendix).

To sum up, the data controller learns (F, G) by minimizing the following combined empirical loss

$$\min_{\theta,\varphi} \frac{1}{n} \sum_i l_{rec}(F, G, x_i) + \lambda \hat{\mathcal{L}}_{DP}(\mu_{n,\sigma}). \tag{6.15}$$

Practical implementation. We minimize the loss (6.15) by stochastic gradient descent. Each mini-batch is split in half: the first half is used to estimate $\mu_{n,\sigma}$ as in (6.10); the second half to estimate the loss in (6.15). At the end of training, we compute a leave-one-out balanced error rate $BER(f_n^{plug})$ for the plug-in auditor on both a test and train

samples and infer an empirical certificate as $\Delta(f_n^{plug}, F) = 1 - 2BER(f_n^{plug})$ (see [36]). The Gaussian noise σ is an hyper-parameter chosen so that empirical certificates estimated on train and test data are similar.

6.3 Experiments

The objective of this experimental section is to demonstrate that (i) our AGWN fair representation method, unlike competitive approaches, generates robust fairness certificates that generalize to unseen data; and, (ii) it is competitive with existing fair representation methods in terms of fairness-accuracy trade-off. All details related to dataset and neural network architectures are in the appendix.

6.3.1 Datasets

6.3.2 Synthetic Datasets

We first consider two synthetic datasets. Our first synthetic data consists of two $3D$ Swiss rolls: one for $S = 0$ and one shifted South-West for $S = 1$. We use $20,000$ samples for training the autoencoder (F, G) and $10,000$ fresh samples to train the downstream processors and $10,000$ to evaluate their demogrpahic disparity. Our second synthetic data is a variant of the DSprites dataset ([89]) that contains 64 by 64 black and white images of various shapes (heart, square, circle). The DSprites dataset has six independent factors of variation: color (black or white); shape (square, heart, ellipse), scales (6 values), orientation (40 angles in $[0, 2\pi]$); x- and y- positions (32 values each). We adapt the sampling to generate a source of potential unfairness as in [21]. We consider shape as the sensitive attribute. We assign to each possible combination of attributes a weight proportional to $\frac{i_{shape}}{3} + \left(\frac{i_X}{32}\right)^3$, where $i_{shape} \in \{0, 1, 2\}$ and $i_X = \{0, 1, ..., 21\}$. we sample $600,000$ combinations of latent factors to train the encoder-decoder; $20,000$ to train the downstream processors; and, $20,000$ to evaluate the disparity of the downstream processors.

Fairness Benchmark Dataset

We also apply our approach of fair representation learning with a AGWN channel to two fair learning benchmarks, Adults[1] and Heritage[2]. The Adults dataset contains $49K$ individuals and includes information on 10 features related to professional occupation, education attainment, race, capital gains, hours worked and marital status. The sensitive attribute is the gender to which individuals self-identify to. The data is split into a $32K$ set to train the auto-encoder; a $13K$ set to train the downstream processors; and, a $3K$ test set to evaluate the disparity of the processors.

The Health Heritage dataset contains $220K$ individuals with 66 features related to age, clinical diagnoses and procedure, lab results, drug prescriptions and claims payment aggregated over 3 years. The sensitive attribute is the gender to which individuals self-identify to. After removing individuals with missing records, we split the data into a $142K$ set to train the auto-encoder; a $17K$ set to train the downstream processors; and, a $17K$ test set to evaluate the disparity of the processors.

6.3.3 Effect of noise on certificate reliability.

We first learn an encoder-decoder mapping (F, G) with an increasing amount of Gaussian noise; estimate an empirical $\Delta(f_n^{plug}, F)-$demographic parity certificate; and then, test whether $\Delta(f_n^{plug}, F)$ is larger than the demographic disparity $\Delta(f_{proc}, F)$ of different downstream processors f_{proc} that predict sensitive attributes from new samples of the representation distribution. Empirical certificates are robust if $\Delta(f_n^{plug}, F) \geq \Delta(f_{proc}, F)$ for any of the processors f_{proc}.

All datasets are split into a train set for training the auto-encoder (F, G); two test sets to first train downstream processors and then evaluate their accuracy.

[1] https://archive.ics.uci.edu/ml/datasets/adult
[2] https://foreverdata.org/1015/index.html

6.3.4 Architectures

Encoder-Decoder

For the DSprites dataset, the autoencoder architecture – taken directly from [21] – includes 4 convolutional layers and 4 deconvolutional layers and uses ReLU activations. For the Swiss Roll dataset and the two real world datasets, the encoder and decoder are made of fully connected layers with ReLU activations. Table 6.1 shows more architectural details for each dataset. Hyperparameter values are in Table 6.2.

Table 6.1: Architecture details. $Conv2d(i, o, k, s)$ represents a 2D-convolutional layer with input channels i, output channels o, kernel size k and stride s. $ConvT2d(i, o, k, s)$ represents a 2D-deconvolutional layer with input channels i, output channels o, kernel size k and stride s. $Linear(i, o)$ represents a fully connected layer with input dimension i and output dimension o. The *tanh* activation is only applied to the last layer of the encoder.

Dataset	Encoder	Decoder	Activation
Swiss Roll	Linear(3, 64) Linear(64, 64), Linear(64, 64)	Linear(3, 64), Linear(64, 3)	ReLU
DSprites	Conv(1, 32, 4, 2), Conv(32, 32, 4, 2), Conv(32, 64, 4, 2), Conv(64, 64, 4, 2), Linear(1024, 128)	Linear(28, 128), Linear(128, 1024) ConvT2d(64, 64, 4, 2), ConvT2d(64, 32, 4, 2) ConvT2d(32, 32, 4, 2) ConvT2d(32, 61, 4, 2)	ReLU
Adults	Linear(10, 64) Linear(64, 10)	Linear(64, 10) Linear(64, 10)	ReLU ReLU
Heritage	Linear(66, 128) Linear(128, 24)	Linear(24, 128) Linear(128, 66)	ReLU

Downstream Processors

The downstream test functions that probe the demographic parity of the representation distribution are fully connected neural networks with 2 to 4 hidden layers with 32 to 128 neurons each. Each test function is trained for 400 epochs with a learning rate of 0.001. After

Table 6.2: Hyperparameter values for training encoder-decoder networks.

Dataset	Number of iterations	Learning rate	σ	λ_{max}		
				AGWN	AdvCE	AdvL1
Swiss Roll	4K	10^{-3}	0.05	10	4	4
DSprites	270K	10^{-4}	0.05	0.025	0.035	0.035
Adults	55K	10^{-3}	0.02	2.6	2.8	2.8
Heritage	55K	0.5×10^{-4}	0.05	2.6	2.6	2.6

the autoencoder is trained, its weights are frozen, and fresh representations are generated by $10,000$ forward passes of the encoder on the test data. The generated fresh representations form the inputs of the test functions.

6.3.5 Comparative Methods

We benchmark the use of an AGWN channel with comparative approaches in fair representation learning based on adversarial auditor trained with (i) a cross-entropy loss (AdvCE, [17]); or, with (ii) a group L1 loss (AdvL1, [16]).

AdvCE.

AdvCE is a fair representation learning method from [17]. The auditor is modeled as an adversarial neural network a that predicts sensitive attributes from samples of the representation distribution and minimizes the following cross-entropy loss:

$$\mathcal{L}_{CE}(a) = -\frac{1}{n}\sum_{i=1}^{n} s_i \log(a(x_i) + (1 - s_i)\log(1 - a(x_i)). \tag{6.16}$$

Moreover, the autoencoder is trained to minimize a loss $\mathcal{L}_{rec} - \lambda\mathcal{L}_{CE}(a)$.

AdvL1

AdvL1 ([16]) replaces the cross-entropy loss by a group L1 loss: instead of (6.16), the adversary minimizes

$$\mathcal{L}_{L1} = \frac{1}{n_0} \sum_{i,s_i=0} a(x_i) - \frac{1}{n_1} \sum_{i,s_i=1} a(x_i), \tag{6.17}$$

and the autoencoder minimizes $\mathcal{L}_{rec} - \lambda \mathcal{L}_{L1}(a)$.

For both AdvCE and AdvL1, the adversarial auditor is modeled as a neural network with 3 hidden layers of 64 neurons each for Adults and Swiss Roll; 3 hidden layers of 128 neurons each for Heritage; and, 3 hidden layers of 256 neurons each for DSprites.

6.4 Results and Discussion

6.4.1 Certificate reliability.

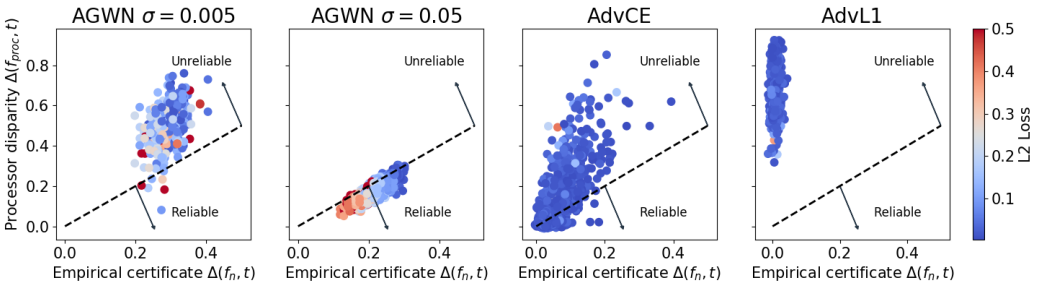


Figure 6.1: Generalization of empirical demographic parity certificates for the Swiss Roll data. Each dot shows empirical demographic parity certificate $\Delta(f_n, F)$ for an encoder $F \in \{AGWN, AdvCE, AdvL1\}$ against an estimate of the disparity $\Delta(f_{proc}, F)$ of downstream processors predicting sensitive attributes. Dots are colored by reconstruction loss.

Figure 6.1 and Figure 6.2 show that the AGWN channel improves how empirical certificates approximate the demographic parity of the representation distribution. As the Gaussian noise σ increases from $\sigma = 0.005$ to $\sigma = 0.05$, the $\Delta(f_n^{plug}, F)$ empirical certificate upper bounds the demographic disparity $\Delta(f_{proc}, F)$ for any of the downstream processors

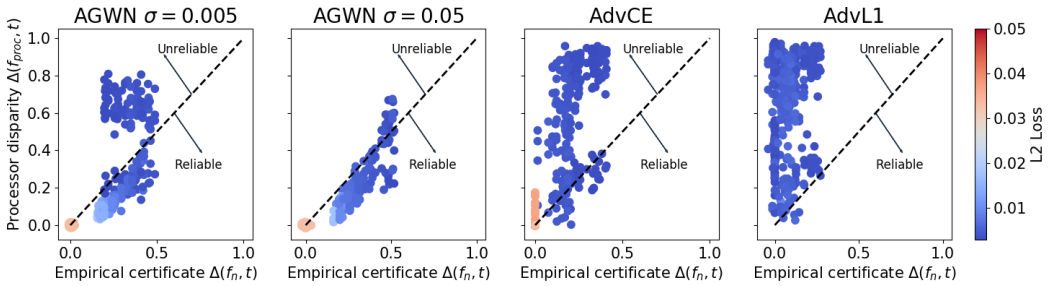


Figure 6.2: Generalization properties of empirical demographic parity certificates for DSprites. See Figure 6.1.

we built, regardless of their complexity. Moreover, the variance of $\Delta(f_n, F) - \Delta(f_{proc}, F)$ decreases as the Gaussian noise increases. This is consistent with the upper bound in Theorem 6.2.1, which decreases with smaller signal-to-noise ratio $||F||_\infty/\sigma$.

6.4.2 Comparative adversarial approaches.

Figure 6.1 and Figure 6.2 also show that for both comparative methods, the empirical certificate $\Delta(f_{adv}, F)$ estimated by the adversarial auditor underestimates significantly the disparity obtained by downstream processors on fresh samples from the representation distribution. For example, for the Swiss Roll dataset, 18.2% of near zero empirical certificates ($\Delta(f_{adv}, F) \leq 0.1$) do not preclude a processor's disparity larger than 0.3.

6.4.3 Real world data.

Figure 6.3 confirms that (i) the AGWN channel is sufficient for empirical certificates to upper-bound the demographic disparity obtained by various downstream processors; and, (ii) that comparative methods (AdvCE, AdvL1) generate empirical fairness certificates that do not bound the disparity of downstream processors.

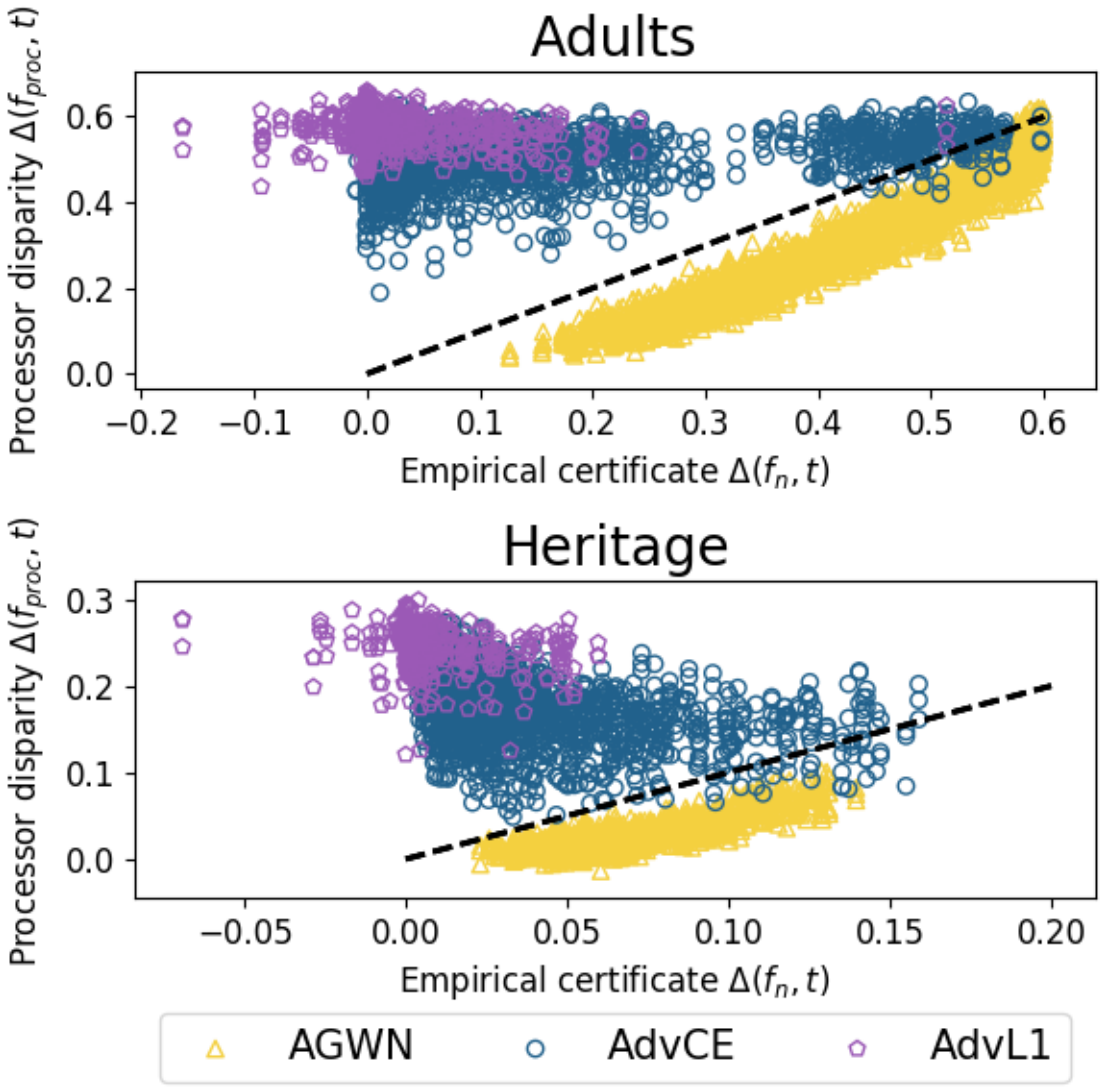


Figure 6.3: Generalization of empirical demographic parity certificates for Adults and Heritage. See Figure 6.1.

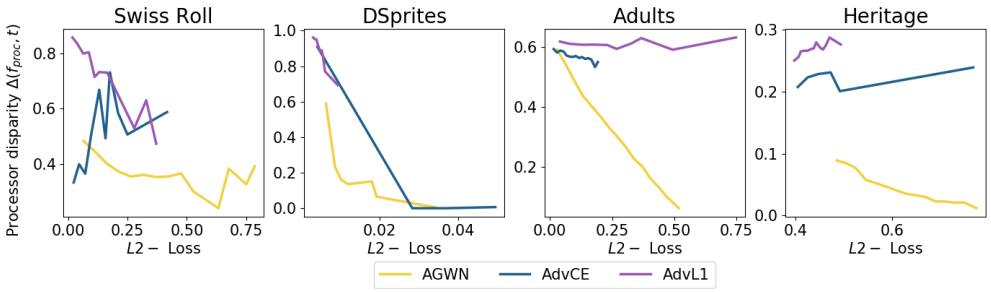


Figure 6.4: Reconstruction loss v.s. worst disparity attained by downstream processors.

