Encoding Fair Representations

THESIS

submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE
in
COMPUTER SCIENCE

Author :	Alexander Latenko
Student ID :	s1427539
Supervisor :	Cor Veenman
2^{nd} corrector :	Peter van der Putten

Leiden, The Netherlands, August 17, 2019

Encoding Fair Representations

Alexander Latenko

Leiden Institute of Advanced Computer Science, Leiden University
Niels Bohrweg 1, 2333 CA Leiden, The Netherlands

August 17, 2019

Abstract

Fair decision making is a difficult problem. Making decision based on certain characteristics about people could be considered unfair, for example, having an individual's race be a factor in the decision of granting a loan. As decision models are becoming more complex, the lack of transparency and our limited understanding of the models make it more difficult to ascertain whether a decision has been made fairly with respect to the relevant subgroups in the population.

One key element for fairness is the data that has been used to arrive at an outcome. Data can be affected by discriminatory practices which, presently, are still an issue globally. If there is much discrimination present in a system, resulting in the system being biased against certain subgroups of the population, then the data collected from the system will contain that same bias. Furthermore, models learning from the data will learn to learn the same bias unless explicitly regulated.

To address this data issue we propose a method for processing the data that removes the sensitive information that enables the discriminatory practices. Just removing an attribute like race will not be enough to ensure that there is no sensitive information left about races in the data. Other attributes or groups of attributes could potentially be proxies for race and would have to be modified as well which makes removing specific information complicated.

We modify the attributes through attribute generalization, which is an anonymization technique used to obscure values by grouping them. Information is lost in this process; our objective is to maximize sensitive information removal and minimize other information losses. The more

information is preserved, the more use the data will have for tasks. We weigh a potential utility of the processed data against the amount of sensitive information remaining to select a suitable trade-off. Attribute generalization is computationally difficult. The technical contribution in this work is to approximate good generalization solutions with neural networks. This contribution includes introducing a derivation, of an existing gradient estimator, for generalization.

Contents

Chapter 1

Introduction

Data collection and data mining has become an important part of a decision making process. The data and decisions made based on the data are sensitive to bias. One alarming source of bias is historical discrimination against certain groups, which can manifest itself in the data and as a result affect the decision making process. This unfairness in the decision process as the result of bias is something we seek to prevent.

The general data protection regulation (GDPR) that has been passed in the EU recognizes the dangers of unregulated data collection and mining [23]. This work touches two important points that are present in the regulation. The first point concerns the usage of **sensitive attributes**, defined as personal data identifying among others a person's gender or ethnic origin. This first point regarding sensitive attributes and their use in fair decision making is the focus of the work. However, we also seek to have a transparent method that may benefit any explanations that need to be made of decision process and we hereby also touch the so-called right to explanation of the GDPR.

One issue that is not addressed in the passed legislation is what requirements have to be fulfilled to prevent unfairness from algorithmic decision making. Multiple interpretations for what fair algorithmic decision-making should entail are possible within GDPR [23]. This is not only limited to the legislation, as the academic literature has produced many differing and even competing notions of fairness, with no general consensus on which ones should be applied [10, 13, 26, 46].

We will be introducing an approach that removes sensitive information, the information about the sensitive attributes from the data. Decision processes based on the data without sensitive information are guaranteed to not discriminate against certain groups because the groups are

no longer distinguishable [29, 40].

Removing sensitive information from the data before processing fits well for data releases. In the context of data releases it is beneficial to keep the data comprehensible even after the removal of sensitive information. Comprehensibility requires that the attributes from the original data are kept. Anyone seeking to use the database, as well as anyone that is subjected to decisions from it, can benefit from increased comprehensibility.

The removal of sensitive information includes obscuring values that are related to the sensitive attributes. These values might be information that could be useful for some task and thus obscuring these values may result in the loss of utility for the task. The contribution in this work is a method that produces data that can guarantee fairness, is comprehensible and also retains as much utility as possible. We refer to the data that has been produced from this method as a fair representation.

1.1 Outline

Chapter 2 will provide background information regarding bias, which includes identification and types of biases, and the various notions of fairness that exist to address (discriminatory) bias. Chapter 3 will explain the meaning of intent in fairness and scope the work with regard to comprehensibility and transparency. Chapter 4 will contain the related work. We will explain the method that has been used to process the data and the conceptual choices that have been made in Chapter 5. We refer to Chapter 6 for the setup of the experiments and the results. The conclusions and possibilities for future work are contained in Chapter 7

Chapter 2

Background

This chapter is split into three parts. The first part shortly summarizes how bias occurs in algorithmic decision makers, either through model implementation choices or data bias. This bias can cause unfair decisions to be made which has led to the conception of notions of fairness. The second part of this chapter will be about these notions of fairness and their limitations. The third part outlines some of the fairness legislation and the context considerations that come with it.

2.1 Bias

We consider bias as systematic differences between the value produced from a model and the ground truth. The main reason why fairness in decision making is a difficult subject is because it is difficult to pinpoint the exact the cause of bias [11, 13]. This challenge of pinpointing the bias increases the importance of fairness notions that allow us to address potential unfairness without knowing the exact bias.

2.1.1 Identifying Bias

It is quite common that we don't know the specifics of the data collection but a decision has to be made nonetheless. To identify bias in the data several tests have been developed [15]. These tests are limited and it remains difficult to guarantee whether or not there is bias in the data [43].

The benchmark test is a commonly-used test to identify the level of bias that may find itself into the data. The benchmark is a expectation

of how the data should be distributed, e.g. the number of drivers of every **protected group** by counting the driver's licenses or looking at census data [15]. With protected group we mean a group of people that have the same sensitive value(s) like race or gender. The test would indicate the level of bias, in for instance traffic searches, by examining if the number of traffic searches for every protected group is in proportion to the number of drivers of that group on the road. However, it is difficult to find unbiased estimators for the true benchmark, for example, instead of driver's licenses we would in reality require the time that each of the protected groups spent on the road.

To avoid the difficulty of finding a representative benchmark for a bias test the outcome test was devised. Outcome tests compare the evaluated groups' rates. In case of traffic stops and searches, an outcome test would evaluate the rate at which different groups would be searched compared to the rate that a search on that group results in identifying criminal activity (e.g. drug possession or drunk driving). The outcome test could indicate discrimination if searches for one group result less often in discovery of criminal activity, possibly indicating that the bar for searching may be too low for the group.

It has been shown that the outcome test may not be able to recognize when there is discrimination occurring [43]. This is due to the difference in risk distributions of groups, we show an example of risk distributions in Figure 2.2. In [43] a threshold test has been proposed to counteract this issue, however, it is not always certain that the threshold test gives a closer reflection of the actual bias than the outcome test.

2.1.2 Bias in COMPAS

COMPAS is a tool that has been used to estimate the risk of recidivism for defendants across the US. A recent study by Propublica assessed, using public records of defendants but not the actual model or data, that COMPAS had different misclassification rates for white and black defendants [3].

The COMPAS tool is an example of where bias is difficult to detect. On the surface the tool seems to be well-calibrated, meaning that the scoring reflects the rate of re-arrest equally well for both the black and white protected groups, as can be seen in figure 2.1. However, calibration does not mean that the tool is bias- or discrimination-free.

The COMPAS tool provides a risk assessment of three groups, we consider the high and low risk ones for simplicity. The model is somewhat

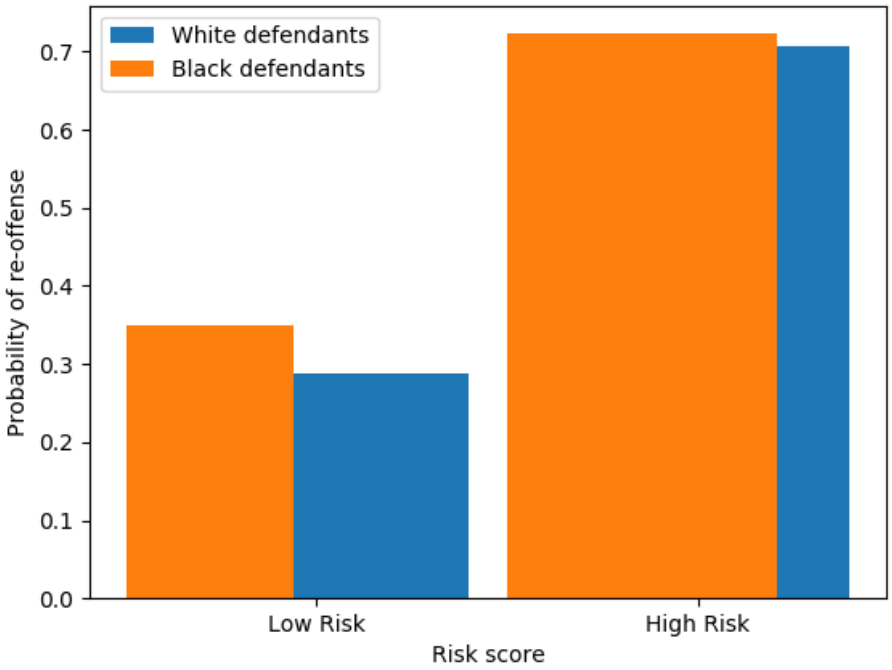


Figure 2.1: *An indication of the probability of recidivism (y-axis) for the high and low risk groups of the COMPAS tool (x-axis). Note that black and white defendants both have similar probability of recidivating per risk group, however, the groups differ significantly in the number of defendants that are considered high or low risk, this is visualized by the width of the bar chart. The probability of recidivating being similar for each score makes this a well-calibrated model regardless of the population distribution.*

calibrated and may even seem as if it is classifying in favour of black defendants in the low category in Figure 2.1. However, only looking at the calibration does not paint the complete picture.

One of the initial reasons for controversy regarding the COMPAS tool was the difference in misclassification rates. It was shown that the false positive rates, on the high risk assessment for individuals that had not recividated, is 44.9% for black defendants and 23.5% of white defendants [3]. This difference in misclassification does not regard the difference in population distributions and has been asserted as insufficient evidence for bias against black defendants in ensuing works [12, 17].

A difference in distributions makes it more difficult to assess whether decision-making is happening fairly. This is also where calibration may

lack as evidence of no bias. Corbett-davies et al. showed that a distribution can be transformed to another calibrated distribution [11]. A simple example would be by taking a distribution of risk, which represents the probability of recidivating of every defendant. For a certain threshold some defendants' probabilities assign them as high risk and some as low risk. However, if we assign score values to defendants, something COMPAS does, we can hypothetically give every defendant the same score. What happens is that every defendant will belong to the same group, the probability of recidivating of the group will be the average of the group. If this average is below the threshold every single individual in the group will be considered low risk.

Corbett-davies et al. visualized a slightly more complex version of such a transformation of scores, we included it in Figure 2.2. Transforming the data this way could for a certain detainment threshold create a large difference proportion of defendants from different protected groups to be detained. For the exact method of transformation we refer the reader to the paper [11].

The COMPAS tool calculates a probability of recidivism of an individual based on the defendant's criminal history and an interview. One relevant source of bias is actually the criminal history and the distinction between re-arrest and re-offense. Re-arrest is a measurable metric and can be subject to bias, e.g. difference in policing of individuals and neighborhoods [21]. Although, re-arrest may be a useful and indicative metric, models using it will suffer from bias that may result in unfairness when considering the actual re-offense rates. Unfortunately the data on actual re-offense rates is not a value that can easily be produced given the existence of unsolved crimes.

2.1.3 Intentional Practices

One reason for fairness is the historical practice of redlining, which consisted of intentionally using proxies of a protected attribute to base a decision on. One controversial case of redlining was using neighborhoods, which were used as proxies for race, as a reason for declining loans which allowed lenders to retain a seemingly non-discriminatory policy [6]. With the amount of data being gathered from various sources, especially in cases of online behaviour, the amount of possible proxies increases and sensitive attributes like gender but also political affiliation are increasingly more available [28, 32]. This is a concern because of possible unintentional disparate impact but also because of for example intentional discrimina-

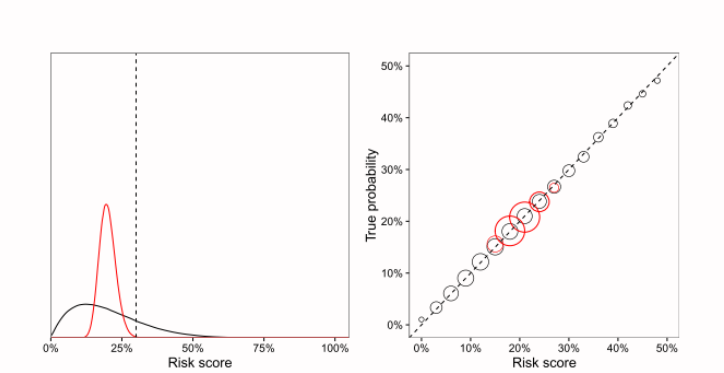


Figure 2.2: *Shows a transformation from black distribution to red distribution while remaining calibrated, effectively averaging the records into a lower score range [11]. The circles in the right image indicate the size of population for both distributions at different risk scores. Note the the distributions are calibrated as the true probability and the risk scores remain the same in the right image. With this transformation the distribution could be placed below the threshold of high risk (vertical line). The risk distributions in COMPAS could be adjusted similarly without breaking calibration.*

tion for purposes like gerrymandering [31]. Dwork et al. [13] provide an overview of practices that exist in discrimination, even though this list is not exhaustive, it summarizes the need for fairness due to intentional discriminatory practices, some of which do throw of the bias tests mentioned above.