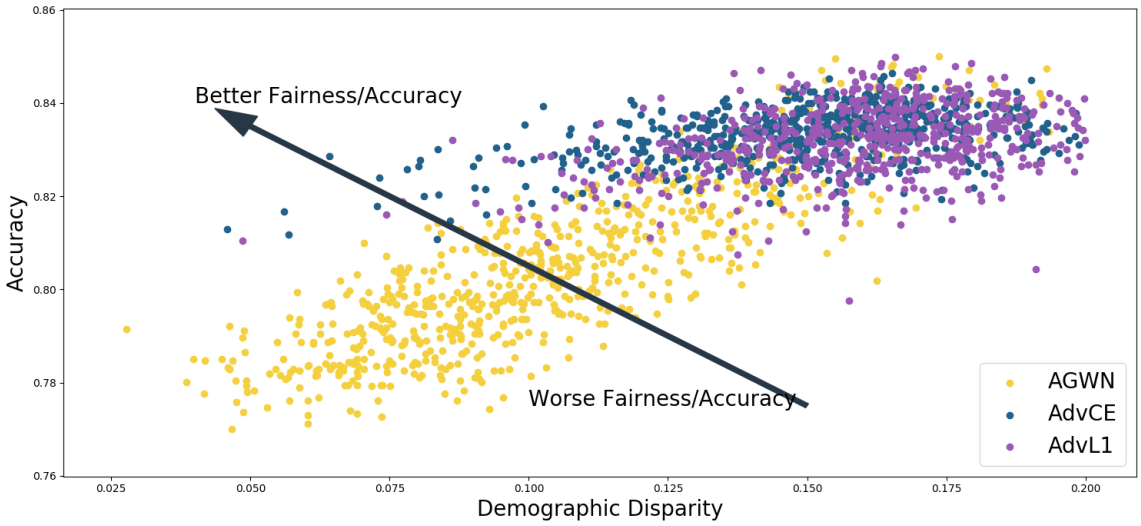


Figure 6.5: Accuracy-fairness trade-off.

6.4.4 Accuracy-fairness trade-off.

For the Swiss Roll dataset (Figure 6.1), AGWN's reliability appears to come at the cost of a larger reconstruction loss for a given empirical fairness certificate. However, it is not the case for DSprites. Moreover, a fair comparison across methods requires to measure reconstruction loss against the worst disparity attained by a downstream processor, i.e. the upper bound of the point clouds in Figure 6.1 and 6.2. Figure 6.4 plots the 95^{th} quantile of the demographic disparity of downstream processors for a given reconstruction loss. It shows that across all datasets, for a given $L2-$loss, the worst demographic disparity of downstream processors is lower when the representations are generated by AGWN than AdvCE or AdvL1. Moreover, for Swiss Roll and Adults, larger reconstruction losses (≥ 0.5 for Swiss Roll; ≥ 0.25 for Adults) with AGWN correspond to low levels of processors' disparity that are never reached by comparative methods.

To explore further how the AGWN channel affects the information contained in the representation, we compare the demographic disparity and the accuracy of downstream processors that predict a task label Y. We retrain the three fair learning methods – AdvCE, AdvL1 and AGWN – on the Adults dataset but leave out the income feature. We map test samples into their corresponding representations and predict whether their income is over $50K$. In Figure 6.5, we sweep the parameter space for different values of the fairness constraint λ in (6.12). Each dot compares the accuracy and the demographic disparity of neural networks of various depth and width. The higher the accuracy of downstream processors for a given level of disparity, the better the fairness-accuracy trade-off. We can draw two conclusions from this experiment. First, for level of disparity between 0.10 and 0.20, AGWN offers the same fairness-accuracy trade-off as AdvL1 or AdvCE. Second, our AGWN method is the only one for which varying the coefficient on the fairness constraint allows to systematically reach low level of disparity (≤ 0.1). Consistent with Figure 6.4, very few simulations of AdvCe and AdvL1 lead to the demographic disparity of downstream processors to be less than 0.075, regardless of the strength of the fairness penalty used during the training of the autoencoder. Although the AGWN channel limits the maximum

amount of information that is transferred from the data to the representation (see [63]), it also allows for a better empirical approximation of demographic parity and thus helps guiding the representation mapping toward the correct fairness-information trade-off.

6.5 Conclusion

This chapter investigates whether a data controller could generate representations of the data with fairness guarantees that would hold for any downstream processor using samples from the representation distribution. We show that for demographic parity certificate to approximate well the demographic parity of all future data processors it is necessary and sufficient to bound the χ^2 mutual information between feature and representation. To meet this condition, we show the benefit of adding an AGWN channel while learning a fair representation of the data.

Our work opens promising research avenues in fair representation learning. An AGWN channel may be only one of many approaches to bound the χ^2 mutual information between feature and representation. A comparison of competitive approaches would be crucial to improve the accuracy-fairness trade-off of learning reliably fair representations.

Chapter 7: Multi-Differential Fairness Auditor for Black Box Classifiers

In this chapter, we construct a tool, **mdfa**, that audits whether a classifier's outcomes depend on sensitive attributes once conditioned on a set of auditing features. First, we introduce a notion of parity, multi-differential fairness, that checks whether for any distribution over the feature space that is balanced across demographic groups, a classifier's outcomes are nearly mean-independent of sensitive attributes within any subset of the feature space. Second, we show that agnostic learning reduces to auditing for multi-differential fairness of black box classifiers and thus, we establish that in the worst-case auditing is NP-hard. Third, we reduce auditing to a weighted learning problem, where the weights are learned to minimize the maximum mean discrepancy between the distributions of features conditioned on sensitive attributes.

Empirically, we apply **mdfa** to a recidivism risk assessment tool and identifies subpopulation of African American defendants with little to no criminal history who are three times more likely to be considered at high risk of violent recidivism than similar individuals of other races.

The work presented in this chapter have been published in [32].

7.1 Individual and Multi-Differential Fairness

Preliminary. An individual i is defined by a tuple $((x_i, s_i), y_i)$, where $x_i \in \mathcal{X}$ denotes i's audited features; $s_i \in \mathcal{S}$ denotes the sensitive attributes; and $y_i \in \{-1, 1\}$ is a classifier f's outcome. The auditor draws m samples $\{((x_i, s_i), y_i)\}_{i=1}^m$ from a distribution on $\mathcal{X} \times \mathcal{S} \times \{-1, 1\}$. Features in $\mathcal{X}$ are not necessarily the ones used to train f, because the auditor may not have access to all features used to train f. Secondly, the auditor may decide to

deliberately leave out some features used to train f because those features – e.g. small geography identifiers – correlate strongly with sensitive attributes and that may break the following assumption of common support.

Assumptions. In our analysis, we assume that the distributions of auditing features conditioned on sensitive attributes have common support.

Assumption 1. *For all* $x \in \mathcal{X}$, $P(S|X = x) > 0$.

Definition 7.1.1. *(Individual Differential Fairness) For* $\delta \geq 0$, *a classifier* f *is* $\delta-$ *differential fair if* $\forall x \in \mathcal{X}, \forall s \in \mathcal{S}, \forall y \in \{-1, 1\}$

$$e^{-\delta} \leq \frac{P(Y = y|S = s, x)}{P(Y = y|S \neq s, x)} \leq e^{\delta} \tag{7.1}$$

The parameter δ controls how much the distribution of the classifier's outcome Y depends on sensitive attributes S given that auditing feature is x; larger δ implies a less differentially fair classifier. $\delta-$ differential fairness bounds the maximum divergence between the distributions $P(Y|S = s, x)$ and $P(Y|S \neq s, x)$:

$$\max_{y \in Y} \ln \left(\frac{P(Y|S = s, x)}{P(Y|S \neq s, x)} \right) \leq \delta$$

Strong notion of individual fairness We could use other notions of distance to define our notion individual differential fairness: possible candidates are the Kullback-Leibler divergence between $P(Y|X = x, S = s)$ and$P(Y|X = x, S \neq s)$

$$KL = E_y \ln \left(\frac{P(Y|S = s, x)}{P(Y|S \neq s, x)} \right); \tag{7.2}$$

or the total variation between $P(Y|X = x, S = s)$ and$P(Y|X = x, S \neq s)$

$$TV = \max_y |P(Y|S = s, x) - P(Y|S \neq s, x)|. \tag{7.3}$$

A definition based on a max divergence is a worst-case analogue of the KL divergence. Moreover, by the Pinsker's inequality, $TV \leq \frac{1}{2}\sqrt{KL}$. Therefore, our definition of differential individual fairness based on max divergence is stronger than counterparts based on total variation or KL divergence.

Relation with Differential Privacy. Differential fairness re-interprets disparate treatment as a differential privacy issue [75] by bounding the leakage of sensitive attributes caused by Y given what is already leaked by the auditing features x. Formally, the fairness condition (7.1) is equivalent to bounding the maximum divergence between the distributions $P(S|Y, x)$ and $P(S|x)$ by δ.

Individual Fairness. Def. (7.1.1) is an individual level definition of fairness, since it conditions the information leakage on auditing features x. Compared to the notion of individual fairness [27], individual differential fairness does not require an explicit similarity metric. This is a strength of our framework since defining a similarity metric has been the main limitation of applying the concept of individual fairness [43].

7.1.1 Multi-Differential Fairness.

Although useful, the notion of individual differential fairness cannot be computationally efficiently audited for. Looking for violations of individual differential fairness requires searching over a set of $2^{|\mathcal{X}|}$ individuals. Moreover, a sample from a distribution over $\mathcal{X} \times \mathcal{S} \times \{-1, 1\}$ has a negligible probability to have two individuals with the same auditing features x but different sensitive attributes s.

Therefore, we relax the definition of individual differential fairness and impose differential fairness for sub-populations. Formally, $\mathbb{C}$ denotes a collection of subsets or group of

individuals G in $\mathcal{X}$. The collection $\mathbb{C}_\alpha$ is α-strong if for $G \in \mathbb{C}$ and $y \in \{-1, 1\}$, $P(Y = y \;\&\; x \in G) \geq \alpha$. Relaxing differential fairness to sub-populations requires to audit only for balanced dataset $\mathcal{D}$ where for all $x \in \mathcal{X}$ and $s \in \mathcal{S}$, $P(S = s|X = x) = P(S \neq s|X = x)$.

Definition 7.1.2. *(Multi-Differential Fairness) Consider a α-strong collection $\mathbb{C}_\alpha$ of sub-populations of $\mathcal{X}$ and a balanced dataset $\mathcal{D} \subset \mathcal{X} \times \mathcal{S} \times \{0, 1\}$. For $0 \leq \delta$, a classifier f is $(\mathbb{C}_\alpha, \delta)$-multi differential fair with respect to $\mathcal{D}$ if $\forall s \in \mathcal{S}, \forall y \in \{-1, 1\}$ and $\forall G \in \mathbb{C}_\alpha$:*

$$e^{-\delta} \leq \frac{P(Y = y|S = s, G)}{P(Y = y|S \neq s, G)} \leq e^{\delta} \tag{7.4}$$

Multi-differential fairness guarantees that the outcome of a classifier f is nearly mean-independent of protected attributes within any sub-population $G \in \mathbb{C}_\alpha$. There are two important restrictions to the definition of multi-differential fairness. First, the fairness condition in Eq. 7.4 applies only to α-strong collection of sub-populations with $P(Y = y, G) \geq \alpha$ for $y \in \{-1, 1\}$. This condition avoids trivial cases where $\{x \in G, Y = y\}$ is a singleton for some y, implying $\delta = \infty$.

Second, the fairness condition in Eq. 7.4 applies only to auditing dataset $\mathcal{D}$ that are balanced with respect to sensitive attributes S. This condition avoids trivial cases where outcomes correlate with sensitive attributes because of data imbalance. For example, suppose that $\mathcal{X} = [0, 1]$, $Y = 1$ if and only if $X \geq 0.5$, $S = 1$ if and only if $X \geq 0.5 + \epsilon$ for some small $\epsilon > 0$. Choose $G = [0.4, 0.6]$. Then, $P[Y = 1|S = 1, G]/P[Y = 1|S = -1, G] = (\epsilon + 0.5)/\epsilon$ can be made arbitrarily large as $\epsilon \to 0$. The issue with unbalanced dataset is that the information related to S leaked by the classifier's outcomes is confounded with the information leaked by membership in G since $(\mathcal{C}_\alpha, \delta)$ multi-differential fairness implies that :

$$\max_{s,y} \ln\left(\frac{P[S = s|Y = y, G]}{P[S \neq s|Y = y, G]}\right) - \ln\left(\frac{P[S = s|G]}{P[S \neq s|G]}\right) \leq \delta \tag{7.5}$$

On the other hand, a balanced distribution does not leak any information on whether a

sensitive attribute is equal to s: $P(S = s|x) = P(S \neq s|x)$.

Collection of Indicators. We represent the collection of sub-populations $\mathbb{C}$ as a family of indicators: for $G \in \mathbb{C}$, there is an indicator $c : \mathcal{X} \to \{-1, 1\}$ such that $c(x) = 1$ if and only if $x \in G$. The relaxation of differential fairness to a collection of groups or sub-population is akin to [41,42,79]. $\mathbb{C}_\alpha$ is the computational bound on how granular our definition of fairness is. The richer $\mathbb{C}_\alpha$, the stronger the fairness guarantee offers by Def. 7.1.2. However, the complexity of $\mathbb{C}_\alpha$ is limited by the fact that we identify a sub-population G via random samples drawn from a distribution over $\mathcal{X} \times \mathcal{S} \times \{-1, 1\}$.

7.2 Auditing as an Agnostic Learning Problem

We first consider the problem of finding violations of multi-differential fairness by sampling from a balanced dataset $\mathcal{D}$, that is from a dataset for which $P(S = s|X = x) = P(S \neq s|X = x)$. This is not a realistic assumption since in most real-world applications, the dataset at hand will be unbalanced and would need to be re-balanced before we can audit for multi-differential fairness. We will tackle the imbalance issue in the next section.

We use the balanced setting to demonstrate the hardness of certifying for the lack of differential fairness. Formally, we reduce auditing for multi-differential fairness to an agnostic learning problem. We observe that if the data distribution is balanced, finding a violation of $(\mathbb{C}_\alpha, \delta)$- multi differential fairness is equivalent to finding a sub-population $G \in \mathcal{C}_\alpha$, a $y \in \{-1, 1\}$ and $s \in \mathcal{S}$ such that

$$P(G, Y = y) \left\{ Pr(S = s|G, Y = y) - \frac{1}{2} \right\} \geq \gamma, \tag{7.6}$$

with $\gamma = \alpha \left(e^\delta/(1 + e^\delta) - 1/2 \right)$. γ combines the size of the sub-population where a violation exists and the magnitude of the violation. We call a $\gamma-$ unfairness certificate any triple (G, y, s) that satisfies Eq. (7.6). Further we postulate that f is $\gamma-$unfair if and only if such

certificate exists. Unfairness for balanced distributions is equivalent to the existence of sub-populations for which sensitive attributes can be predicted once the classifier's outcomes are observed.

Searching for γ-unfairness certificate reduces to mapping the auditing features $\{x_i\}$ to the labels $\{s_i y_i\}$.

Lemma 7.2.1. *Let $s \in \mathcal{S}$. Suppose that the data is balanced. f is $\gamma-$ multi-differential unfair for $y \in \{-1, 1\}$ if and only there exists $c \in \mathbb{C}_\alpha$ such that $Pr(ySY = c) \geq 1-\rho(y)+4\gamma$, where $\rho(y) = P(S = yY)$.*

Lemma 7.2.1 allows us to reduce searching for a (G, y, s) unfairness certificate to predicting where sensitive attribute and outcomes of f (if $y = 1$) or outcomes of $\neg f$ (if $y = -1$) coincide. Since f is a black-box classifier, the function $g(x, s) = sf(x)$ is only accessed via a sample $\mathcal{D}_n = \{x_i, s_i, y_i\}_{i=1}^n$. Therefore, searching for unfairness certificate is akin to learn from a finite sample of $\mathcal{D}_n$ a hypothesis $c \in \mathcal{C}$ that approximates well the predictions from $g : \mathcal{X} \times \mathcal{S} \to \{-1, 1\}$. Note that there is no guarantee that $g \in \mathcal{C}$ and thus the learning is agnostic. The optimal membership indicator in $\mathcal{C}$ has an error rate $\min_{c\in\mathcal{C}} P(g(x, s) \neq c(x))$ that is not necessarily zero. The next result shows a necessary and sufficient condition on $\mathcal{C}$ to learn from a finite sample a sub-population c with an error rate that approximates well the optimal one.

Theorem 7.2.2. *Let $\epsilon, \beta > 0$ and $\mathcal{C} \subset 2^{\mathcal{X}}$. Let $\gamma^{'} \in (\gamma-\epsilon, \gamma+\epsilon)$. The following statements are equivalent:*

(i) There exists an algorithm that by using $O(\log(|\mathcal{C}|), \log(\frac{1}{\eta}), \frac{1}{\epsilon^2})$ samples $\{(x_i, s_i), y_i\}$ drawn from a balanced distribution D outputs with probability $1 - \eta$ a $\gamma^{'}$-unfairness certificate if y_i are outcomes from a $\gamma-$unfair classifier;

(ii) $\mathcal{C}$ is agnostic learnable: there exists an algorithm that with $O(\log(|\mathcal{C}|, \log(\frac{1}{\eta}), \frac{1}{\epsilon^2})$ samples $\{x_i, s_i, o_i\}$ drawn from a balanced distribution D outputs with probability $1 - \eta$, $Pr_D[h(x_i) = o_i] + \epsilon \geq max_{c\in\mathbb{C}} Pr_D[c(x_i) = o_i]$.