2.1.4 Model Biases

Bias is not only limited to data, models can also be affected by bias. In this context we don't mean deliberately discriminatory elements in the model but instead unintentional bias that models suffer from. For instance uncertainty bias, which is when one group is underrepresented in the data and as such, there will be more uncertainty regarding the group. In case of a risk-averse algorithm it is possible for there to be a disparate impact on the low confidence (underrepresented) group. This is closely related to underestimation [30] which is lack of convergence of a model towards the true distribution due to a lack of samples.

2.2 Fairness Notions

Bias is arguably impossible to remove from any real-life complex system. Instead of addressing the bias directly the fairness literature seeks to treat the protected groups equally. The goal is to ensure that the bias, or the decision process in general, does not strongly mistreat certain protected groups over others.

There is no unified notion of fairness in decision making. Generally, decision makers would like to make decisions based solely on attributes that are most useful for accuracy. When attributes are highly related to a protected class, for a myriad of possible reasons, there is a risk of **disparate impact**. Disparate impact being the disproportionate (mis)treatment of certain protected groups. We seek to obscure or remove the sensitive information from the data such that the protected groups are indistinguishable and thereby reducing disparate impact between protected groups.

For the sake of completeness we include the formal definitions of the other notions of fairness as well. For the following definitions, to ease the notation, we have s representing a binary sensitive attribute, y as a binary decision variable and x as the feature vector representing all other attributes, we use the capital variants to refer to all cases s, x and y in the data D.

- *Statistical parity* is the simplest notion of fairness. It basically requires that similar proportions of individuals of every protected group receive a positive decision $y = 1$, for this we want to minimize:

$$1 - \frac{p(y = 1|s = 0)}{p(y = 1|s = 1)}.$$

  We assume w.l.o.g. that $p(y = 1|s = 1) > p(y = 1|s = 0)$. Statistical parity on its own does not consider utility.

- *Calibration* takes into account an evaluation function $f(x)$ that results in some scoring. The proportion of individuals with a positive decisions for both protected groups needs to be the same for all possible scorings $f \in F$:

$$p(y = 1|F = f, s = 0) \approx p(y = 1|F = f, s = 1).$$

  Figure 2.1 shows an example of a fairly calibrated model, where the x-axis represent the possible scorings. It requires similar proportions of positive values per scoring, even though it is not infallible to bias

or optimal in accuracy as has been shown above. It also allows disparate impact to be present because it only looks at the population proportions within a single scoring. The difference in cumulative scorings and population sizes is not considered.

- *Equal classification rates* can have many different variants. We refer to Berk et al. [5] for an overview of some possible notions of fairness that can be taken from the confusion matrix. Additionally we note that notions based on the confusion matrix are very closely related to the model or its output and less so in regard to the data. When considering the model determining the level of fairness versus utility can be done more precisely for that model.

- *Fairness through independence* is the notion of fairness applied in this work and requires the removal of information about the sensitive attributes from the data that is going to be used in the decision process. Removing S from the data, this will only ensure that future decisions $\hat{Y}$, that are solely based on X, will give us $(S \perp\!\!\!\perp \hat{Y})|X$. If we in addition ensure that $S \perp\!\!\!\perp X$, the decisions $\hat{Y}$ will be independent from sensitive attributes due to contraction,

  $$P(S|\hat{Y}, X) = P(S|X) = P(S)$$

  this results in $S \perp\!\!\!\perp \hat{Y}$. This notion will allow us to release data and guarantee a level of statistical parity on it.

Applying only a single fairness notion would regard the degree the protected groups differ in the given measure. For statistical parity one could simply provide equal probabilities for members of different protected groups, this could potentially cause relevant information to be lost and result in mistreatment or even discrimination against individuals that happen to be in the better-performing group. The concept of equal classification rates avoids the problem of mistreatment of better-performing groups by requiring similar accuracy for both groups. Nonetheless, a significant drop in accuracy may be needed to equalize classification rates, which could potentially be considered unfair for the better-performing individuals, regardless of which group they are in. Furthermore, the classification rate equalization does not address the bias present in for example COMPAS, where the metric of re-arrest is known to be a biased version of the target value re-offense.

It is important to combine fairness notions, with either other notions of fairness or with utility functions to prevent discrimination. The specific

tradeoff between fairness and utility is difficult to define and may vary depending on the sample size and or information about known biases in the decision process historically or even presently.

2.3 Fairness Legislation

Fairness has many definitions, none of which could encompass all situations. A selection has to be made what could be considered fair in which context. This is where knowledge from disciplines such as law, economics and ethics should be used to determine the contexts where the fairness is applied and what kind of societal impact it will have. Furthermore, these definitions have to be assimilated into a formal set of rules.

Two important terms in the fairness literature are disparate impact and disparate treatment. Disparate impact, the disproportionate (mis)treatment of protected groups, can occur without it being intentional and is not necessarily illegal. Disparate treatment is about the intentionally treating someone differently because of their sensitive attributes.

Defining regulations to address the notions of disparity is not a simple task. The terms can be in conflict with each other. In the past the 80%-rule has been introduced as to prevent disparate impact in hiring practices [1]. However, it has also been ruled that preventing disparate impact can result in illegal disparate treatment, for example the practice of positive discrimination. Such a case, where actions were taken with the intent to prevent disparate impact, has been brought before Supreme Court of the United States and the actions were deemed as unlawful disparate treatment [2].

The terms disparate impact and disparate treatment come from legal practices and have found an important place specifically in the governing of employment practices and their tension with legitimate metrics of hiring and promoting. One example of tension in promoting practises is shortly after Title VII of the Civil Rights Acts, which prohibited discrimination based on sensitive attributes, promotion based on seniority had a large disparate impact and was used for discrimination [4]. This was due to the recency that black Americans *had* to be hired and thus generally did not possess seniority on the workfloor. Even in this context the discrimination on promotion based on seniority was not considered illegal. Barocas and Selbst examine such practices of US employment laws [4]. On the European side, Hacker published a work regarding the fairness in algorithmic decision making under EU law [25]. In it he discusses the issue of indirect discrimination, the issue of discovering it for learning models and

the shortcoming in the EU anti-discrimination laws to deal with it. Most of it stems from unclarity regarding whether the results of machine learning models can be justified as a reason for disparate impact and how difficult it would be to prove claims against a machine learning model.

Chapter 3

Problem formulation

This work is about processing the data such that we can guarantee a level of fairness. In addition the objective is to keep the data useful. For this purpose we will be introducing the concepts of comprehensibility and truthfulness in this chapter. First we elaborate on the notion of fairness that we use and why it is needed.

3.1 Task-independent Fairness

Our notion of fairness pertains to the removal of sensitive attributes, introduced as fairness through independence in the previous Chapter. The removal of differences between protected groups also removes possible statistical parity between those groups, as the groups will then be indistinguishable.

A reason for this approach is that we do not require knowledge of the tasks that will be run on the data. Instead, we only look at the dependencies between the processed data and the sensitive attributes. With this it will be possible to process the data to guarantee minimum statistical parity in a public release, regardless of what kind of tasks are performed on the data.

One subtle notion in the fairness literature is the concept of intent in discrimination based on sensitive attributes. Using sensitive attributes to improve for example the statistical parity between a better-scoring group A and another group B, will require from the decision-maker to either disproportionately mistreat group A, preferentially treat group B or a combination of both. This is known as positive discrimination and generally considered unlawful (not to be confused with positive action). By using

knowledge of both the task and the sensitive attributes, or strong proxies, to determine a decision a decision-maker shows the intent to discriminate on the sensitive attributes.

In a task-independent pre-processing scenario we avoid positive discrimination. As we can split the decision process into two separate parts, the first part with knowledge of the data, including the sensitive attributes, but not of the task. The second part with knowledge of the task but not of the sensitive attributes. Both parts cannot show the intent to discriminate based on the sensitive attributes. The first part of the process does not contain any information with regard to the tasks to be run on the data and such cannot provide preferential treatment for any group. The second part of the process cannot identify the sensitive attributes and thus cannot use those to treat groups differently.

Task independent fairness is stronger than fairness through unawareness that has been subject of criticism in the fairness literature [26], which is defined as decision-making without using the sensitive attributes. This strength does come with the issue that encoding the protected groups to be completely indistinguishable may require too much processing of the data, causing it to have little utility. The rest of this chapter will discuss restrictions to be placed on the processing to ensure that the data remains useful.

The concept of task-independent fairness is not necessarily binary, some degree of awareness of possible tasks could be used to transform the data. An example of partial awareness is knowing which attributes should be considered as protected attributes and could have discriminatory effects within a domain of tasks. If certain attributes are not considered as potentially discriminatory then the amount information that has to be removed from the data will be smaller. However, with this non-binary definition another question and issue presents itself: which attributes should be considered as sensitive attributes and for which tasks? To properly answer this question reliably awareness of the context of the data release will be required. This work is focused on defining a generic model of task independent fairness and, although this is very important, we will not be looking for which contexts what may lead to discrimination.

3.2 Objective

First let us define the setting we are working in. We assume to know the binary sensitive attribute $s \in S$ and the related attribute vector $x \in X$, with $X \coprod S$ representing the sets covering the data. We are interested in find-

ing a publishable representation of X that protects the sensitive attributes while simultaneously not losing too much (potential) utility. Theoretically we are interested in the tradeoff between $I(X;S)$, mutual information between attributes X and sensitive attributes S, and some utility function. We also assume no knowledge regarding the task, so the utility function will have to be based on some distortion or distance metric in the data. The second objective is to have the published representation be comprehensible and transparent.

3.3 Comprehensibility

When doing a modification to the data we have to take into account the data user, the entity that receives the representation of the data to do analysis on. The data user can be another part of the algorithm or be a physical stakeholder who has done a data request. We want to take into account two important concerns of a data user, namely, the meaning and the trustworthiness of the data.

To address the first concern we want to keep the meaning of the features in the data. If an attribute originally had for example countries as values, these values should not be transformed to something other than countries. Take any individual with some attribute vector x of length n with $x_0, \ldots, x_{n-1}$ being separate attributes. We limit the modification for an attribute i to values from a set W_i such that

$$\forall w_i \in W_i : w_i \in dom(X_i)$$

holds. With $dom(X_i)$ we mean the domain of all possible values for the attribute x_i. The domain of possible values can be defined as the values in the data or we could further extend it with domain knowledge of possible values. We demonstrate a transformation that leads to a comprehensible dataset in Table 3.1, with the European value as an example of use of domain knowledge. We do not consider domain knowledge in this work but for less generic use-cases it may be effective to take into account the hierarchical relations in the data.

The comprehensible transformation for the numeric attributes can be extended, in addition to the use of the set notation above, to also include ranges. We can instead of using a value w_i use a set of ranges R. Where R contains tuples T consisting of two bounding values $t_0 \leq t_1$. Then we could say that a new value or set of values R is comprehensible if and only if

$$\forall T \in R : max(X_i) \geq t_0, t_1 \geq min(X_i).$$

In Table 3.1 we can generalize rating to a set of ranges between 0 and 5. In this scenario we could take the domain knowledge regarding the grading system to set maximum and minimum bounds for generalization. Without domain knowledge the maximum and minimum bounds could be simply set to the smallest and largest values present in the data. This may result in a smaller range of values to be used than there are actually available in reality. In this work we will be assuming no domain knowledge of the features and all possible values we generalize to are limited to within the values present in the data.

Generalizing a single numerical value into multiple ranges may not always make sense. If we take the ratings in Table 3.1 as an example then generalizing a single value, for example 4, to a set of multiple ranges {1-2, 4-5} would make it only more difficult to interpret the data. To prevent this issue we limit the we limit the number of ranges to $|R| = 1$ in this work.

Gender	Nationality	Rating
Male	German	2.2
Female	Dutch	4.9
Male	Bulgarian	4.1

(a) *Original data with gender as sensitive attribute.*

Gender	Nationality	Rating
M/F	European	2.2
M/F	Dutch/Belgian	4-5
M/F	Bulgarian	4-5

(b) *Example of a possible transformation of that data.*

Table 3.1: *Example of a comprehensible and truthful generalization. One possible origin of such data could be a platform of any kind with user reviews. It is not unimaginable that human bias may result in certain genders receiving max rating more often than others. If the difference is slight the disparate impact could be reduced by generalizing the rating scores into a set of ranges.*

3.4 Truthfulness

A common approach for generating a dataset with desired fairness properties is through the addition of noise. However, this addition of noise can generate unwanted relations between data points, which may result in erroneous hypotheses being reinforced by noisy data [19]. For transparency's sake it is important to minimize the degree of noise or false data being put into the data. That is why next to the concept of comprehensibility we also include the concept of truthfulness.

Closely related to comprehensibility, truthfulness requires the values, instead of the domain, to remain the same after transformation. We keep

the original values through generalization. Instead of adding noise the values that are strong proxies for the sensitive attributes are obscured by putting them into a more abstract representation. An example of generalization can be found in Table 3.1, where the nationality is generalized to slightly bigger sets. This attribute generalization example may seem like the addition of noise but one could also imagine a generalization occurring to a known hierarchical element, for instance generalizing a German nationality to a European one. Both are true, however, one is a less informative value that still provides some information.