Our reduction to agnostic learning means that the granularity of any auditing algorithm for differential fairness is limited by the difficulty to approximate g with concepts from the class $\mathcal{C}$ while using a finite sample. Indeed, $\mathcal{C}$ needs an agnostic learner for the search of an unfairness certificate to be poly-logarithmic in $|\mathcal{C}|$.

Theorem 7.2.2 shows the hardness of auditing a black-box classifier for multi-differential fairness. In the worst-case, for many classes $\mathcal{C}$, agnostic learning and thus auditing is NP-hard [103]. It means that there is a computational limit on how granular multi-differential fairness can be: richer $\mathcal{C}$ – larger $|\mathcal{C}|$ – means that auditing searches for more complex sub-populations, but at the cost of looser guarantees on the generalization of the error rate.

7.3 A Learning Algorithm to Audit for Multi-Differential Fairness

Using a sample $\mathcal{D}_n$ from a balanced dataset $\mathcal{D}$, our reduction suggests to solve the following empirical loss minimization:

$$\min_{c \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(c \neq s_i y_i). \tag{7.7}$$

In practice, we optimize over a family $\mathcal{H}$ from $\mathcal{X}$ to $[0, 1]$ and then estimate an unfairness certificate by using the sub-population indicator $c(x) = \text{sign}(h(x))$. Moreover, since the $0-1$ loss is not differentiable, we will use a convex and differentiable proxy l for it and minimize the empirical loss

$$R(h) \triangleq \frac{1}{n} \sum_{i=1}^{n} l(h(x_i), s_i y_i), \tag{7.8}$$

where $h \in \mathcal{H}$.

At issue is that the auditor has only access to a sample $\mathcal{D}_n^{'}$ from a unbalanced dataset $\mathcal{D}^{'}$. Therefore, it will need first to balance the data and then, solve the empirical loss minimization (7.7).

7.3.1 Unbalanced Data

To extend our auditing approach to unbalanced dataset, we propose to first rebalance the data. Assume that the minority demographic group corresponds to $S = 1$ and denote π_s the density $P(x, S = s)$.

Definition 7.3.1 (Re-weigthing)**.** *A function $u : \mathcal{X} \to \mathbb{R}$ is a valid re-weigthing if for all $x \in \mathcal{X}$, $\pi_1(x) > 0$ implies $u(x) > 0$. We denote the re-weighted density $\pi_1^u(x) = u(x)\pi_1(x)$.*

We would like to choose a re-weighting scheme such that $\pi_1^u(x) = \pi_{-1}(x)$. A natural candidate for u is to choose for $w(x) = \pi_{-1}(x)/\pi_1(x)$. However, in practice we do not have direct access to $\pi_s(x)$. One approach is to directly estimate the density $P[S = s|x]$. This method is used in propensity-score matching methods [31] in the context of counterfactual analysis. But, exact or estimated importance sampling results in large variance in finite sample [104]. Instead, we use a kernel-based matching approach [105, 106].

Given a re-weigthing function u, we denote $R^u(h)$ the risk

$$R^u(h) \triangleq E_{x,y\sim\pi_1}[]u(x)l(h(x), y)] + E_{x,y\sim\pi_{-1}}[]l(h(x), -y)], \tag{7.9}$$

and $R_n^u(h)$ its empirical counterpart.

Integral Probability Metric

Our approach relies on integral probability metrics (IPM) [107] to measure the distance between two probability measures P and Q. For a family $\mathcal{G}$ of functions $g : \mathcal{X} \to \mathbb{R}$, we define

$$IPM_{\mathcal{G}}(P, Q) = \sup_{g\in\mathcal{G}} \left| \int_{\mathcal{X}} gdP - \int_{\mathcal{X}} gdQ \right|. \tag{7.10}$$

If the family of functions is rich enough, integral probability metrics define a true metric over the set of probability measures, i.e. $IPM_{\mathcal{G}}(P, Q) = 0$ implies $P = Q$. In this chapter, we choose for $\mathcal{G}$ the unit ball of functions in a universal reproducing Hilbert kernel space.

For this choice of $\mathcal{G}$, $IPM_{\mathcal{G}}$ is a true metric named the maximum mean discrepancy [105] and has an empirical estimator with a convergence rate independent of the dimension of $\mathcal{X}$ [107].

Reproducing Kernel Hilbert Space

Formally, we consider a reproducing kernel Hilbert space $\mathcal{G}_k$ with kernel k such that $||k||_\infty < \infty$. $\mathcal{G}$ is then defined as the unit ball in this reproducing kernel Hilbert space: $\mathcal{G} = \{g \in \mathcal{G}_k | ||g||_k \leq 1\}$. We choose the kernel k such that the reproducing kernel Hilbert space has the universal approximating property:

Definition 7.3.2 (Universal kernel)**.** *A kernel k has the universal approximating property if given any compact subset $\mathcal{Z}$ of $\mathcal{X}$, any $\epsilon > 0$ and any continuous bounded function $l : \mathcal{Z} \to \mathbb{R}$, there exists $g_{l,\epsilon} \in \mathcal{G}_k$ such that $||g_{l,\epsilon} - l||_\infty \leq \epsilon$.*

Example of universal kernels are Gaussian radial-based kernels. Kernel universality allows to control the shift in multi-differential unfairness certificate γ when the re-weigthing u is used instead of w^*.

We measure the gap between π_1^u and π_{-1} by measuring:

$$IPM_{\mathcal{G}}(\pi_1^u, \pi_{-1}) = \sup_{g \in \mathcal{G}} \left| E_{x \sim \pi_1^u} g(x) - E_{x \sim \pi_{-1}} g(x) \right| \tag{7.11}$$

The next result bounds the error in the fairness risk $R_u(h)$ while using a re-weighting u instead of w:

Lemma 7.3.1. *Suppose that k is a universal kernel. Let $\epsilon > 0$. For $h \in \mathcal{H}$,*

(i) If $l \circ h \in \mathcal{G}_k$, for any re-weighting u

$$|R^u(h) - R^w(h)| \leq ||l \circ h||_k IPM_{\mathcal{G}}(\pi_1^u, \pi_{-1}) \tag{7.12}$$

(ii) If $l \circ h$ is a bounded continuous function from $\mathcal{X}$ to $\mathbb{R}$, there exists $g_{l,\epsilon} \in \mathcal{G}_k$ such that

for any re-weighting u

$$|R^u(h) - R^w(h)| \leq \epsilon + ||g_{l,\epsilon}||_k IPM_{\mathcal{G}}(\pi_1^u, \pi_{-1}). \tag{7.13}$$

The bound of Lemma 7.3.1 is tighter if (i) π_1^u is closer to π_{-1} according to (7.11); (ii) the loss $l \circ h$ belongs to the reproducing kernel space $\mathcal{G}_k$; (iii) the loss $l \circ h$ is smooth enough in x. Importance sampling w leads to a tight bound in expectation since $\pi_1^w = \pi_{-1}$, but are not practical since $P(S|x)$ is not observed. The norm $||g||_k$ or $||l \circ h||_k$ measures the complexity or smoothness of the loss $l \circ h$. Note that $||g||_k$ hides a dependence in ϵ, since the smaller ϵ, the larger $||g||_k$ is likely to be.

Using Lemma 7.3.1, we upper bound the error between the empirical risk with re-weighting u and the risk $R^w(h)$ if the data is re-balanced using importance sampling weights w.

Theorem 7.3.2. *Let $\delta > 0$. Consider a sample $\mathcal{D}_n = \{(x_i, s_i, y_i)\}_{i=1}^n$, with $n_s = |\{i|s_i = s\}|$. Assume that $u : \mathcal{X} \times \mathcal{S}$ is a re-weighting function. Assume that exists $B > 0$ such that $||g_{l,\epsilon}||_k \leq B$ for all $\epsilon > 0$, where $g_{l,\epsilon} \in \mathcal{G}_k$ is defined as in 7.3.2 and $\mathcal{G}_k$ is the reproducing kernel Hilbert space associated with kernel k. Assume that $\sup_{g \in \mathcal{G}} |g(x)| \leq \nu$, Then, for any $\delta > 0$ with probability at least $1 - \delta$, for any $h \in \mathcal{H}$*

$$\begin{aligned} |R^w(h) - R_n^u(h)| \leq & B \times IPM_{\mathcal{G}}(\pi_{n,1}^u, \pi_{n,-1}) + 2^{5/4} V(u) \left(\frac{d_{\mathcal{H}} \log \frac{2n}{d_{\mathcal{H}}} + \log \frac{16}{\delta}}{n} \right)^{3/8} \\ & + B \sqrt{18\nu^2 \log \frac{8}{\delta} ||k||_\infty} \left(\frac{1}{\sqrt{n_1}} + \frac{1}{\sqrt{n_{-1}}} \right), \end{aligned} \tag{7.14}$$

where $\pi_{n,1}^u, \pi_{n,-1}$ are the empirical estimates of π_n^u, π_{-1};

$$V(u) = \max\{\sqrt{E_{\pi_1^u}[u^2(x) l^2(h(x))}, \sqrt{E_{\pi_{n,1}^u}[u^2(x) l^2(h(x))}\}; \tag{7.15}$$

and, $d_{\mathcal{H}}$ is the pseudo-dimension of $\{l \circ h | h \in \mathcal{H}\}$.

The upper-bound in Theorem 7.3.2 makes explicit that higher variance of the term $u(x,s)l \circ h(x)$ degrades the approximation of $R^w(h)$ by $R^u_n(h)$. This is consistent with results in [104]: choosing $u \neq w$ generates a bias in R^u – larger $IPM_{\mathcal{G}}(\pi^u_1, \pi_{-1})$ – but could lower the variance – smaller $V(u)$.

Moreover, the upper bound in Theorem 7.3.2 depends on the pseudo-dimension $d_{\mathcal{H}}$ of the class of auditors $\mathcal{H}$. The larger $d_{\mathcal{H}}$, the more granular is the fairness criteria for which the auditor searches for fairness violations, but this granularity comes at the cost of a higher variance of $R^u(h)$. This is consistent with our hardness result in Theorem 7.2.2.

The term B captures the complexity of the $l \circ h$.

7.3.2 Auditing Algorithm for Unbalanced Data

Motivated by our theoretical insights of Theorem 7.3.2, we propose an auditing algorithm that given a sample $\mathcal{D}_n$, (i) learns the re-weighting u by mininizing a combination of $IPM_{\mathcal{G}}(\pi^u_1, \pi_{-1})$ and the variance of u; and, (ii) learns an unfairness certificate by minimizing $R^u_n(h)$. The training objective of our algorithm is then

$$\begin{aligned}\mathcal{L}(h,u;\beta,\lambda) = \underbrace{\frac{1}{n}\sum_{i=1:s_i=1}^{n} u(x_i)l \circ h(x_i) + \frac{1}{n}\sum_{i=1:s_i=-1}^{n} l \circ h(x_i)}_{R^u_n(h)} \\ + \beta \underbrace{IPM_{\mathcal{G}}(\pi^u_{n,1}, \pi_{n,-1}) + \lambda||u(x)l \circ h(x)||_2}_{\mathcal{L}_{bal}(h,u;\beta,\lambda)}\end{aligned} \tag{7.16}$$

where $\mathcal{L}_{bal}(h,u;\beta,\lambda)$ includes two regularization terms, a term $IPM_{\mathcal{G}}(\pi^u_{n,1}, \pi_{n,-1})$ to penalize discrepancies between π^u_1 and π_{-1}; and, a term $||u(x)l \circ h(x)||_2$ to penalize the variance of the re-weighted loss. (β, λ) are hyperparameters.

7.3.3 Worst-Case Violation

Solving the empirical minimization Eq. (7.16) allows certifying whether any black box classifier is multi-differential fair, but the solution of Eq. (7.16) does not distinguish a large sub-population S with low value of δ from a smaller sub-population with larger value of δ. For example, consider two sub-populations of same size G_0 and G_δ for $\delta > 0$. Assume that there is no violation of multi-differential fairness on G_0, but a $\delta-$ violation on G_δ. The risk minimization Eq. (7.16) will pick indifferently G_δ and $G_\delta \cup G_0$ as unfairness certificates, although G_0 mixes the violation G_δ with a sub-population without any violation of differential fairness.

Worst-Case Violation Algorithm (WVA). At issue in the previous example is that for the sub-population G_0, choosing $c = 1$ or $c = -1$ will lead to the same empirical risk Eq. (7.16). To force $c(x) = -1$ for $x \in G_0$, our approach is to add a regularization term that penalizes large value of h and thus, forces the solution of (7.16) to assign small values to h whenever $s = y$ with probability $1/2$. Formally, the training objective of our algorithm is

$$\mathcal{L}(h, u; \beta, \lambda) + \frac{\mu_t}{n} \sum_{i=1}^{n} u(x_i) h(x_i) \tag{7.17}$$

where μ_t is hyperparameter for which we choose a following annealing strategy: starting from $\mu_0 = 0$, at each iteration t, we increase the value of μ_t by a step ξ. This temperature schedule terminates whenever the resulting sub-population $G_t = \{x \in \mathcal{X} | \operatorname{sign}(h_t(x)) = 1\}$ is less than α. At the second to the last iteration T, we estimate the accuracy of the classifier $c_T = \operatorname{sign}(h_T)$ at predicting the labels SY when sampling from the dataset re-weighted by u_T.

7.3.4 Mdfa Auditor

Putting the building blocks together allows us to design a fairness diagnostic tool **mdfa** that identifies efficiently the most severe violation of multi-differential unfairness.

Estimating the maximum mean discrepancy

For our choice of $\mathcal{G}$ as the unit ball of functions in a reproducing Hilbert kernel space, the integral probability metric $IPM_{\mathcal{G}}(\pi_1^u, \pi_{-1})$ is the maximum mean discrepancy between π_1^u and π_{-1} and its empirical estimator is available in closed form ([107])

$$IPM_{\mathcal{G}}(\pi_{n,1}^u, \pi_{n,-1}) = \sqrt{\sum_{i,j=1}^{n} \tilde{s_i}\tilde{s_j}k(x_i, x_j)}, \tag{7.18}$$

where k is the kernel associated with the reproducing Hilbert kernel space; $\tilde{s_1} = u(x)/n_1$; $\tilde{s_{-1}} = -1/n_{-1}$, where n_s is the number of sample points with $S = s$. We choose a normalized Gaussian kernel, with a scale hyperparameter σ.

Architecture

Inputs are a dataset with a classifier's outcomes (labels ± 1) along with auditing features. **mfda** models h and u as neural networks, whose depth and width depends on the data at hand. That is, **mfda** is made of two parallel neural networks that are combined at the outset to compute the weighted risk R^u and the imbalance loss $\mathcal{L}_{bal}$. We solve the empirical loss (7.17) by stochastic gradient descent. For every τ iterations of the stochastic descent, we increase the value of the parameter μ by a step ξ if the size of $\{x \in \mathcal{X} | \operatorname{sign}(h_t(x)) = 1\}$ is larger than α.

Cross-Validation

The auditor chooses the minimum size α of the worst-case violation they would like to identify. The advantage of our approach is that, although we do not have ground truth for unfair treatment, we can propose heuristics to cross-validate our choice of regularization parameters λ, β and σ used in Eq. (7.17). First, we split 70%/30% the input data into a train and test set. Starting with $\mu = 0$ and using a $5-$fold cross-validation, **mdfa** is trained on four folds and a grid search looks for regularization parameters σ, β and λ that minimize

the imbalance loss $\mathcal{L}_{bal}$ with $h = 1$. Once the values for σ, β and λ are set, we minimize the complete loss (7.17).

7.4 Experimental Results

7.4.1 Synthetic Data

A synthetic data is constructed by drawing independently two features X_1 and X_2 from two normal distributions $N(0, 1)$. We consider a binary protected attribute $\mathcal{S} = \{-1, 1\}$ drawn from a Bernouilli distribution with $S = 1$ with probability $w(x) = \frac{e^{\mu*(x_1-x_2))^2}}{1+e^{\mu*(x_1+x_2))^2}}$. μ is the imbalance factor. $\mu = 0$ means that the data is perfectly balanced. The data is labeled according to the sign of $(X_1 + X_2 + e)^3$, where is e is a noise drawn from $N(0, 0.2)$. The audited classifier f is a logistic regression classifier that is altered to generate instances of differential unfairness. For $x_1^2 + x_2^2 \leq 1$, if $S = -1$, the classifier's outcomes Y is changed from -1 to 1 with probability $1-\nu \in (0, 1]$; if $S = 1$, all $Y = -1$ are changed to $Y = 1$. For $\nu = 0$, the audited classifier is differentially fair; however, as ν increases, in the half circle $\{(x_1, x_2)|x_1^2 + x_2^2 \leq 1 \text{ and } y = -1\}$ there is a fraction ν of individuals with $S = 1$ who are not treated similarly as individuals with $S = -1$.