Truthfulness solely requires that the original set of values V_i for some individual is present in the set of values $V_i \subseteq V_i'$ that is produced by the transformation. When the data is generalized the data user retains certainty that the new representation can be trusted even though it might be less informative.

The advantage of generalization is that it remains truthful and comprehensible while losing a limited amount of information. An issue when transforming to continuous representations of categorical values is that it either does not protect very well or is potentially misleading. If we take as example the Dutch female from Table 3.1 and we can produce a continuous representation that has Dutch as the highest value, which may not obscure the value well. The other option is that the value for the Dutch nationality is lower than another value, for instance Belgian, causing the representation to be misleading, as someone who was Dutch is now likelier to have another nationality in the new representation. Generalization avoids this by simply putting the Dutch person into a larger pool of nationalities.

3.5 Distance

Comprehensibility and truthfulness allow us to address some of the concerns a user may have when deciding to operate based on data from a data release. However, these notions do not take into account the loss of information that may result from a transformation.

Truthfulness requires the original values to still be included in the set of new values and comprehensibility requires the transformation to remain in the same domain. What both of these do not capture is that values remain truthful and in the same domain but generalize to very large sets. We could generalize for every individual all attribute values to $dom(X_i)$, basically completely generalizing all values. This would remain comprehensible and truthful but would not leave any distinguishing information because every record would be exactly the same.

A distance or utility function is required to ensure that some information is kept in the data. We do not know the task, so we require a distance function between the original data and the transformed data that represents the loss in information.

Chapter 4

Related work

The problem we are addressing is an optimization problem in the context of fairness. Several works in the literature have addressed similar challenges within and outside the context of fairness. This chapter describes what those works are and how they relate to this work.

4.1 Bounding Predictability

Transforming the data as little as possible and simultaneously removing as much sensitive information as possible is the conceptual objective in this work. Computing the precise amount of sensitive information that is remaining in the data is a costly operation [33]. Mcnamara et al. [40] showed that the possible statistical parity can be tightly bound by the performance of a Bayes-optimal classifier that seeks to identify the sensitive attributes. Naturally computing a Bayes-optimal classifier is an intractable task but the theorem does allow us to approximately bound the statistical parity by the performance of a classifier that seeks to identify the sensitive attributes. We will be using a well-trained classifier as a measure of the possible discrimination.

4.2 Fairness Models

Many approaches exist with fairness as an objective. The task-independent fairness allows us to work for data release and also prevents disparate treatment as we have defined in Section 3.1. However, it is also possible to use in- and post-processing approaches which allow for more utility to

be kept but require more knowledge of the tasks at hand. This section describes the pre-processing approaches that relate the most to our problem, we refer to [5, 18] for more extensive overviews of the fairness methods.

4.2.1 Adversarial Models

In an adversarial approach two models have a competing objective function. Usually one model produces data and another model evaluates the data [22]. In the fairness literature this is an encoder or generator that produces data and an adversary that evaluates the data. The evaluation in this context is how well the adversary can learn sensitive information from the data. The performance of the adversary can be used as an approximation of the statistical parity as mentioned above. The competing objective is that the encoder trains to minimize the performance of the adversary whereas the adversary does the opposite.

Recently several adversarial approaches have been introduced to address the issue of unfairness [7, 14, 39, 40]. Most of these approaches address some of the objectives defined in our problem statement but none of them consider the comprehensibility of the data. None of these approaches use categorical values (i.e. do attribute generalization). Instead these approaches encode into continuous representation of the original values, which loses the comprehensibility. Madras et al. introduced adversarial training for a fairness with an encoder and decoder [39]. They used the adversary for fairness evaluation and the combination of encoder and decoder to form a distance term. Works that precede that of Madras et al. have used encoders to form fair representations [35, 47], without using an adversary to estimate the fairness. Beutel et al. [5] investigated the effect of skew in sensitive features and target labels on the fairness of a adversarial model but also on the performance and standard deviation of the model. Mcnamara et al. [40] introduced the bound discussed in the previous section but also gives some theoretical properties of the euclidean distance.

4.2.2 Heuristic Models

Aside from using an adversary as an estimation of the degree of fairness other heuristics have been used as well. One straightforward but fairly effective approach is by repairing the distribution of attributes such that they are the same for sensitive features [16]. This can be done optimally for single attributes but taking a multivariate combination of attributes

requires will have to be approximated [29].

Previous work has used attribute generalization to produce data which requires that every unique set of feature values points to sensitive values in about the same proportion as they are found in the data [42]. They show this problem to be NP-hard and use a heuristic algorithms to search for good solutions.

Another approach used convex constraints to find a randomized mapping of the feature values that would result in more fair data [8]. These constraints consist of a general distance constraint between the original data and transformed data, and a discrimination constraint, which requires the sensitive attributes to have similar target label distributions. In addition they also use a individual distortion constraint that required individuals that were similar in the original data to be similar up to a threshold in the transformed data.

Chapter 5

Method

This chapter is about the general model used to encode the data. We propose an adversarial approach that consists of two models that can be trained intermittently, structured as a GAN [22]. The main technical contribution of this work lies in the encoder model used to produce comprehensible representations. We implemented two methods of gradient estimation for categorical values and compare them for this objective. One is a sampling approach [20], whereas the other is our variant of a an approach that works by relaxing the categorical values [27, 37].

We visualize our model in Figure 5.1. Both the encoder and adversary are neural networks. We search for a encoding of the data, for this purpose we devise an objective function consisting of two terms. The two terms represent the utility and fairness of the data.

The convergence speed of the model can be controlled with a temperature parameter τ that is gradually moved towards 0 [27, 37]. With this we have some control regarding thoroughness of the search, with which we can choose between rapid convergence to local minima and high variance but better global convergence [44].