Results

First, we test whether **mdfa** identifies the worst-case violation of multi-differential fairness that occurs in the sub-space $\{(x_1, x_2)|x_1^2 + x_2^2 \leq 1 \text{ and } y = -1\}$. In Figure 7.1 (left), **mdfa** is trained using a neural network with one hidden layer of 32 neurons on a unbalanced data ($\mu = 0.2$). The true value of δ varies from 0 to 0.7 (i.e ν varying from 0 to 0.5). Figure 7.1 compares the estimated value δ against the true one δ_{true} and shows that **mdfa**'s estimate is unbiased, since the plots aligns wells with the 45^o diagonal but at very low value of δ_{true}. Increasing the sample size reduces the variance of **mdfa**'s estimator. In Figure 7.1, right, we test the effect of the complexity of the auditor on the bias and variance of **mdfa**'s

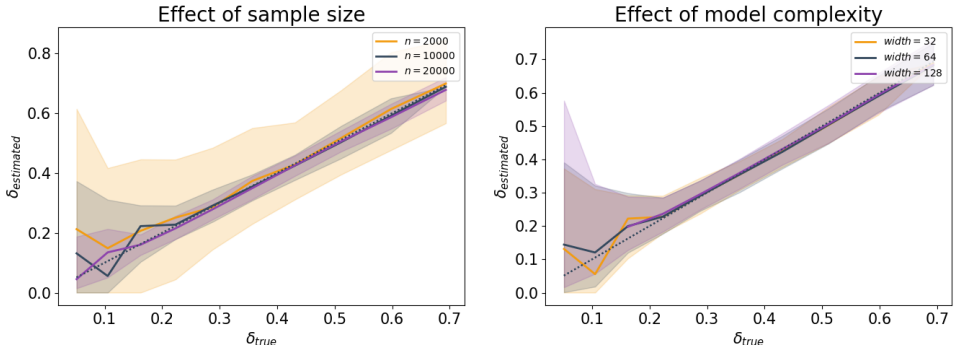


Figure 7.1: Performances of **mdfa** on synthetic data. Shaded area shows the 90% confidence interval of $\delta_{estimated}$ that is obtained by simulating 100 synthetic data for a given value of ν. The balancing factor μ is set to -0.2.

estimator. Models with wider hidden layers appear to have more variance.

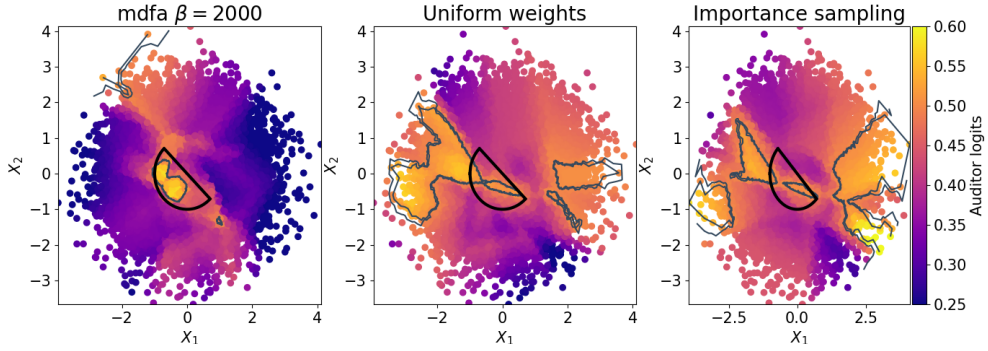


Figure 7.2: Auditing performances for different balancing schemes. The data is colored by the outputs of the last layer of the auditor neural network, once activated by a sigmoid function. The gray contour represents the area identified by the auditor as violation of multi-differential fairness. The black semi-circle represents the true region with a violation of multi-differential fairness.

We compare our balancing approach MMD to alternative re-balancing approaches: (i) uniform weights with $u(x) = 1/n_1$ for all x and (ii) importance sampling with exact weights $w(x)$. UW applies **mdfa** without rebalancing. IS uses an estimate of the probabilities $P(S = s|X = x)$ obtained by training a neural network that predicts S from X. In Figure

7.2, we apply a sigmoid activation to the last layer of the auditor neural network and color the data points with these logits. Figure 7.2 shows that only **mdfa** is able to indentify with higher logits the region of the feature space where the violation of multi-differential fairness occurs. Absent of a re-weighting scheme, the auditor errs to identify the region $\{(x_1, x_2)|x_1^2 + x_2^2 \leq 1 \text{ and } y = -1\}$. Using importance sampling weights directly does not perform well: this confirms previous observations in the literature that in finite sample, the variance of the importance sample weights can be detrimental to a re-balancing approach.

7.4.2 Case Study: COMPAS

We apply our method to the COMPAS algorithm, widely used to assess the likelihood of a defendant to become a recidivist ([5]). The research question is whether without knowledge of the design of COMPAS, **mdfa** can identify group of individuals that could argue for a disparate treatment. The data collected by ProPublica in Broward County from 2013 to 2015 contains $7K$ individuals along with a risk score and a risk category assigned by COMPAS. We transform the risk category into a binary variable equal to 1 for individuals assigned in the high risk category (risk score between 8 and 10). The data provides us with information related to the historical criminal history, misdemeanors, gender, age and race of each individual.

Worst Violations. We run **mdfa** on 100 different 70/30% train/test splits and report average value of auditing features and recidivism risk for the whole population and the worst-case subpopulation in Table 7.1. The first two columns show that the distribution of features in the whole population is disperse and differs between African American (AA) and Other. This is due to the data imbalance issue (c.f. Section 3). The probability of being classified as high risk is 0.14 for African-American, thereby 2.7 times higher than for non-African American. However, it is unclear whether that difference could be explained either by the distribution imbalance or by the classifier's disparate treatment. The two last columns in Table 7.1 show that in the sub-population "violation" extracted by **mdfa**,

Table 7.1: Identifying the worst-case violation of differential fairness in the COMPAS risk score. The sensitive attribute is whether the individual is self-identified as African American (AA) or not (Other). ( ) indicates standard deviation.

Variable	Population		Violation	
	AA	Other	AA	Other
Prior Felonies	4.44	2.46	0.79	0.67
	(5.58)	(3.76)	(0.24)	(0.17)
Charge Degree	0.31	0.4	0.74	0.74
	(0.46)	(0.49)	(0.23)	(0.2)
Juvenile Felonies	0.1	0.03	0.01	0.0
	(0.49)	(0.32)	(0.02)	(0.02)
Juvenile Misdemeanor	0.14	0.04	0.01	0.01
	(0.61)	(0.3)	(0.02)	(0.01)
High Risk	0.14	0.05	0.06	0.02
	(0.35)	(0.22)	(0.04)	(0.01)

the distribution of features is narrower and similar for African-American and non-African American: the sub-population is made of individuals with little criminal and misdemeanor history. However, African American are still three times more likely to be classified as high risk. A policy implication of **mdfa** findings is that a judge using COMPAS may discount its assessment for African-American with little criminal history.

7.4.3 Group Fairness vs. Multi-Differential Fairness

We evaluate whether previous fairness correcting approaches protect small group of individuals against violation of differential fairness. We consider two techniques: (i) [36]'s disparate impact repair with a logistic classification ($DI - LC$) and (ii) [108]'s reduction with a logistic regression ($Red - LC$). We use **mdfa** to identify sub-population G with worst-case violations and measure sub-population disparate treatment as $DT_G = P(Y = 1|S = 1, G)/P(Y = 1|S = -1, G)$. We compare DT_G to its aggregate counterpart computed on the whole population $DI = P(Y = 1|S = 1)/P(Y = 1|S = -1)$.

Data. The experiment is carried on three datasets from [28, 109]): **Adult** with $48,840$ individuals; **German** with 1000 individuals; and, **Crimes** with 1994 communities. In

Adult the prediction task is whether an individual's income is less than $50K$ and the sensitive attribute is gender; in **German**, the prediction task is whether an individual has bad credit and the sensitive attribute is gender; in **Crimes**, the task is to predict whether a community is in the 70^{th} percentile for violent crime rates and the sensitive attribute is whether the percentage of African American is at least 20%. For each data, each repair technique produces a prediction; then, **mdfa** is trained on 70% of the data and computes estimates for disparate treatment DT_G on the remaining 30% of the data. The experiment is repeated with 100 train/test splits.

Results. In Table 7.2, even despite the fairness correction applied by $DI-LC$ and $Red-LC$, **mdfa** still finds sub-populations G for which DT_G is significantly larger than one. It indicates the existence of group of individuals who are similar but for their sensitive attributes and who are treated differently by the classifier trained by either $DI-LC$ or $Red-LC$. The repair techniques reduce the aggregate disparate impact compared to the baseline (LC), since DI is closer to one for $DI-LC$ and $Red-LC$ across all datasets. However, in the **Adult** dataset, DT_G remains between 1.44 and 1.6 after repair: **mdfa** identifies a group G of Females that are $44\%-60\%$ more likely to be classified as low-income than Males with similar characteristics. In **Crimes** dataset, disparate treatment DT_G is around 5.7 for both $DI-LC$, $R-LC$: this means that there exist communities with dense African-American populations that are six times more likely to be classified at high risk than similar communities with lower percentages of African Americans.

7.5 Conclusion

In this chapter, we present **mdfa**, a tool that measures whether a classifier treats differently individuals with similar auditing features but different sensitive attributes. We hope that **mdfa**'s ability to identify sub-populations with severe violations of differential fairness could inform decision-makers when to discount the classifier's outcomes. It also provides the victims with a framework to contest a classifier's outcomes.

Table 7.2: Worst-case violations of multi-differential fairness identified by **mdfa** for classifiers trained with standard fairness repair techniques. ( ) indicates standard deviation.

Repair Technique	Adult		German		Crimes	
	DT_G	DI	DT_G	DT	DT_G	DI
LC	1.88	1.08	1.26	1.07	5.76	1.0
	(0.4)		(0.14)		(3.16)	
DI-LC	1.44	0.99	1.1	1.04	5.74	1.0
	(0.32)		(0.08)		(2.19)	
Red-LC	1.6	1.03	1.04	1.01	5.24	1.0
	(0.25)		(0.21)		(0.89)	

Avenues for future research are to investigate (i) the properties of a classifier trained under a multi-differential fairness constraint; and, (ii) the possibility to extend our approach to re-balance distributions in order to make counterfactual inference [85] in the context of algorithmic fairness.

Chapter 8: Conclusions

8.1 Summary of Findings

In this thesis, we present two types of methods to mitigate potentially unfair outcomes of a black-box classifier: (i) unsupervised fair representation learning and (ii) auditing.

8.1.1 Unsupervised Fair Representation Learning

This thesis expands recent contributions in fear representation learning to the unsupervised setting: the user generates a fair encoding of the data without knowledge of how the resulting encoding will be used. This unsupervised setting differs from previous work in fair representation learning that generates encoding tailored to a specific task. The advantage of all purpose fair representations is that the transformation can be performed at the time of collecting the data, potentially on distributed devices.

Our first step toward unsupervised fair representation learning is to make tight connections with rate-distortion problems. We show that encoding a data into a representation that does not leak information related to a sensitive attribute is equivalent to solving a rate-distortion problem, provided that a side channel gives the decoder direct access to the sensitive attribute. We formalize the notion of unfairness-distortion functions as the minimum mutual information between sensitive attribute and representation for a given level of distortion. Chapter 4 proves that unfairness-distortion functions can be completely derived from rate-distortion functions.

This result allows to solve fair information bottleneck via compression-based technique (chapter 3); explore at test time multiple points in the fairness-information plane with one single trained model; and, propose the first application of the fair representation paradigm to images via hierarchical quantization (chapter 5).

A second step toward practical deployment of fair representation models is to earn the user's trust. In chapter 6, we ask which statistical guarantees we can offer to the user that the representation hides her sensitive attribute to any classifier that will use the data. The answer is striking: there is no guarantee unless the χ^2- mutual information between the data and the representation is finite. This result leads to a simple solution toward certifying the fairness property of a representation: we show that fair representation learning models can offer statistical hiding guarantees by adding a Gaussian noise to the representation.

A natural question is whether our compression-based approaches to fair representation learning can offer the same statistical hiding properties since they are aiming at minimizing the Shannon mutual information between the data and the representation. The rationale would be that both Shannon mutual information and χ^2- mutual information controls how much information is encoded in Z. It turns that in theory, there exist representations with finite Shannon mutual information and infinite χ^2- mutual information. However, in practice, we find empirically (chapter 3,4 and 5) that compression-based techniques generate representations with fairness properties that hold against diverse downstream classifiers.

8.1.2 Auditing Black Box Classifiers

The second part of this thesis looks at the other extremity of the data science pipeline, once a black-box classifier has used some data or its representation as inputs. In chapter 7, we propose a method, mdfa, to estimate whether there exist sub-populations or subsets of the feature space where the classifier has exacerbated existing biases encoded in the features themselves.

Mdfa identifies sub-populations, if any, for which a black-box classifier leaks more information about the sensitive attributes than the features themselves. Identifying these 'fairness leakages' is a first and important step toward empowering victims of classifier discriminatory outcomes and allowing contestability.

8.2 Limitations and Ethical Implications

This thesis considers classifiers as black-boxes and describe or control the fairness properties of the black-box via its inputs – in fair representation learning – and its outputs – in auditing. Abstracting away the inner working of the decision making system is typical in computer science. We discuss in this section the benefits, ethical implications and limitations of this abstraction.

A benefit of our black-box abstraction is that it is flexible with respect to the data science pipeline. Our proposed unsupervised pre-processing methods operate as portable plug-and-play tools that can be readily deployed across media (tabular data in chapter 3 and 4; image data in chapter 5). It opens avenues to deploy the tools on edge devices, which is critical for practical application of fairness paradigms [78, 110].

Our pre-processing – and to some extend, auditing – approaches also abstract away how to define fairness by re-casting fairness as a privacy issue: we design systems so that we can measure / control how much information about a sensitive attribute the release of a data leaks (chapter 3 to 6); and, how exacerbated the privacy leakage is after the data is ingested by a classifier (chapter 7). This formalism is inspired by differential privacy [75] and allows to quantify how robust are the fairness guarantees offered by a system. For example, in chapter 6, we show that with proper noise structure, organizations can offer certifiable guarantees on how much their pre-processed data hides the sensitive attributes.

We are aware that our level of abstraction bounds the scope of our work since it does not account much for the context surrounding the data and the data mining tasks. This thesis does not model the social components that interact with the system. We do not provide guidance in how to articulate and solve the socio-cultural tensions and debates inherent to decision making systems that affect individual well-being. For example, we do not expand on how to define sensitive attributes in the first place. Nor do we propose any framework to analyze how social biases are encoded in a data. As discussed in [111], there could be a debate on whether the limitations discussed here lead the proposed methods to

a solutionism trap. Ideally, fair machine learning needs heterogeneous engineering [112] to build sociotechnical systems that consider how technology interacts with social actors [113]. We argue here that this thesis offers technical and mathematical guidelines on what fairness constraints are feasible under the reasonable assumption that decision making systems are black-boxes. To some extent, an hybrid approach to fair machine learning benefits from understanding how a bank of diverse tools can address different social concerns.

8.3 Future Research

This thesis opens promising avenues for research in fair machine learning in the context of black-box classifiers. In this section, we present two main themes: (i) extension of unsupervised fair representation learning to distributed systems; (ii) development of additional auditing tools to build fairness unit tests for developers and foster stakeholder engagement.

8.3.1 Federated Fair Representation Learning

Unsupervised fair representation learning offers a flexible pre-processing technique for any data owner/controller to offer fairness guarantees without knowing the future of the data. However, in this thesis, we assume that the data owner/controller is a central entity that collects a sample of data to train a fair encoder-decoder and potentially deploys the trained encoder-decoder to edge devices. However, in many applications, data owners are decentralized entities that do not necessarily consent sharing their data with a central server.

One area for future research is to train the fair encoder-decoder within a federated learning framework [110] that exchanges parameters or gradient values between decentralized entities, but does allow a central server to access the data. Main challenges to federated fair representation learning include (i) data distribution shifts across diverse clients [114]; (ii) collaboration of decentralized devices; and, (iii) privacy of the sensitive attribute. It is an open question to understand how distribution shifts between clients would affect the aggregate fairness properties of the learned representation; and whether each client would

have an incentive to collaborate. Privacy of the sensitive attribute is also essential to foster collaboration between edge devices: a user would agree to participate to a federated representation learning framework only if sharing gradient or parameter updates does not reveal its sensitive attribute.

8.3.2 Pre-and-Post Auditing

In the presence of black-box classifiers, we could expand upon our auditing work in Chapter 6 along three research directions: (i) individualized post-mortem auditing; (ii) auditing of utility-based decisions making systems; and, (iii) pre-emptive auditing of a data to prevent discriminatory outcome of classifiers using the data.

Chapter 7's auditing tool is not individualized and remains at the aggregate level since membership in the audited sub-group is defined by an indicator function that belongs to a class of functions with limited dimensions. A natural follow up is to explore the possibility of a more granular auditing tool that would estimate disparate treatment at the individual level. In many real world, auditing for individual disparate treatment faces three challenges: (i) we do not observe the counterfactual outcome for the same individual but with different sensitive attributes; (ii) the distributions differ when conditioned on sensitive attributes; and, (iii) the outcomes are binary and most of the literature on individual treatment effect has focused on continuous outcomes.

Moreover, our definition of multi-differential fairness in Chapter 7 along with other popular definitions (statistical parity [27], equalized odds and opportunity [38]) only applies to binary settings and implicitly assumes that the utility of an individual is equal to one when the algorithm's outcome is one and equal to zero otherwise. This is a limitation since in many societal applications of machine learning, utilities are heterogeneous across individuals and this heterogeneity could be systematic across demographic groups [115, 116]. However, necessary trade-offs exist for fairness in the context of utility-based decision making system: in a work adjacent to this thesis, we show that many reasonable definitions of equitable outcomes cannot hold simultaneously except under stringent conditions [34].

Lastly, in standard machine learning pipelines [117] users discover unfairness issues after the models are trained, validated, and sometimes deployed [118]. This is inefficient whenever the problem resides in the data itself and could have been diagnosed ex-ante before model development. Unfairness in machine learning is first and foremost rooted in the data itself. Of interest is the development of tools that would identify which data presents fairness risk when ingested by a data pipeline.

Appendix A: Proofs of Results in Single Shot Fair Representation Learning

A.1 Proof of Theorem 2.1

First, we show the following identity:

Lemma A.1.1. $I(Z,S) = I(Z,X) + H(X|Z,S) - H(X|S)$.

Proof. The proof of Lemma A.1.1 relies on multiple iterations of the chain rule for mutual information:

$$
\begin{aligned}
I(Z,S) &\overset{(a)}{=} I(Z,\{X,S\}) - I(Z,X|S) \\
&\overset{(b)}{=} I(Z,X) + I(Z,S|X) - I(Z,X|S) \\
&\overset{(c)}{=} I(Z,X) - I(Z,X|S) \\
&\overset{(d)}{=} I(Z,X) - I(X,\{Z,S\}) + I(X,S) \\
&\overset{(e)}{=} I(Z,X) - H(X) + H(X|Z,S) \\
&+ H(X) - H(X|S) \\
&= I(Z,X) + H(X|Z,S) - H(X|S)
\end{aligned}
$$

where (a), (b) and (d) use the chain rule for mutual information; and, (c) uses the fact that Z is only encoded from X and from S, so $H(Z|X,S) = H(Z|X)$ and $I(Z,S|X) = H(Z|X) - H(Z|X,S) = 0$. And (e) uses the fact that $I(X,S) = H(X) - H(X|S)$ and $I(X,\{Z,S\}) = H(X) - H(X|Z,S)$. □

Lemma A.1.1 implies that if the distortion is $d(X, \{Z, S\}) = H(X|Z, S)$, the unfairness-distortion function is given by

$$\begin{aligned} I(D) = &\min_F I(Z, X) + H(X|Z, S) - H(X|S) \\ &\text{s.t. } H(X, \{Z, S\}) \leq D \end{aligned} \tag{A.1}$$

Second, a fundamental theorem in rate-distortion shows that if the distortion is $d(X, \{Z, S\}) = H(X|Z, S)$ the rate-distortion function is given by

$$R(D) = \min_F I(X, Z) \text{ s.t } H(X|Z, S) \leq D, \tag{A.2}$$

and that $R(D)$ is a non-increasing convex function. The next Lemma shows how solution of the minimization problem (A.2) solves the minimization problem (A.1) whenever $\frac{\partial R(D)}{\partial D} \leq -1$

Lemma A.1.2. *Let $D \geq 0$ be a distortion value. Assume that $\frac{\partial R(D)}{\partial D} \leq -1$. A solution F^* of the minimization* (A.2) *for D is also solution of* (A.1).

Proof. At the optimum, the constraint in (A.2) is binding and thus, that $H_{f^*}(X|Z, S) = D$, where the subs-script F^* reminds that the code Z depends on F^*. Consider now a solution g^* of the minimization (A.1) for a distortion D. We consider two cases: case (I) the constraint is binding for g^* in (A.1); case (II) the constraint is not binding for g^* in (A.1).

case (I): $H_{g^*}(X|Z, S) = D$ and we have

$$\begin{aligned} I(D) &= I_{g^*}(Z, X) + H_{g^*}(X|Z, S) - H(X|S) \\ &= I_{g^*}(Z, X) + D - H(X|S) \\ &\overset{(a)}{\geq} I_{f^*}(Z, X) + D - H(X|S), \end{aligned} \tag{A.3}$$

where (a) uses the fact that F^* is solution of (A.2) and that $H_{g^*}(X|Z,S) \leq D$. Therefore, since $H_{f^*}(X|Z,S) \leq D$, F^* is also solution of (A.1).

case (II): Let denote $D^{'}$ the value of the distortion achieved by g^*. Then, $D^{'} = H_{g^*}(X|Z,S) < D$. We have

$$\begin{aligned} I(D) &= I_{g^*}(Z,X) + H_{g^*}(X|Z,S) - H(X|S) \\ &= I_{g^*}(Z,X) + D^{'} - H(X|S) \\ &\overset{(a)}{\geq} R(D^{'}) + D^{'} - H(X|S), \end{aligned} \tag{A.4}$$

where (a) follows from the definition of $R(D^{'})$. By convexity of the rate-distortion function, we have that

$$\begin{aligned} R(D^{'}) - R(D) &\overset{(a)}{\geq} \frac{\partial R(D)}{\partial D}(D^{'} - D) \\ &\overset{(b)}{\geq} (D - D^{'}), \end{aligned} \tag{A.5}$$

where (a) uses the convexity of $R(D)$ and that $D^{'} < D$ and (b) uses that $\frac{\partial R(D)}{\partial D} \leq -1$. Hence, by combining (A.4) and (A.5), we have

$$I(D) \geq R(D) + D - H(X|S) = I_{f^*}(Z,X) + D - H(X|S). \tag{A.6}$$

Therefore, F^* is also solution of the minimization (A.1) since $H_{f^*}(X|Z,S) \leq D$. □

It follows from Lemma A.1.2 that we have by definition of F^*, if $\frac{\partial R(D)}{\partial D} \leq -1$

$$I(D) = I_{f^*}(Z,X) + D - H(X|S) = R(D) + D - H(X|S), \tag{A.7}$$

which proves the first part of the statement in Theorem 2.1. Moreoover, if $\frac{\partial R(D)}{\partial D} < -1$,

$\frac{\partial I(D)}{\partial D} = \frac{\partial R(D)}{\partial D} + 1 < 0$, hence $I(.)$ is decreasing for D such that $\frac{\partial R(D)}{\partial D} < -1$.

To prove that if $\frac{\partial R(D)}{\partial D} \geq -1$, $I(D) = 0$, we first prove the following Lemma:

Lemma A.1.3. *Let D^* denote the value of D such that $\frac{\partial R(D)}{\partial D} = -1$. For $D^* \leq D$, $I(D) = I(D^*)$.*

Proof. Let $D > D^*$. Let g^* be a solution of the minimization (A.1) for D. Note that a solution of (A.1) for D^* respects the constraint of the minimization (A.1) for D and thus, $I(D^*) \geq I(D)$. Let $D^{'}$ denote $H_{g^*}(X|Z,S)$. Then, by definition of the rate-distortion objective value (A.2), we have

$$\begin{aligned} I(D) &= I_{g^*}(Z,X) + D^{'} - H(X|S) \\ &\geq R(D^{'}) + D^{'} - H(X|S). \end{aligned} \tag{A.8}$$

If $D^{'} < D^*$, then we already know that $I(D^{'}) = R(D^{'}) + D^{'} - H(X|S)$ and that $I(D^{'}) > I(D^*) \geq I(D)$. Moreover, by inequality (A.8), $I(D) \geq I(D^{'})$, thus $I(D^{'}) > I(D) \geq I(D^{'})$, which is a contradiction. If $D^{'} = D^*$, we already know that $I(D) \leq I(D^*) = R(D^*) + D^* - H(X|S) = I(D^{'}) \leq I(D)$ and thus that $I(D) = I(D^*)$.

It remains to look at the case $D^{'} > D^*$. Consider $D^{"} \in [D^*, D^{'}]$. By convexity of $R(D)$ we have

$$\begin{aligned} R(D^*) - R(D^{'}) &\leq \frac{\partial R(D^*)}{\partial D}(D^* - D^{'}) \\ &\overset{(a)}{=} D^{'} - D^*, \end{aligned} \tag{A.9}$$

where (a) comes the fact that $\frac{\partial R(D^*)}{\partial D} = -1$. It results that by the inequality (A.7) $I(D) \geq R(D^*) + D^* - H(X|S)$. Moreover, we already know that $R(D^*) + D^* - H(X|S) = I(D^*)$. Hence $I(D^*) \geq I(D) \geq I(D^*)$, which proves the equality in Lemma A.1.2. □

Lemma A.1.4. *Let $D^{**} = H(X|S)$. We have $I(D^{**}) = 0$.*

Proof. Consider an encoder g that generates a random variable Z independent of X. Then $H_g(X|Z,S) = D^{**}$ and $I_g(Z,X) = 0$. Therefore, g respect the constraint of the minimization (A.1) for D^{**} and $I(D^{**}) \leq I_g(Z,X) + H_g(X|Z,S) - H(X|S) = 0$. Hence, $I(D^{**}) = 0$. □

By combining Lemma A.1.2 and A.1.4, we can show that $I(D) = 0$ for $D \geq D^{**}$.

A.2 Lower Bound on $I(Z,S)$

When constructing unfairness-distortion curves, we approximate the mutual information $I(Z,S)$ with an adversarial lower bound. For any approximation $q(s|Z)$ of $p(s|Z)$, we have

$$\begin{aligned} I(Z,S) &= H(S) - H(S|Z) \\ &= H(S) - E_{s,z}[-\log q(s|z)] + KL(p(s|z)||p(s|z) \\ &\geq H(S) - E_{s,z}[-\log q(s|z)], \end{aligned} \tag{A.10}$$

where the inequality comes from the non-negativity of the Kullback-Leibler divergence $KL(p|q)$. Therefore, we lower bound $I(Z,S)$ with

$$H(S) - \min_q E_{s,z}[-\log q(s|z)], \tag{A.11}$$

where the minimum is taken over classifiers that predict S from Z.

A.3 Bit Disparity

In the main test, we make the following claim:

Lemma A.3.1. *For $\mathcal{S} = \{0, 1\}$, $I(Z, S) \geq g(\pi, \Delta(b))$, where g is an increasing non-negative convex function and $\pi = P(S = 1)$.*

Proof. The proof is based on a result from [91] applied to a classifier c_b that returns 1 if and only $b = 1$: the demographic disparity of c_b is exactly $\Delta(b)$ and thus, by Theorem 2.1 in [91], there exists a non-negative, convex and increasing function g such that

$$I(Z, S) \geq g(\pi, \Delta(b)). \tag{A.12}$$

□

Appendix B: Proofs of Results in Learning Smooth and Fair Representations

B.1 Proof of Theorem 1

The proof of Theorem 1 uses the following lemma (from [36]) that links the demographic parity of a test function f and its balanced error rate $BER(f)$,

$$BER(f, F) = \frac{P(f(Z) = 1|S = 0) + P(f(Z) = 0|S = 1)}{2}, \tag{B.1}$$

where we make the dependence on the representation mapping F explicit in $BER(f, F)$.

Lemma B.1.1. *[36] A representation space $(\mathcal{Z}, \mu_F)$ satisfies an $\Delta^*(F)-$ demographic parity certificate if and only if*

$$BER^*() \triangleq \min_{f:\mathcal{Z}\to\{0,1\}} BER(f, F) \geq \frac{1 - \Delta}{2}. \tag{B.2}$$

Therefore, a representation space $(\mathcal{Z}, \mu_F)$ can be stamped with a $\Delta^*(F)-$ demographic parity certificate with $\Delta^*(F) \equiv 1 - 2BER^*(F)$.

To prove the result in Theorem 1, we consider a deterministic transformation F.

Lemma B.1.2. *Suppose that F is a deterministic mapping from $\mathcal{X}$ to $\mathcal{Z}$. Denote K the size of $F(\mathcal{X})$ with $K \leq \infty$. Then, for all distribution μ_x over the features $\mathcal{X}$ such that for all $z \in F(\mathcal{X})$, $P(F(X) = z) > 0$, $I_{\chi^2}(X, Z) = K - 1$.*

Proof. First, since F is a function, $P(Z = z|X = x)$ is equal to one if and only if $F(x) = z$.

Therefore,

$$
\begin{aligned}
I_{\chi^2}(X,Z) &= E_x\left(1-\frac{1}{P(Z=F(x))}\right)^2 P(Z=F(x)) \\
&= E_x\left[\frac{1}{P(Z=F(x)}\right]-1 \\
&= \sum_{z\in F(\mathcal{X})}\left[\frac{P(X,F(X)=z)}{P(Z=z)}\right]-1 \\
&= K-1
\end{aligned}
\tag{B.3}
$$

□

Now for a given distribution μ_x over $\mathcal{X}$ and a given transformation F, there are two cases: $I_{\chi^2}(X,Z)=\infty$ and $I_{\chi^2}(X,Z)<\infty$. Let denote $I_{\chi^2}(X,Z)$ by I_{χ^2}.

B.1.1 Case $I_{\chi^2}<\infty$

By lemma B.1.2, $F(\mathcal{X})$ is finite and $F(\mathcal{X})=\{z_1,z_2,...,z_K\}$, with $K\leq\infty$ and $z_k\neq z_{k'}$ for $k\neq k^{'}$.

For each $k\in\{1,...,K\}$, we choose one $x_k\in\mathcal{X}$ such that $F(x_k)=z_k$. We parametrize a family of joint distributions $\mu(b)$ over $[0,1]\times\{0,1\}$ as follows: X is uniformly distributed over $\{x_1,...,x_K\}$; and, for $b\in(0,1)$, the sensitive attribute is given by k^{th} binary expansion of b, where $X=x_k$. By Lemma B.1.2, the χ^2 squared mutual information between X and $F(X)$ is the same for any b and equal to $K-1$. Moreover, since the sensitive attribute is a function of $F(X)$, $\Delta_b^*(F)=1$, where the subscript indicates that demographic parity is computed using the joint distribution $\mu(b)$ over (Z,S).

Let B denote a random variable uniformly distributed on $[0,1]$. For any auditor f_n,

$$\begin{aligned}
\sup_{b\in[0,1]} E_{\mathcal{D}_n(b)} BER(f_n, F) &\overset{(a)}{\geq} E_B E_{\mathcal{D}_n(B)} BER(f_n, F) \\
&= E_{X,B} P[f_n((F(X), \mathcal{D}_n(B)) \neq \\
&\quad S | F(X_1), ..., F(X_n), S_1, ...S_n, F(X)] \\
&\overset{(b)}{\geq} \frac{1}{2} P\left(\cap_{i=1}^n [F(X) \neq F(X_i)]\right) \\
&\overset{(c)}{=} \frac{1}{2}\left(1 - \frac{1}{K}\right)^n \\
&\overset{(d)}{=} \frac{1}{2}\left(1 - \frac{1}{I_{\chi^2}(X,Z)}\right)^n
\end{aligned} \tag{B.4}$$

where (a) uses that the suppremum is larger than the average; (b) that for $Z \notin \{Z_1, ..., Z_n\}$, the sensitive attribute has a Bernouilli distribution with probability $1/2$; (c) that X and then Z is uniformly distributed; and, (d) that $I_{\chi^2}(X,Z) \leq K$ by Lemma B.1.2. Since $I_{\chi^2}(X,Z)$ is equal for all b, it follows from Lemma B.1.1 that

$$\sup_{b\in(0,1)} \Delta^* - \Delta(f_n, F) \geq \left(1 - \frac{1}{I_{\chi^2}(Z,X)}\right)^n. \tag{B.5}$$

Note that μ does not depend on μ_x. Therefore, for all auditors f_n,

$$\sup_{\mu} E_{\mathcal{D}_n} |\Delta^* - \Delta(f_n, F)| \geq \left(1 - \frac{1}{I_{\chi^2}}\right)^n. \tag{B.6}$$

B.1.2 Case $I_{\chi^2} = \infty$

By Lemma B.1.2, if for a distribution μ over $\mathcal{X} \times \{0,1\}$, $I_{\chi^2}(Z,X) = \infty$, then there exists an infinite countable set $\{a_k\}$ of $\mathcal{X}$ such that F takes a different value at each a_k. We choose

X to take value in $\{a_k\}_{k\geq 1}$ such that $P(a_k) = p_k$ for $k \geq 0$ where the sequence $\{p_k\}_{k=1}^{\infty}$ will be chosen later on. As in the previous case, we parametrize a family of distributions over $\mathcal{X} \times \{0,1\}$ by $b \in (0,1)$ such that for $X \in \{a_1, ...\}$, the sensitive attribute S is the k^{th} term of b's binary expansion, where $X = a_k$. Because S is a deterministic function of X, $\Delta^*(F) = 1$.