5.1 Objective function

We are interested in finding the minimum mapping of sets for a required level of fairness. We define this as an optimization problem, with X as input representing the original data, we seek a representation Z that contains little information on the sensitive attributes but retains all other information as much as possible. We assign the information left w.r.t. the sensitive attributes as the performance of the adversary trained to minimize the en-

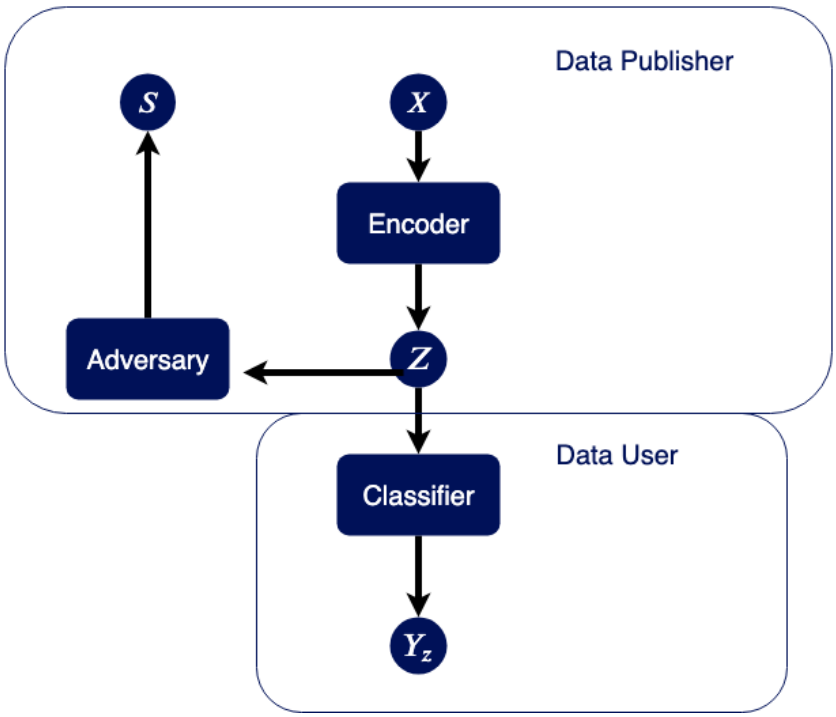


Figure 5.1: *Overview of the model. The data publisher uses the encoder and adversary to search for a representation Z that is close to X and does not contain much information about S. The data user will employ the data to some task producing classification Y. The data publisher does not know the specicifics regarding the data user's process*

tropy between predicted values and the true sensitive values. With the entropy score representing the adversary's uncertainty in predicting the sensitive attributes. We indicate the level of utility as the distance between the input and the representation. This can be described with the loss function:

$$L = \alpha\, d(X, Z) - H(\hat{S}, S) \tag{5.1}$$

With α balancing the trade-off between potential utility and the fairness. The potential utility has been defined as a distance function between the original and transformed data. The fairness bound is defined as H,the cross-entropy between the adversary's prediction of sensitive values $\hat{S}$ and the truth values S. The encoder should be trained using the above objective function, whereas the discriminator tries to optimize its performance on correctly predicting the S.

5.1.1 Distance Objective

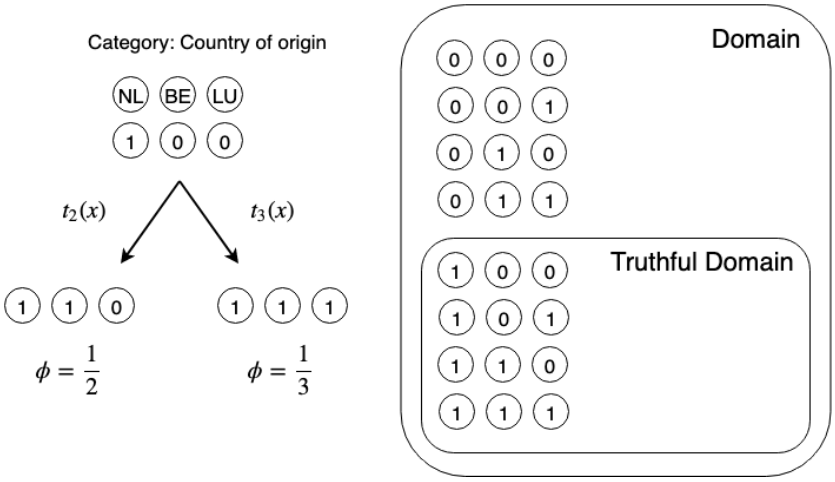


Figure 5.2: *An example of an attribute generalization using a one-hot encoded category. The probabilities below the generalized values represent the probability of the attribute being the original attribute if we assume a uniform distribution. For a comprehensible representation we will have to stay in the domain and for a truthful one in the truthful domain.*

We do not consider specific tasks and as such we do not have utility functions, instead we shall use a distance function between the original data X and the new representation Z. This distance function represents the general information that has been lost through the transformation.

Let's take a vector c of length n that represents the categorical values of an individual in a one-hot encoding, c_i is 1 if the individual has the property of category i otherwise it is 0. Let's assume that i is a property the individual has and thereby $P(c_i = 1|c) = 1$ will be true. When we generalize we are obscuring the value and value i will no longer be the only value that is a 1, we show this in Figure 5.2.

Without including external information, we effectively change from knowing the true value with complete certainty to $P(c_i = 1|z) = \phi$. Where z is the generalized representation of c, meaning it contains more 1's and ϕ is the posterior probability given z for i being the original 1, this value is the uniform probability over all the 1's.

If we are solely implementing a truthful generalization we know that for any increase in the euclidean distance $d(c, z)$ the uniform probability ϕ will decrease, resulting in an increase in the relative entropy or the Kullback–Leibler divergence: $D_{KL}(P(c_i|c)||P(c_i|z)$ [36].

When truthfulness is not regarded the above relation is less useful because the case $P(c_i|z) = 0$ is possible, i.e. the original value was removed in the transformation.

We can still consider the information we lose because we generalize records into larger groups. If we consider any value c_j, that we know is going to be generalized, and an associated value y.

The original conditional probability of $p(y|c_j)$ moves as c_j generalizes to a group of categories $t(c_j)$ with the new group being larger:$|c_j| < |t(c_j)|$. As more values are generalized to a single group the difference:

$$\sum_{j=0}^{n-1} abs\left(p(\hat{y}|t(c_j) - p(y|c_j)\right)$$

grows on average [34]. The distance between the conditional probabilities from the original values and the new values grow as more generalization is applied.

With little knowledge regarding the task the objective is to minimize the addition of new values into the data. For this the euclidean distance suffices as an indication of possible utility. When certain knowledge is known about the data or task beforehand, a more advanced distance function might be preferred.

5.2 Adversarial Encoder

The approach we use to produce a encoding of the data consists of two models. The first model is an encoder that will be transforming the original data. The second model is the adversary which will be evaluating how fair the encoded data can be. Both these models will be trained intermittently.

5.2.1 Adversary

We train the adversary to predict sensitive attributes from the data to estimate how much sensitive information can still be retrieved from the (encoded) data. The adversary objective is to minimize the entropy between its probability predictions and the truth values. We seek to minimize the error for every sensitive group and therefore train the model by (under/over)sampling all sensitive groups equally. The adversary is a neural network with a continuous loss function in the form of the entropy

and therefore we can easily derive gradients on its objective. The gradients are used for training of the adversary with backpropagation. We can backpropagate the adversary's gradient through the encoder as well [22].