Let B denote a random variable uniformly distributed on $[0,1]$. For a sample point X_i, we denote k_i such that $X_i = a_{k_i}$. For any auditor f_n,

$$\begin{aligned}
\sup_{b\in[0,1]} E_{\mathcal{D}_n(b)} BER(f_n, F) &\overset{(a)}{\geq} E_B E_{\mathcal{D}_n(B)} BER(f_n, F) \\
&= E_{X,B} P[f_n((F(X), \mathcal{D}_n(B)) \neq \\
&S|F(X_1), ..., F(X_n), S_1, ...S_n, F(X)] \qquad \text{(B.7)} \\
&\overset{(b)}{\geq} \frac{1}{2} P\left(\cap_{i=1}^{n} [k \neq k_i]\right) \\
&\overset{(c)}{=} \frac{1}{2} \sum_{k=1}^{\infty} p_k (1-p_k)^n
\end{aligned}$$

It remains to show that for all $\epsilon > 0$, we can choose $\{p_k\}$ such that the right hand side of inequality (B.7) is at least $1/2(1-\epsilon)$. Let $\epsilon > 0$. We choose p_k as follows. First, pick $K > \frac{1}{1-(1-\epsilon)^{1/n}}$.Then, let $p_k = 1/K$ for $1 \leq k \leq K$ and $p_k = 0$ elsewhere. It follows that

$$\sup_{b\in[0,1]} E_{\mathcal{D}_n(b)} BER(f_n, F) \geq \frac{1}{2}\left(1 - \frac{1}{K}\right)^n \geq \frac{1}{2}(1-\epsilon). \qquad \text{(B.8)}$$

Therefore, using Lemma B.1.1, we can conclude that for all $\epsilon > 0$, there exists a distribution over $\mathcal{X} \times \{0,1\}$ such that for all auditors f_n

$$\Delta^*(F) - \Delta(f_n, F) \geq 1 - \epsilon. \qquad \text{(B.9)}$$

Therefore,

$$\sup_{\mu} \Delta^*(F) - \Delta(f_n, F) \geq 1 = \left(1 - \frac{1}{I_{\chi^2}}\right)^n. \tag{B.10}$$

B.1.3 Final Step

Therefore, by combining both cases $I_{\chi^2} < \infty$ and $I_{\chi^2} = \infty$, we have that for all distribution μ_x over the features $\mathcal{X}$,

$$\sup_{\mu} \Delta^*(F) - \Delta(f_n, F) \geq 1 = \left(1 - \frac{1}{I_{\chi^2}}\right)^n, \tag{B.11}$$

which implies the result in theorem 1.

B.2 Proof of Corollary 1

Suppose that $\inf_{f_n \in \mathcal{F}_n} \sup_{\mu} E_{\mathcal{D}_n} |\Delta^* - \Delta(f_n, F)| \leq \epsilon_n$ for some $\epsilon_n > 0$. Let $f_n \in \mathcal{F}_n$ be the auditor that reaches the minimum.

We have, for any distribution μ over $\mathcal{X} \times \{0, 1\}$,

$$\begin{aligned}
\left(1 - \frac{1}{I_{\chi^2}(Z, X)}\right)^n &\leq \sup_{\mu} \left(1 - \frac{1}{I_{\chi^2}(Z, X)}\right)^n \\
&\overset{(a)}{\leq} \sup_{\mu} E_{\mathcal{D}_n} |\Delta^* - \Delta(f_n, F)| \\
&\leq \epsilon_n,
\end{aligned} \tag{B.12}$$

where (a) uses Theorem 1. The result follows directly from equation (B.12).

B.3 Examples of Representation Mappings without Finite Sample Guarantees

Injective mappings. Suppose that F is injective from $[0,1]^D$ to $\mathbb{R}^d$.

Consider X distributed over the countable and infinite set $\{1, 1/2, ...1/k, ...\}$ with $p_k = \kappa/k^2$ and $k^{-1} = \sum_{k=1}^{\infty} 1/k^2$. By lemma B.1.2, $I_{\chi^2}(X,Z) = \infty$ and thus, by Corollary 1, there exists a distribution such that $\Delta^*(F) - \Delta(f_n, F) = 1$ for all f_n.

Large $F(\mathcal{X})$. Suppose that $|\{F(x)|x \in \mathcal{X}\}| \geq n/(\ln(n))^{\alpha}$, for some $\alpha < 1$.

By Lemma B.1.2, $I_{\chi^2}(X,Z) \geq n/(\ln(n))^{\alpha} - 1$ and thus, by Corollary 1, if $\inf_{f_n \in \mathcal{F}_n} \sup_{\mu} E_{\mathcal{D}_n} |\Delta^*(F) - \Delta(f_n, F)| = \epsilon_n$, then

$$\begin{aligned} \frac{n}{(\ln(n))^{\alpha}} - 1 \leq I_{\chi^2}(X,Z) &\leq \frac{1}{1 - \epsilon_n^{\frac{1}{n}}} \\ &\overset{(a)}{\leq} \frac{n}{-\ln(\epsilon_n)}, \end{aligned} \tag{B.13}$$

where (a) uses that $e^{-x} \geq 1 - x$. Therefore, $\epsilon_n \geq e^{-(\ln(n))^{\alpha}} = \omega(n^{-s})$ for $s > 0$, since $\alpha < 1$.

B.4 Proof of Theorem 2

The proof Theorem 2 relies on a upper bound of $\Delta^*(F) - \Delta(f_n, F)$ that uses the total variation distance $TV(\mu_F^s, \mu_n^s)$ between class conditional densities and their empirical counterpart:

$$TV(\mu_F^s, \mu_n^s) = \int |\mu_F^s - \mu_n^s| dz. \tag{B.14}$$

Lemma B.4.1. *Consider a sample $\{(z_i, s_i)\}_{i=1}^n$ from a representation distribution μ_F induced by a representation rule F. Suppose that μ_n^0 and μ_n^1 are empirical density estimators of $P(Z|S=0)$ and $P(Z|S=1)$ respectively. Denote f_n the following auditing plug-in*

decision: for $z \in \mathcal{Z}$, $f_n(z) = 1$ *if and only if* $\mu_n^1(z) > \mu_n^0(z)$. *Therefore, for all* n

$$\Delta(f_n, F) \leq \Delta^*(F) \leq \Delta(f_n, F) + 2 \sum_{i=0,1} TV(\mu_F^i, \mu_n^i). \tag{B.15}$$

Proof. Let f^* denote the auditing rule that minimzes the balance error rate. Using [101] (ch 2), we show that for any auditing rule f_n

$$\begin{aligned} 2 - \int \eta_{f_n(z)}(z)\mu_F(dz) &= 2 - \sum_{i=0,1} \int_{f_n(z)=i} \eta_i(z)\mu_F(dz) \\ &= 2 - \sum_{i=0,1} \int_{f_n(z)=i} P(z|S=i)dz \\ &= 2BER(f_n), \end{aligned} \tag{B.16}$$

where $\eta_i(z)$ is the balanced posteriori probability $\eta_i(z) = P(S = i|Z = z)/P(S = i)$. Moreover,

$$\begin{aligned} 2BER(f^*) &= 2 - P(f^*(z) = 1|S = 1] \\ &\quad - P(f^*(z) = 0|S = 0) \\ &= 2 - \int_{z,\mu_F^1 > \mu_F^0} \mu_F^1(dz) - \int_{z,\mu_F^0 > \mu_F^1} \mu_F^0(dz) \\ &= 2 - \int \max_i \eta_i(z)\mu_F(dz). \end{aligned} \tag{B.17}$$

Let denote $\eta_{n,i}$ the empirical estimate of η_i. Using equations (B.16) and (B.17), the proof of lemma B.4.1 relies on the fact that

$$
\begin{aligned}
BER(f_n) - BER(f^*) &= \int \max_i \eta_i(z)\mu_F(dz) \\
&\quad - \int \eta_{f_n(z)}(z)\mu_F(dz) \\
&= \int (\max_i \eta_i(z) - \max_i \eta_{n,i}(z))\mu_F(dz) \\
&\quad + \int (\eta_{n,f_n(z)}(z) - \eta_{f_n(z)}(z))\mu_F(dz) \\
&\overset{(a)}{\leq} \sum_{i=0,1} \int |\eta_i(z) - \eta_{n,i}(z)|\mu_F(dz) \\
&= \sum_{i=0,1} \int |\mu_t^i(z) - \mu_n^i(z)|dz,
\end{aligned}
\tag{B.18}
$$

The inequality (a) comes from the following observation. If the maxima are attained for the same $i \in \{0, 1\}$, then the right hand side integrand is equal to 0. Otherwise, suppose without loss of generality that $\max \eta_i(z)$ is reached for $i = 0$, then the right hand side integrand is

$$
\begin{aligned}
\eta_0(z) - \eta_{n,1}(z) + \eta_{n,1}(z) - \eta_1(z) &= \eta_0(z) - \eta_{n,0}(z) \\
&\quad + \eta_{n,1}(z) - \eta_1(z) \\
&\quad + \eta_{n,0}(z) - \eta_{n,1}(z) \\
&\quad \leq |\eta_0(z) - \eta_{n,0}(z)| \\
&\quad + |\eta_1(z) - \eta_{n,1}(z)|,
\end{aligned}
\tag{B.19}
$$

where the inequality follows $\max_i \eta_{n,i}(z) = \eta_{n,1}(z)$. The same argument can be applied

when $\max \eta_i(z) = \eta_1(z)$. The result in lemma B.4.1 follows from (B.18). $\square$

The second part of the proof of theorem 2 is to show that the total variation distance between μ_n^s and μ_F^s is $O(1/\sqrt{n_s})$ for some empirical estimate of μ_F^s:

Lemma B.4.2. *Consider a representation mapping $F : \mathcal{X} \to \mathcal{Z}$ and its induced distribution μ_F. Assume that $I_2(Z, X) < \infty$. Then, for $s = 0, 1$, define μ_n^s as*

$$\mu_n^s(z) = \frac{1}{n_s} \sum_{i=1, s_i=s}^{n} P(z|X = x_i) \tag{B.20}$$

The total variation between μ_t^s and μ_n^s can be bounded as follows:

$$E_{\mathcal{D} \sim \mathcal{X}^n} \left[TV(\mu_F^s, \mu_n^s) \right] \leq \sqrt{\frac{I_2(Z, X)}{n_s}}.$$

The upper bound of the total variation distance uses a Monte Carlo integration argument. For a sample $\mathcal{D}_n = \{x_i\}_{i=1}^n$, denote $\phi(z, x_i)$ the probability $P(Z = z|X = x_i)$. Therefore, $\mu_F(z) = E_{x \sim \mathcal{X}}[\phi(z, x)]$ and if μ_n^s is defined as in (B.20), $\mu_F^s(z) = E_{\mathbf{X}, S=s}[\mu_n^s]$, where $\mathbf{X} = \{x_i\}_{i=1}^n \sim \mathcal{X}^n$. Denote

$$\mathcal{E}^s(\mathbf{X}) = \int \left| \mu_F(z) - \frac{1}{n_s} \sum_{i=1, s_i=s}^{n} \phi(z, x_i) \right| dz, \tag{B.21}$$

with $n_s = |\{i|s_i = s\}|$. We have

$$
\begin{aligned}
E_{\mathbf{X}}[\mathcal{E}^s(\mathbf{X})] &\overset{(a)}{\leq} E_{\mathbf{X}}\left[\sqrt{\int \left(\frac{\mu_F(z) - \mu_n^s(z)}{\mu_F(z)}\right)^2 \mu_F(z)dz}\right] \\
&\overset{(b)}{=} \frac{1}{n_s} E_{\mathbf{X}}\left[\sqrt{\int \sum_{i=1,s_i=s}^{n} \left(\frac{\mu_F(z) - \phi(z,x_i)}{\mu_F(z)}\right)^2 \mu_F(z)dz}\right] \\
&\overset{(c)}{\leq} \frac{1}{n_s}\sqrt{E_{\mathbf{X}}\left[\int \sum_{i=1,s_i=s}^{n} \left(\frac{\mu_F(z) - \phi(z,x_i)}{\mu_F(z)}\right)^2 \mu_F(z)dz\right]} \\
&\overset{(d)}{=} \frac{1}{n_s}\sqrt{\sum_{i=1,s_i=s}^{n} E_{\mathbf{X}}\left[\int \left(\frac{\mu_F(z) - \phi(z,x_i)}{\mu_F(z)}\right)^2 \mu_F(z)dz\right]} \\
&\overset{(e)}{=} \sqrt{\frac{I_2(Z,X)}{n_s}},
\end{aligned}
\tag{B.22}
$$

where (a) applies Cauchy-Schwarz inequality; (b) uses the fact that the samples are independently drawn and that $E_{x_i}[\phi(z,x_i)] = \mu_F(z)$; (c) that the squared-root is concave; (d) that expectation and integral can be interchange; and, (e) the definition of the chi-squared mutual information between Z and X.

Putting lemma B.4.1 and B.4.2 together, we get the upper bound in theorem 2.

B.5 χ^2 versus Classic Mutual Information

Features are uniformly distributed over $[0,1]$ and $F(x) = i$ for $x \in [1/(i+1), 1/i))$ and $i > 0$. For each $i > 0$, the sensitive attribute is constant over $[1/(i+1), 1/i))$ and equal to 1 with probability $1/2$.

Form Lemma B.1.2, it is clear that $I_{\chi^2}(X,Z) = \infty$. On the other hand, we can show that the classic mutual information between X and Z, $I_{Sh}(X,Z)$ is bounded. Since F is

deterministic,

$$
\begin{aligned}
I_{Sh}(X,Z) &= \sum_{i=1}^{\infty} \frac{\ln(i(i+1))}{i(i+1)} \\
&\leq \frac{\ln(2)}{2} + \int_1^{\infty} \frac{\ln(x(x+1))}{x^2} dx \\
&\overset{(a)}{=} \frac{\ln(2)}{2} + 1 + \int_1^{\infty} \frac{1}{x(x+1)} dx \\
&\overset{(b)}{\leq} \frac{\ln(2)}{2} + 2 < \infty,
\end{aligned}
\tag{B.23}
$$

where (a) and (b) use integration by part and (b) the fact that $1/x \geq 1/(x+1)$.

B.6 Proof of Theorem 3

We only prove the upper bound on the χ^2 mutual information since the remaining results in Theorem 3 follow directly from Theorem 2.

Since the mapping $(p,q) \to q(p/q-1)^2$ is convex and since Z is an infinite mixtures of Gaussians, we have that for $x \in \mathcal{X}$

$$
\begin{aligned}
&\int \left(\frac{\mu_{t*\sigma}(z|X=x)}{\mu_{t*\sigma}(z)} - 1 \right)^2 \mu_{t*\sigma}(z) dz \\
&\leq \int \int \left(\frac{\mu_{t*\sigma}(z|X=x)}{\mu_{t*\sigma}(z|X=x^{'})} - 1 \right)^2 \mu_{t*\sigma}(z|X=x^{'}) dz \mu(dx^{'}) \\
&, \overset{(a)}{=} \int \chi^2(z|X=x)||z|X=x^{'}) \mu(dx^{'}),
\end{aligned}
\tag{B.24}
$$

where we use Fubini Theorem to invert the summation over z and $x^{'}$ and (a) uses the definition of the χ^2 divergence between $p(z|X=x)$ and $p(z|X=x^{'}$. Since both $p(z|X=x)$ and $p(z|X=x^{'}$ are Gaussians with variance σ^2 and mean $F(x)$ and $F(x^{'})$, respectively, the

integrand in the right hand side of (B.24) can be computed analytically as

$$\chi^2(z|X=x)||z|X=x^{'}) = \\ \frac{1}{2}\left[\exp\left(\frac{||F(x)-F(x^{'})||_2}{\sigma^2}\right)-1\right]. \tag{B.25}$$

Therefore,

$$\begin{aligned} I_{\chi^2}(X,Z) &\leq \frac{1}{2}E_{x,x'}\left[\exp\left(\frac{||F(x)-F(x^{'})||_2^2}{\sigma^2}\right)\right] \\ &\leq \frac{1}{2}\exp\left(\frac{2||F||_\infty^2}{\sigma^2}\right). \end{aligned} \tag{B.26}$$

B.7 Proof of Theorem 4

By [102], we know that the balanced error rate of the optimal auditor f^* is given by

$$\begin{aligned} BER(f^*) &= \frac{1}{2}\int \min(\eta(z,0),\eta(z,1))\mu_{t*\sigma}(dz) \\ &= \frac{1}{4}\int(\eta(z,0)+\eta(z,1))\mu_{t*\sigma}(dz) \\ &- \frac{1}{4}\int|\eta(z,0)-\eta(z,1)|\mu_{t*\sigma}(dz) \\ &\overset{(a)}{=} \frac{1}{2}-\frac{1}{4}\int|\eta(z,0)-\eta(z,1)|\mu_{t*\sigma}(dz), \end{aligned} \tag{B.27}$$

where (a) uses the definition of $\eta(z,s) = P(Z=z|S=s)/P(z)$. Therefore, by Lemma B.1.1,

$$\mathcal{L}_{DP}(\mu_{t,\sigma}) = \frac{1}{2}\int|\mu_{t,\sigma}^0(z)-\mu_{t,\sigma}^1(z)|dz \tag{B.28}$$

and that

$$\mathcal{L}_{DP}(\mu_{n,\sigma}) = \frac{1}{2}\int |\mu^0_{n,\sigma}(z) - \mu^1_{n,\sigma}(z)|dz. \tag{B.29}$$

Therefore, for any F and any features distribution μ over the features $\mathcal{X}$,

$$\begin{aligned}
|\mathcal{L}_{DP}(\mu_{n,\sigma}) - \mathcal{L}_{DP}(\mu_{t,\sigma})| &\overset{(a)}{\leq} \int |(\mu^0_{t,\sigma}(z) - \mu^1_{t,\sigma}(z)) \\
&\quad -(\mu^0_{n,\sigma}(z) - \mu^1_{n,\sigma}(z))|dz \\
&\overset{(b)}{\leq} \int |(\mu^0_{t,\sigma}(z) - \mu^0_{n,\sigma}(z))|dz \\
&\quad + \int |(\mu^1_{t,\sigma}(z) - \mu^1_{n,\sigma}(z))|dz \\
&\overset{(c)}{\leq} \exp\left(\frac{||F||^2_\infty}{\sigma^2}\right)\left(\sqrt{\frac{1}{n_0}} + \sqrt{\frac{1}{n_1}}\right),
\end{aligned} \tag{B.30}$$

where (a) and (b) are consequences of triangular inequalities; and (c) follows from the definition of total variation distance, the upper bound in lemma B.4.2 and theorem 3.

B.8 Monte Carlo Approximation

Lemma B.8.1. *Let $m > 0$ and $n > 0$. Consider a sample $\{(x_i, s_i)\}$ and a noise vector $\{noise_{ji}\}$ of $n \times m$ draws from a d-dimensional Gaussian $\mathcal{N}(0, \sigma I_d)$. Denote $\mu_{n,\sigma}$ the empirical density as in* (B.20) *and for $i = 1, ..., n$ and $j = 1, ..., m$ $z_{ij} = F(x_i) + noise_{ij}$. If*

$$\hat{\mathcal{L}}_{DP}(\mu_{n,\sigma}) = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} |\eta_n(z_{ij}, 1) - \eta_n(z_{ij}, 0)| \tag{B.31}$$

then $\hat{\mathcal{L}}_{DP}(\mu_{n,\sigma})$ *is an unbiased estimator of* $\mathcal{L}_{DP}(\mu_{n,\sigma})$ *and*

$$E_{noise}\left[(\hat{\mathcal{L}}_{DP}(\mu_{n,\sigma}) - \mathcal{L}_{DP}(\mu_{n,\sigma}))^2\right] \leq \frac{8||F||_\infty^2 + 4\sigma^2}{\sigma^2}\frac{1}{nm}. \tag{B.32}$$

Proof. First, $\hat{\mathcal{L}}_{DP}(\mu_{n,\sigma})$ is an unbiased estimator of $\mathcal{L}_{DP}(\mu_{n,\sigma})$ because

$$\begin{aligned} E_{noise}\left[\hat{\mathcal{L}}_{DP}\right] &= \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} E_{noise}[|\eta_n(z_{ij},1) - \eta_n(z_{ij},0)|] \\ &= \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} \mathcal{L}_{DP}(\mu_{n,\sigma}) \\ &= \mathcal{L}_{DP}(\mu_{n,\sigma}). \end{aligned} \tag{B.33}$$

Therefore, the mean squared error can be written as

$$\begin{aligned} &E_{noise}\left[(\hat{\mathcal{L}}_{DP}(\mu_{n,\sigma}) - \mathcal{L}_{DP}(\mu_{n,\sigma}))^2\right] \\ &= \frac{1}{n^2 m}\sum_{i=1}^{n} var_{noise}\left[k(x_i + noise)\right], \end{aligned} \tag{B.34}$$

where $k(z) = |\eta_n(z,1) - \eta_n(z,0)|$. Moreover, by Gaussian Poincare inequality,

$$\begin{aligned} var_{noise}\left[k(x_i + noise)\right] &\overset{(a)}{\leq} \sigma^2 E_{noise}||\nabla k(x_i + noise)||_2^2 \\ &\overset{(b)}{=} 2\sigma^2 \sum_s E_{noise}\left[||\nabla \log(\mu_{n,\sigma}^s(z,s))||_2^2\right] \end{aligned} \tag{B.35}$$

where (a) uses the fact that the noise is Gaussian with standard deviation σ; (b) that $z = x_i + noise$ and that $\nabla\eta_n(z,s) = \eta_n(z,s)\nabla\log(\mu_{n,\sigma}^s(z,s)) + (1-\eta_n(z,s)\nabla\log(\mu_{n,\sigma}^s(z,1-s))$. Moreover, for $s = 0, 1$

$$\begin{aligned}\nabla \log(\mu^s_{n,\sigma}(z,s)) &\overset{(a)}{=} \sum_{i=1}^{n} \nabla \log(\phi(z,x_i)) P(X=x_i|z) \\ &= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(z-F(x_i))P(X=x_i|z),\end{aligned} \tag{B.36}$$

where (a) denotes the Gaussian density with mean $F(x)$ and standard deviation σ as $\phi(z,x)$. Therefore,

$$||\nabla \log(\mu^s_{n,\sigma}(z,s))||_2 \leq \frac{||z||_2 + ||F||_\infty}{\sigma^2}. \tag{B.37}$$

□

Moreover, $z \sim \mu_{n,\sigma}$, which is a mixture of n Gaussians, each with a non-central second moment equal to $\sigma^2 + ||F(x_i)||^2$. Therefore,

$$E_{noise}||z||_2^2 \leq \sigma^2 + ||F||_\infty^2. \tag{B.38}$$

By combining (B.35), (B.36) (B.37) and (B.38), we obtain that

$$var_{noise}\left[k(x_i + noise)\right] \leq 4\frac{2||F||_\infty^2 + \sigma^2}{\sigma^2}, \tag{B.39}$$

and thus that

$$E_{noise}\left[(\hat{\mathcal{L}}_{DP}(\mu_{n,\sigma}) - \mathcal{L}_{DP}(\mu_{n,\sigma}))^2\right] \leq 4\frac{2||F||_\infty^2 + \sigma^2}{\sigma^2 nm}. \tag{B.40}$$

Appendix C: Proofs of Results in Multi-Differential Fairness for Black-Box Classifiers

C.1 Lemma 7.2.1

Proof. Denote $\langle x, x^{'} \rangle$ the average inner product between x and $x^{'}$:

$$\langle x, x^{'} \rangle = E[xx^{'}]. \tag{C.1}$$

Observe that for two variable $U, V \in \{-1, 1\}$, $Pr[V = U] = \frac{1+\langle U,V \rangle}{2}$.

Suppose that there exists $c \in \mathbb{C}_\alpha$ such that for some y and s in $\{-1, 1\}$,

$$LHS \triangleq P[c = 1, Y = y] \left(P[S = s|C = 1, Y = y] - \frac{1}{2} \right) \geq \gamma. \tag{C.2}$$

Without loss of generality, we can assume $s = 1$ (otherwise consider $S^{'} = -S$ instead of S).

$$\begin{aligned} LHS &= P[Y = y, S = 1, c = 1] - \frac{1}{2} P[c = 1, Y = y] \\ &= E\left[\frac{1+c}{2} \frac{yY+1}{2} \frac{S+1}{2} \right] - \frac{1}{2} E\left[\frac{1+c}{2} \frac{yY+1}{2} \right] \\ &= \frac{1}{2} E\left[S \frac{Yy+1}{2} \frac{1+c}{2} \right]. \end{aligned} \tag{C.3}$$

Therefore, the left-hand side in Eq. (7.6) can be written $\frac{1}{2} \left\langle \frac{c+1}{2}, S \frac{1+yY}{2} \right\rangle$.

Moreover, $\langle S, 1 \rangle = \langle S, c \rangle = 2Pr[S = c] - 1 = 0$, since $Pr[S = s|x] = Pr_w[S \neq s|x]$. Lastly,

$$\begin{aligned} \langle yY, S \rangle &= 2E\left[\frac{1+yYS}{2} \right] - 1 \\ & \\ & 2Pr[Y = yS] - 1. \end{aligned} \tag{C.4}$$

Therefore,

$$LHS = \frac{1}{4}Pr[ySY = c] + \frac{1}{4}Pr[Y = yS] - \frac{1}{4} \tag{C.5}$$

The result from lemma 7.2.1 follows since $\rho(y) = Pr[Y = yS]$. □

C.2 Theorem 7.2.2

Proof. $(i) \Rightarrow (ii)$. Denote (x_i, s_i, y_i) a sample from a balanced distribution D over $\mathcal{X} \times \mathcal{S} \times \{-1, 1\}$. Denote $c^* \in \mathbb{C}$ such that $Pr[c^*(x_i) = y_i] = max_{c \in \mathbb{C}} Pr[c(x_i) = y_i] = opt$. Construct a function f such that for (x_i, s_i, y_i), $f(x_i, s_i) = s_i y_i$. Therefore, $f(x_i)s_i = y_i$ and $Pr[c^* = s_i f(x_i)] = Pr[c^* = y_i] = opt$: by lemma 7.2.1, c^* is a γ-unfairness certificate, with $\gamma = \frac{opt+\rho-1}{4}$ and $\rho = Pr[y_i = 1]$. By (i), the certifying algorithm outputs a $(\gamma - \epsilon/4)-$ unfairness certificate $c \in \mathbb{C}$ with probability $1 - \eta$ and $O(\log(|\mathcal{C}, \log(\frac{1}{\eta}), \frac{1}{\epsilon^2})$ sample draws. Hence, by lemma 7.2.1, $Pr[c(x_i) = y_i] = Pr[c(x_i) = f(x_i)y_i] = 4(\gamma - \epsilon/4) + 1 - \rho = opt - \epsilon$, which concludes $(i) \Rightarrow (ii)$

$(ii) \Rightarrow (i)$. Suppose that f is a γ-unfair. Denote $y_i = f(x_i, s_i)$. Samples $\{(x_i, s_i), y_i\}$ are drawn from a balanced distribution over $\mathcal{X} \times \mathcal{S} \times \{-1, 1\}$. By lemma 7.2.1, there exists $c \in \mathbb{C}$ such that $Pr[c(x_i) = s_i y_i] = 4\gamma + 1 - \rho_r$, with $r = \pm$. Assume, without loss of generality $r = +$. Then, $\max_{c'} Pr[c^{'}(x_i) = s_i y_i] \geq 4\gamma + 1 - \rho_+$. By (ii), there exists an algorithm that with probability $1-\eta$ and $O(\log(|\mathcal{C}, \log(\frac{1}{\eta}), \frac{1}{\epsilon^2})$ sample draws outputs $c^{''} \in \mathbb{C}$ such that $Pr[c^{''}(x_i) = s_i y_i] \geq \max_{c'} Pr[c^{'}(x_i) = s_i y_i] - \epsilon/4$. Therefore $Pr[c^{''}(x_i) = s_i y_i] \geq 4(\gamma - \epsilon) + 1 - \rho_+$. By lemma 7.2.1, $c^{''}$ is a $(\gamma - \epsilon)-$ unfairness certificate for f, which concludes $(ii) \Rightarrow (i)$. □

C.3 Lemma 7.3.1

Let $\epsilon > 0$. Let $h \in \mathcal{H}$. For a re-weighting scheme u,

$$\begin{aligned} |R^u(h) - R^w(h)| &= \left| \int_{\mathcal{X}\times\{0,1\}^2} l \circ h \; dP^u - \int_{\mathcal{X}\times\{0,1\}^2} l \circ h \; dP^w \right| \\ &\overset{(a)}{=} \left| \int_{\mathcal{X}\times\{0,1\}^2} l \circ h \; d\pi_1^u - \int_{\mathcal{X}\times\{0,1\}^2} l \circ h \; d\pi_1 \right|, \end{aligned} \tag{C.6}$$

where (a) comes from the fact that $\pi_{-1}^u = 1 = \pi_{-1}$.

Then, if $l \circ h \in \mathcal{G}_k$, then $l \circ h / ||l \circ h||_k \in \mathcal{G}$ and (i) follows from the definition of $IPM_{\mathcal{G}}$.

Suppose now that $l \circ h$ is a bounded continuous function from $\mathcal{X}$ to $\mathbb{R}$. Since k is universal, $\mathcal{G}_k$ is dense in the space of bounded continuous functions and there exists $g \in \mathcal{G}_k$ such that

$$||g_{l,\epsilon} - l \circ h||_\infty \leq \frac{\epsilon}{2\pi_m}, \tag{C.7}$$

where $\pi_m = \max\{E[\pi_1^u], E[\pi_1]\}$. Therefore, using (C.6), we have

$$\begin{aligned} |R^u(h) - R^w(h)| &\leq \left| \int_{\mathcal{X}\times\{0,1\}^2} g \; d\pi_1^u - \int_{\mathcal{X}\times\{0,1\}^2} g \; d\pi_1 \right| \\ &\quad + \left| \int_{\mathcal{X}\times\{0,1\}^2} (g - l \circ h) \; d\pi_1^u - \int_{\mathcal{X}\times\{0,1\}^2} (g - l \circ h) \; d\pi_1 \right| \\ &\overset{(a)}{\leq} \left| \int_{\mathcal{X}\times\{0,1\}^2} g \; d\pi_1^u - \int_{\mathcal{X}\times\{0,1\}^2} g \; d\pi_1 \right| + \epsilon, \end{aligned} \tag{C.8}$$

where (a) uses the definition of g. The result (ii) follows from the fact that $g/||g||_k \in \mathcal{G}$ and from the definition of $IPM_{\mathcal{G}}$.

C.4 Theorem 7.3.2

To prove the result in Theorem 7.3.2, we first need a result from [104] (Corollary 2) to bound the difference between the risk $R^u(h)$ and $R^u_n(h)$ for a re-weighting scheme u and reason about the generalization of . We restate the result as follows:

Theorem C.4.1 ([104])**.** *Consider a sample $\mathcal{D}_n = \{(x_i, s_i, y_i)\}_{i=1}^n$. Assume that $u : \mathcal{X} \times \mathcal{S}$ is a re-weighting function. Assume that $d_{\mathcal{H}} < \infty$ is the pseudo-dimension of $\{l \circ h | h \in \mathcal{H}\}$. Then, for $\delta > 0$, with probability $1 - \delta$, for any $h \in \mathcal{H}$, we have:*

$$|R^u(h) - R^u_n(h)| \leq 2^{5/4} V(u) \left(\frac{d_{\mathcal{H}} \log \frac{2n}{d_{\mathcal{H}}} + \log \frac{8}{\delta}}{n} \right)^{3/8}, \tag{C.9}$$

where

$$V(u) = \max\{\sqrt{E_{\pi_1^u}[u^2(x) l^2(h(x))}, \sqrt{E_{\pi_{n,1}^u}[u^2(x) l^2(h(x))}\}; \tag{C.10}$$

.

We also need a result from [107] to bound the difference between the integral metric probability $IPM_{\mathcal{G}}$ and its empirical counterpart.

Theorem C.4.2. *[107] Consider a sample $\mathcal{D}_n = \{(x_i, s_i, y_i)\}_{i=1}^n$, with $n_s = |\{i | s_i = s\}|$. Assume that $u : \mathcal{X} \times \mathcal{S}$ is a re-weighting function. Assume that $\sup_{g \in \mathcal{G}} |g(x))| \leq \nu$ and that $||k||_\infty < \infty$. Then, for any $\delta > 0$, with probability at least $1 - \delta$*

$$|IPM_{\mathcal{G}}(\pi_1^u, \pi_{-1}) - IPM_{\mathcal{G}}(\pi_{n,1}^u, \pi_{n,-1})| \leq \sqrt{18\nu^2 \log \frac{4}{\delta} ||k||_\infty} \left(\frac{1}{\sqrt{n_1}} + \frac{1}{\sqrt{n_{-1}}} \right), \tag{C.11}$$

The result in Theorem 7.3.2 follows from results C.4.1 and C.4.2 by replacing δ by $\delta/2$

and observing that

$$
\begin{aligned}
|R^w(h) - R^u_n(h)| &\leq |R^w(h) - R^u(h)| + |R^u_n(h) - R^u(h)| \\
&\overset{(a)}{\leq} |R^w(h) - R^u(h)| + ||l \circ h||_k IPM_{\mathcal{G}}(\pi^u_1, \pi_{-1}) \\
&\leq |R^w(h) - R^u(h)| + ||l \circ h||_k IPM_{\mathcal{G}}(\pi^u_{n,1}, \pi_{n,-1}) \\
&+ \|l \circ h||_k |IPM_{\mathcal{G}}(\pi^u_1, \pi_{-1}) - IPM_{\mathcal{G}}(\pi^u_{n,1}, \pi_{n,-1})|.
\end{aligned}
\tag{C.12}
$$

where (a) uses Lemma 7.3.1.

Bibliography

Bibliography

[1] C. K. Sønderby, T. Raiko, L. Maaløe, S. K. Sønderby, and O. Winther, "Ladder variational autoencoders," *Advances in neural information processing systems*, vol. 29, 2016.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[3] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[4] C. Calsamiglia, "Decentralizing equality of opportunity," *International Economic Review*, vol. 50, no. 1, pp. 273–290, 2009.

[5] ProPublica, "How we analyzed the compas recidivism algorithm," *ProPublica*, 2016.

[6] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, ser. Proceedings of Machine Learning Research, S. A. Friedler and C. Wilson, Eds., vol. 81. New York, NY, USA: PMLR, 23–24 Feb 2018, pp. 77–91. [Online]. Available: http://proceedings.mlr.press/v81/buolamwini18a.html

[7] J. Gardner, C. Brooks, and R. Baker, "Evaluating the fairness of predictive student models through slicing analysis," in *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*. ACM, 2019, pp. 225–234.

[8] S. Pfohl, B. Marafino, A. Coulet, F. Rodriguez, L. Palaniappan, and N. H. Shah, "Creating fair models of atherosclerotic cardiovascular disease risk," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 2019, pp. 271–278.

[9] D. Ensign, S. A. Friedler, S. Neville, C. Scheidegger, and S. Venkatasubramanian, "Runaway feedback loops in predictive policing," in *Conference on Fairness, Accountability and Transparency*, 2018, pp. 160–171.

[10] L. vs. State of Wisconsin, *Supreme Court of the State of Wisconsin*, 2016.

[11] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *International Conference on Machine Learning*, 2013, pp. 325–333.