5.2.2 Discrete Representation

The objective is to use the encoder for producing a discrete representation and to evaluate it with the adversary. The encoder is a network that takes the input and produces an output containing a sample of categorical values. One challenge in this approach is that backpropagation through the discrete representation is difficult. Generally neural networks are not eligible for such problems. However, recent literature has introduced forms of gradient estimation that allow neural networks to train for discrete values better.

One way of handling discrete values is by adding stochastic elements to the network. The stochastic elements allows for a discrete value to be selected and be passed forward, basically sampling a selection of categories or other discrete values to pass forward. For the backward pass an estimation can be made of the gradient based on the probability and the performance of the sample. There are two types of estimators: biased and unbiased.

5.2.3 Unbiased Estimation

A common concept that is often built upon for unbiased estimation is the likelihood ratio [20] as in the REINFORCE algorithm [45]. Which basically determines the sampling space by likelihood and performance of samples. Let's consider a stochastic network with the objective to minimize the loss: $L(\theta) = \mathbb{E}_{p_\theta(x)}[f(x)]$. Where x is a discrete random variable whose probability is given by $p_\theta(x)$, a continuous differentiable function with respect to θ. The gradient of the objective function is given by:

$$\frac{\delta L}{\delta \theta} = \nabla_\theta \mathbb{E}_{p_{\theta(x)}}[f(x)] = \mathbb{E}_{p_{\theta(x)}}[\nabla_\theta \log p_\theta * f(x)]. \tag{5.2}$$

We know the gradient of the evaluation function and we estimate the gradient of $p_\theta(x)$ by sampling, however, this may result in high variance. Variance reductions techniques have been introduced to combat this variance and allow gradient estimation to achieve more consistent results [24, 41].

5.2.4 Biased Estimation

Biased estimators can be used to avoid the variance problem but come with the issue that they may converge to a worse local minima more easily. A determination has to be made what an appropriate computation time is for convergence.

Recently, the concept of the concrete distribution [37] also published as the Gumbel-Softmax [27] has been introduced. The biased gradient estimation is computed from a continuous relaxation of an discrete activation function. This relaxation is a smooth function that allows for computation of the gradients. Using a temperature parameter and the softmax function one can move between smooth activation and an argmax function:

$$\pi_k = \frac{e^{\frac{x_k}{\tau}}}{\sum_{i=1}^{K} e^{\frac{x_i}{\tau}}}$$

Noise from the $g \sim Gumbel(0, 1)$ distribution can be added to x to sample with softmax probabilities from x [38].

5.3 Approaches encoder

The goal is for the encoder to find a discrete representation. The encoder uses some input IN and weights θ that result in a continuous or partial solution C. We want to search for a discrete solution. There are two common routes for this. Discretize C with an activation function $d()$, which can be a heaveside function which means we only have a gradient at a single point and would require sampling to train the network. Another route is to use a smooth function $s()$ to provide a temporary solution to search with before eventually using $d()$ to provide the final representation. This is the conceptual difference between biased and unbiased estimation for gradient estimation.

5.3.1 Multi-softmax

One approach for working with discrete latent representations is by relaxing them to the continuous space. Instead of using the heaveside activation function we use a smooth activation function, the softmax in our case, to produce a continuous latent representation. We visualize the difference in the two approaches in the first two networks in Figure 5.3.

The first network in Figure 5.3 uses the heaveside function and has no backpropagation error next to it because the gradient only exists at one

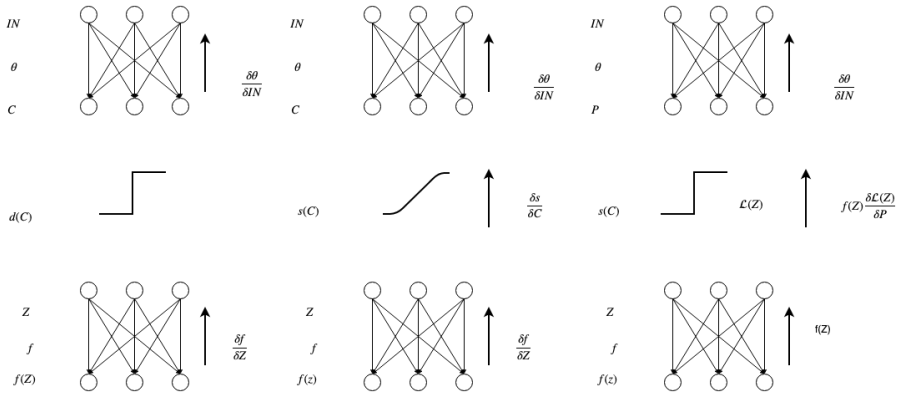


Figure 5.3: *An overview of discrete latent representation propagation. Three methods are shown for backpropagation. The upper networks represents the encoder, whereas the lower network represents the adversary. The first image shows the discretization between the encoder and adversary which does not allow gradients to move smoothly in the backwards pass from the adversary to the encoder. The second image shows the multi-softmax explained in Section 5.3.1. To get an accurate evaluation of Z and the smooth learning with $s(C_1)$ we slowly move $s(C_1)$ towards $d(C_1)$ with the temperature parameter, where $d()$ is the heaveside function shown in the first image. The final image shows how the reinforce method propagates the gradient with discrete samples, explained in Section 5.3.2*

point and is not useful. The second network has a smooth activation function which allows us to compute the gradient of the activation function for the representation: $\frac{\delta s}{\delta C}$.

With the gradient we can use regular backpropagation to optimize the loss function. The evaluation of the continuous variant may not be very telling of how the discrete representation would perform. So in order to evaluate the discrete representation the smooth activation function is slowly moved towards the heaveside function using the temperature parameter during training. With this we can optimize over the continuous space before slowly moving to the discrete space to find the representation to actually evaluate, this has been done similarly in [27, 37].

The gumbel-softmax or concrete distribution has been implemented to work with solely single label latent dimensions [27, 37]. We introduce the multi-softmax to allow for a multi-label problem to be addressed. This is necessary for generalizing from a one-hot encoded representation to a multi-label one like in Figure 5.2.