[12] X. Gitiaux and H. Rangwala, "Learning smooth and fair representations," 2020, unpublished.

[13] ——, "Learning smooth and fair representations," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 253–261.

[14] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," 2000.

[15] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel, "The variational fair autoencoder," 2015.

[16] D. Madras, E. Creager, T. Pitassi, and R. Zemel, "Learning adversarially fair and transferable representations," 2018.

[17] H. Edwards and A. Storkey, "Censoring representations with an adversary," 2015.

[18] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," 2016.

[19] B. Dai and D. Wipf, "Diagnosing and enhancing vae models," *arXiv preprint arXiv:1903.05789*, 2019.

[20] X. Gitiaux and H. Rangwala, "Fair representations by compression," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 13, 2021, pp. 11 506–11 515.

[21] E. Creager, D. Madras, J.-H. Jacobsen, M. A. Weis, K. Swersky, T. Pitassi, and R. Zemel, "Flexibly fair representation learning by disentanglement," *arXiv preprint arXiv:1906.02589*, 2019.

[22] A. Vahdat and J. Kautz, "Nvae: A deep hierarchical variational autoencoder," *Advances in Neural Information Processing Systems*, vol. 33, pp. 19 667–19 679, 2020.

[23] R. Child, "Very deep vaes generalize autoregressive models and can outperform them on images," *arXiv preprint arXiv:2011.10650*, 2020.

[24] F. Mentzer, E. Agustsson, M. Tschannen, R. Timofte, and L. V. Gool, "Practical full resolution learned lossless image compression," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10 629–10 638.

[25] K. Gregor, F. Besse, D. Jimenez Rezende, I. Danihelka, and D. Wierstra, "Towards conceptual compression," *Advances In Neural Information Processing Systems*, vol. 29, 2016.

[26] Z. Goldfeld, K. Greenewald, Y. Polyanskiy, and J. Weed, "Convergence of smoothed empirical measures with applications to entropy estimation," *arXiv preprint arXiv:1905.13576*, 2019.

[27] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*. ACM, 2012, pp. 214–226.

[28] M. Kearns, S. Neel, A. Roth, and Z. S. Wu, "An empirical study of rich subgroup fairness for machine learning," *arXiv preprint arXiv:1808.08166*, 2018.

[29] ——, "Preventing fairness gerrymandering: Auditing and learning for subgroup fairness," in *International Conference on Machine Learning*, 2018, pp. 2569–2577.

[30] P. R. Rosenbaum and D. B. Rubin, "Reducing bias in observational studies using subclassification on the propensity score," *Journal of the American statistical Association*, vol. 79, no. 387, pp. 516–524, 1984.

[31] ——, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.

[32] X. Gitiaux and H. Rangwala, "mdfa: Multi-differential fairness auditor for black box classifiers," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, S. Kraus, Ed. ijcai.org, 2019, pp. 5871–5879. [Online]. Available: https://doi.org/10.24963/ijcai.2019/814

[33] J. Vasquez Verdugo, X. Gitiaux, C. Ortega, and H. Rangwala, "Faired: A systematic fairness analysis approach applied in a higher educational context," in *LAK22: 12th International Learning Analytics and Knowledge Conference*, 2022, pp. 271–281.

[34] T. Mashiat, X. Gitiaux, H. Rangwala, P. J. Fowler, and S. Das, "Trade-offs between group fairness metrics in societal resource allocation," *arXiv preprint arXiv:2202.12334*, 2022.

[35] D. S. Massey, J. Rothwell, and T. Domina, "The changing bases of segregation in the united states," *The Annals of the American Academy of Political and Social Science*, vol. 626, no. 1, pp. 74–90, 2009.

[36] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 259–268.

[37] D. Ensign, F. Sorelle, N. Scott, S. Carlos, and V. Suresh, "Decision making with limited feedback," in *Proceedings of Algorithmic Learning Theory*, ser. Proceedings of Machine Learning Research, F. Janoos, M. Mohri, and K. Sridharan, Eds., vol. 83. PMLR, 07–09 Apr 2018, pp. 359–367. [Online]. Available: http://proceedings.mlr.press/v83/ensign18a.html

[38] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," 2016.

[39] P. Besse, E. del Barrio, P. Gordaliza, and J.-M. Loubes, "Confidence intervals for testing disparate impact in fair learning," *arXiv preprint arXiv:1807.06362*, 2018.

[40] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," *arXiv preprint arXiv:1609.05807*, 2016.

[41] U. Hebert-Johnson, M. P. Kim, O. Reingold, and G. N. Rothblum, "Calibration for the (computationally-identifiable) masses," *arXiv preprint arXiv:1711.08513*, 2017.

[42] M. P. Kim, O. Reingold, and G. N. Rothblum, "Fairness through computationally-bounded awareness," *arXiv preprint arXiv:1803.03239*, 2018.

[43] A. Chouldechova and A. Roth, "The frontiers of fairness in machine learning," *arXiv preprint arXiv:1810.08810*, 2018.

[44] Y. Ritov, Y. Sun, and R. Zhao, "On conditional parity as a notion of non-discrimination in machine learning," *arXiv preprint arXiv:1706.08519*, 2017.

[45] J. Rawls, *A Theory of Justice: Revised Edition*. Harvard University Press, 1999.

[46] T. Calders and I. Žliobaitė, "Why unbiased computational processes can lead to discriminative decision procedures," in *Discrimination and privacy in the information society*. Springer, 2013, pp. 43–57.

[47] P. Gordaliza, E. Del Barrio, G. Fabrice, and J.-M. Loubes, "Obtaining fairness using optimal transport theory," in *International Conference on Machine Learning*, 2019, pp. 2357–2365.

[48] F. Calmon, D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, and K. R. Varshney, "Optimized pre-processing for discrimination prevention," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 3992–4001.

[49] Y. Bengio, A. C. Courville, and P. Vincent, "Unsupervised feature learning and deep learning: A review and new perspectives," *CoRR*, vol. abs/1206.5538, 2012. [Online]. Available: http://arxiv.org/abs/1206.5538

[50] A. M. Alaa, M. Weisz, and M. Van Der Schaar, "Deep counterfactual networks with propensity-dropout," *arXiv preprint arXiv:1706.05966*, 2017.

[51] J. Song, P. Kalluri, A. Grover, S. Zhao, and S. Ermon, "Learning controllable fair representations," *arXiv preprint arXiv:1812.04218*, 2018.

[52] A. Jaiswal, R. Brekelmans, D. Moyer, G. V. Steeg, W. AbdAlmageed, and P. Natarajan, "Discovery and separation of features for invariant representation learning," *arXiv preprint arXiv:1912.00646*, 2019.

[53] F. Locatello, G. Abbati, T. Rainforth, S. Bauer, B. Schölkopf, and O. Bachem, "On the fairness of disentangled representations," in *Advances in Neural Information Processing Systems*, 2019, pp. 14 584–14 597.

[54] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *Journal of Machine Learning Research*, vol. 13, no. Mar, pp. 723–773, 2012.

[55] Y. Li, K. Swersky, and R. Zemel, "Learning unbiased features," 2014.

[56] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 335–340.

[57] D. Xu, S. Yuan, L. Zhang, and X. Wu, "Fairgan: Fairness-aware generative adversarial networks," in *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018, pp. 570–575.

[58] K. Kurach, M. Lucic, X. Zhai, M. Michalski, and S. Gelly, "A large-scale study on regularization and normalization in gans," 2018.

[59] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[60] A. Achille and S. Soatto, "Emergence of invariance and disentanglement in deep representations," 2017.

[61] A. Jaiswal, D. Moyer, G. Ver Steeg, W. AbdAlmageed, and P. Natarajan, "Invariant representations through adversarial forgetting." in *AAAI*, 2020, pp. 4272–4279.

[62] A. Jaiswal, Y. Wu, W. AbdAlmageed, and P. Natarajan, "Unsupervised adversarial invariance," 2018.

[63] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[64] E. Agustsson, F. Mentzer, M. Tschannen, L. Cavigelli, R. Timofte, L. Benini, and L. V. Gool, "Soft-to-hard vector quantization for end-to-end learning compressible representations," in *Advances in Neural Information Processing Systems*, 2017, pp. 1141–1151.

[65] F. Mentzer, E. Agustsson, M. Tschannen, R. Timofte, and L. Van Gool, "Conditional probability models for deep image compression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[66] A. v. d. Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," *arXiv preprint arXiv:1601.06759*, 2016.

[67] A. Van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves *et al.*, "Conditional image generation with pixelcnn decoders," in *Advances in neural information processing systems*, 2016, pp. 4790–4798.

[68] L. Theis, W. Shi, A. Cunningham, and F. Huszár, "Lossy image compression with compressive autoencoders," *arXiv preprint arXiv:1703.00395*, 2017.

[69] Y. Choi, M. El-Khamy, and J. Lee, "Variable rate deep image compression with a conditional autoencoder," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3146–3154.

[70] Z. Cui, J. Wang, B. Bai, T. Guo, and Y. Feng, "G-vae: A continuously variable rate deep image compression framework," *arXiv preprint arXiv:2003.02012*, 2020.

[71] V. Kostina and E. Tuncel, "Successive refinement of abstract sources," *IEEE Transactions on Information Theory*, vol. 65, no. 10, pp. 6385–6398, 2019.

[72] Y. Elazar and Y. Goldberg, "Adversarial removal of demographic attributes from text data," *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018. [Online]. Available: http://dx.doi.org/10.18653/v1/d18-1002

[73] H. Gonen and Y. Goldberg, "Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them," 2019.

[74] L. Oneto, M. Donini, A. Maurer, and M. Pontil, "Learning fair and transferable representations," 2019.

[75] C. Dwork, A. Roth *et al.*, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[76] M. Mohri, G. Sivek, and A. T. Suresh, "Agnostic federated learning," *arXiv preprint arXiv:1902.00146*, 2019.

[77] W. Du, D. Xu, X. Wu, and H. Tong, "Fairness-aware agnostic federated learning," *arXiv preprint arXiv:2010.05057*, 2020.

[78] P. P. Liang, T. Liu, L. Ziyin, R. Salakhutdinov, and L.-P. Morency, "Think locally, act globally: Federated learning with local and global representations," *arXiv preprint arXiv:2001.01523*, 2020.

[79] M. Kearns, S. Neel, A. Roth, and Z. S. Wu, "Preventing fairness gerrymandering: Auditing and learning for subgroup fairness," *arXiv preprint arXiv:1711.05144*, 2017.

[80] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi, "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment," in *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2017, pp. 1171–1180.

[81] J. R. Loftus, C. Russell, M. J. Kusner, and R. Silva, "Causal reasoning for algorithmic fairness," *arXiv preprint arXiv:1805.05859*, 2018.

[82] M. Jagielski, M. Kearns, J. Mao, A. Oprea, A. Roth, S. Sharifi-Malvajerdi, and J. Ullman, "Differentially private fair learning," *arXiv preprint arXiv:1812.02696*, 2018.

[83] J. Foulds and S. Pan, "An intersectional definition of fairness," *arXiv preprint arXiv:1807.08362*, 2018.

[84] Y. Mansour, M. Mohri, and A. Rostamizadeh, "Domain adaptation: Learning bounds and algorithms," *arXiv preprint arXiv:0902.3430*, 2009.

[85] F. Johansson, U. Shalit, and D. Sontag, "Learning representations for counterfactual inference," in *International conference on machine learning*, 2016, pp. 3020–3029.

[86] D. T. Braithwaite and W. B. Kleijn, "Bounded information rate variational autoencoders," *arXiv preprint arXiv:1807.07306*, 2018.

[87] H. Berard, G. Gidel, A. Almahairi, P. Vincent, and S. Lacoste-Julien, "A closer look at the optimization landscapes of generative adversarial networks," *arXiv preprint arXiv:1906.04848*, 2019.

[88] D. Lopez-Paz and M. Oquab, "Revisiting classifier two-sample tests," *arXiv preprint arXiv:1610.06545*, 2016.

[89] L. Matthey, I. Higgins, D. Hassabis, and A. Lerchner, "dsprites: Disentanglement testing sprites dataset," https://github.com/deepmind/dsprites-dataset/, 2017.

[90] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[91] U. Gupta, A. Ferber, B. Dilkina, and G. Ver Steeg, "Controllable guarantees for fair outcomes via contrastive information estimation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 9, 2021, pp. 7610–7619.

[92] T. Salimans, A. Karpathy, X. Chen, and D. P. Kingma, "Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications," *arXiv preprint arXiv:1701.05517*, 2017.

[93] A. Jaiswal, D. Moyer, G. V. Steeg, W. AbdAlmageed, and P. Natarajan, "Invariant representations through adversarial forgetting," 2019.

[94] P. C. Roy and V. N. Boddeti, "Mitigating information leakage in image representations: A maximum entropy approach," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2586–2594.

[95] D. Moyer, S. Gao, R. Brekelmans, A. Galstyan, and G. Ver Steeg, "Invariant representations without adversarial training," in *Advances in Neural Information Processing Systems*, 2018, pp. 9084–9093.

[96] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3730–3738.

[97] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[98] A. A. Alemi, B. Poole, I. Fischer, J. V. Dillon, R. A. Saurous, and K. Murphy, "Fixing a broken elbo," *arXiv preprint arXiv:1711.00464*, 2017.

[99] R. Brekelmans, D. Moyer, A. Galstyan, and G. Ver Steeg, "Exact rate-distortion in autoencoders via echo noise," in *Advances in Neural Information Processing Systems*, 2019, pp. 3884–3895.

[100] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[101] L. Devroye, L. Györfi, and G. Lugosi, *A probabilistic theory of pattern recognition*. Springer Science & Business Media, 2013, vol. 31.

[102] M.-J. Zhao, N. Edakunni, A. Pocock, and G. Brown, "Beyond fano's inequality: bounds on the optimal f-score, ber, and cost-sensitive risk and their implications," *Journal of Machine Learning Research*, vol. 14, no. Apr, pp. 1033–1090, 2013.

[103] V. Feldman, V. Guruswami, P. Raghavendra, and Y. Wu, "Agnostic learning of monomials by halfspaces is hard," *SIAM Journal on Computing*, vol. 41, no. 6, pp. 1558–1590, 2012.

[104] C. Cortes, Y. Mansour, and M. Mohri, "Learning bounds for importance weighting," in *Advances in neural information processing systems*, 2010, pp. 442–450.

[105] A. Gretton, A. J. Smola, J. Huang, M. Schmittfull, K. M. Borgwardt, and B. Schöllkopf, "Covariate shift by kernel mean matching," in *Dataset shift in machine learning*. MIT Press, 2009, pp. 131–160.

[106] C. Cortes, M. Mohri, M. Riley, and A. Rostamizadeh, "Sample selection bias correction theory," in *International Conference on Algorithmic Learning Theory*. Springer, 2008, pp. 38–53.

[107] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, G. R. Lanckriet *et al.*, "On the empirical estimation of integral probability metrics," *Electronic Journal of Statistics*, vol. 6, pp. 1550–1599, 2012.

[108] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach, "A reductions approach to fair classification," *arXiv preprint arXiv:1803.02453*, 2018.

[109] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth, "A comparative study of fairness-enhancing interventions in machine learning," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 329–338.

[110] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020.

[111] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, "Fairness and abstraction in sociotechnical systems," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 59–68.

[112] J. Law, W. Bijker, T. P. Hughes, and T. Pinch, "Technology and heterogeneous engineering: The case of portuguese expansion," *The social construction of technological systems: New directions in the sociology and history of technology*, pp. 105–127, 2012.

[113] B. Latour, *Science in action: How to follow scientists and engineers through society*. Harvard university press, 1987.

[114] F. Zhang, K. Kuang, Z. You, T. Shen, J. Xiao, Y. Zhang, C. Wu, Y. Zhuang, and X. Li, "Federated unsupervised representation learning," *arXiv preprint arXiv:2010.08982*, 2020.

[115] H. Heidari, M. Loi, K. P. Gummadi, and A. Krause, "A moral framework for understanding fair ml through economic models of equality of opportunity," in *Proceedings of the Conference on Fairness, Accountability and Transparency*, 2019, pp. 181–190.

[116] S. Hossain, A. Mladenovic, and N. Shah, "Designing fairly fair classifiers via economic fairness notions," in *Proceedings of The Web Conference 2020*, 2020, pp. 1559–1569.

[117] S. Alla and S. K. Adari, "What is mlops?" in *Beginning MLOps with MLFlow*. Springer, 2021, pp. 79–124.

[118] D. Pessach and E. Shmueli, "Algorithmic fairness," *arXiv preprint arXiv:2001.09784*, 2020.

Curriculum Vitae

Xavier Gitiaux graduated from high school at Lycée Marcelin Berthelot, Saint Maur, France in 1999. He received his Bachelor of Science from Ecole Polytechnique, Palaiseau, France in 2005, and his Master of Art from the University of Colorado, Boulder, in 2011. He joined George Mason University as PhD candidate in the Fall 2018. He worked as a visiting scientist at the Massachusetts Institute of Technology Joint Program from 2007 to 2009. He lectured for various courses in Economics at the University of Colorado, Boulder, from 2011 to 2014. He served as a Senior Economist for the Denver Regional Council of Government from 2015 to 2018, as a research scientist for the Frontier Development Lab from June to August 2019 and as a research assistant for the GMU Quantum Science Engineering Center from April to August 2020. He was also employed as a data scientist intern at Microsoft from May to July 2021.