If we consider the output of an encoder as n neurons with activation values $\{o_i \dots o_n\}$. The multi-softmax activation basically consists of two

activation functions to be applied on the continuous output of the encoder, one function for the k largest activations and one for the $n-k$ other activations. If we sort the activations of the output of the encoder in a decreasing manner then we can define the multi-softmax as follows:

$$\pi_i = \begin{cases} \frac{e^{\frac{o_i}{\tau}}}{\sum_{h=0}^{n} e^{\frac{o_h}{\tau}}} & i = 0 \vee i > k \\ \\ \frac{\beta_i}{\sum_{h=0}^{n} e^{\frac{o_h}{\tau}}} & 0 < i \leq k \end{cases}$$

with

$$\beta_i = \begin{cases} e^{o_i} & i = 0 \\ e^{o_i} + e^{\frac{1}{\tau}}(\beta_{i-1} - e^{o_i}) & i \geq 1 \end{cases}$$

The idea is similar to the idea of the gumbel-softmax where there is a temperature parameter with which we can move from a smooth activation to a step size activation function. The difference between the two is that as the temperature decreases towards 0 the gumbel-softmax moves to a step size function that activates a single neuron in a class. The multi-softmax results in a step size function that activates k neurons, i.e. makes a selection k categorical values. Furthermore, we defined the multi-softmax such that the neurons with the highest activations are most likely to be sampled, as the temperature approaches zero the difference between π_i and π_{i+1} becomes smaller but always with $\pi_i \geq \pi_{i+1}$. We do require the k, the number of values to generalize to, for this definition to work.

We show in Figure 5.4 how such activation looks like at different temperatures. The distance between the output of the multi-softmax of the higher activation π_n and the lower activation π_{n-1} decreases for the k highest values as the temperature moves to zero. There is a limit on the accuracy of the temperature for small numbers so in a practical setting a threshold is used to reach discrete values.

Lastly, the gumbel-softmax lends its name to the addition of gumbel noise that is added to the activation value o, before applying the activation function. This noise combined with the argmax function had the property of sampling with softmax probabilities [38]. This property is lost in the multi-softmax, nonetheless it remains useful to add noise to the activations allow for sampling of the categorical values.

5.3.2 Sampling

The second approach we implemented does not evaluate the function in a relaxation but with the actual discrete values. To learn this well we require

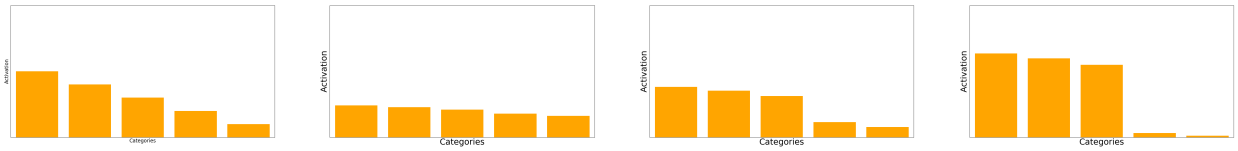


Figure 5.4: *An image of the multi-softmax activation of a sample for different temperature parameters. Note that as the temperature decreases the k highest, in this case 3, move towards 1 and the other values towards 0.*

an estimation of the gradient to propagate to the encoder. For this purpose we will use the REINFORCE method [45]. The encoder produces a continuous output of values from which we will be sampling. The methods takes the k highest values as the active values, but first it adds noise from the uniform distribution $unif(a, b)$ for proportional sampling. We use the log likelihood of the activated sample z given the continuous output c:

$$\mathcal{L}(z|c + unif(a, b)).$$

We compute both the distance loss and the adversary loss on π and backpropagate the gradient computed from Equation 5.2. The third network in Figure 5.3 uses this method, where we still have a heaveside function but we are able to backpropagate due to the computed gradient.

The sampling approach allows us to remain in the discrete space, which is the space we want to end up in. However this generally comes at the cost of much variance, whereas a continuous relaxation should be able to converge more easily given that it has well-defined gradients [39]. On the other hand the expectation is that the sampling approach does not get stuck in local minima as easily. We will be examining the tradeoff in convergence of both approaches in the next chapter.

Chapter 6

Experimental Evaluation

This section will describe the experiments that were conducted to evaluate our method. We split this into several parts. The first part, Section 6.1 elaborates on the settings and the data used. Section 6.2, evaluates the performance of the gradient estimator that we have introduced. In Section 6.3 we seek to evaluate the whether this method can produce satisfactory solutions and if the distance we are using is on average representative for the utility on a task run on the data. There is no clear all-encompassing definition of a satisfactory solution in fairness, as a substitute we will be evaluating what distance-fairness trade-off our model makes. Lastly, in Section 6.4 we will be comparing our method to other state-of-the-art methods for addressing fairness. There are some difference in the objectives of other fairness methods that affect the trade-off performance of the methods.

6.1 Settings

For the experiments we defined the encoder as a network with input nodes being equal to the cardinality of the data put in $|X|$. We used 100 hidden nodes and an output layer with again $|X|$ output nodes. The 100 hidden nodes may have to be updated for datasets with more attributes. The activation function that was used is ReLU. The adversary was defined as a network that took the $|X|$ output of the encoder as the input and produced one sigmoidal output, its prediction of the sensitive value. Again the ReLu activation function was used for the two hidden layers with 100 hidden nodes.

A grid search has been performed for the hyperparameter settings. We summarize the parameter values tested for the experiments in Table 6.1.

Furthermore, we used a parameter k to set number of values we generalize to. This value has been simply set to a static value for all categories but it is possible to use a different k for every category, which would require a more extensive parameter search.

Unless mentioned otherwise, we used the Adult income dataset for the evaluation of our methods*. Which consists of 48842 instances with 14 attributes. We chose the binary attribute gender as sensitive attribute. For task evaluations we used a binary target label that represents whether an income is over 50k. We quantized the age and the years of education attributes in the same manner as in [8].

Evaluation	Parameter	Values
Static & Dynamic		
	lr_e	$\{1e{-}1, 1e{-}2, 1e{-}3, 1e{-}4\}$
	τ_0	$\{0.9, 0.7, 0.6, 0.06\}$
	r	$\{1e{-}3, 1e{-}4, 1e{-}5\}$
Dynamic		
	lr_d	$\{1e{-}1, 1e{-}2, 1e{-}3, 1e{-}4\}$
	d_{runs}	$\{10, 25, 50\}$
	e_{runs}	$\{2, 5, 10, 25\}$

Table 6.1: *The search space of parameters used for the adversarial approach. We did a somewhat wide search to find a well-performing models, for some parameters settings it can quickly be determined that they are not effective. The d and e refer to the discriminator/adversary and encoder respectively. The runs is the number of batches performed on that model between freezes.*

6.2 Static Evaluation

The encoder is trained by putting a moving adversary against it. Both the encoder and adversary are updated for a chosen number of loops. As mentioned in the previous sections there are many possible attribute generalizations, giving the encoder a large search space for every update of the adversary. The encoder has to find solutions in this large search space for every iteration of the adversarial model. One desirable property to have for this process is to quickly find a good solution. This Section will be about evaluating the speed of the implemented approach for a single setting of the adversary.

*Available at https://archive.ics.uci.edu/ml/datasets/adult

6.2.1 Setup

One issue in the evaluation of the adversarial model is that there are two dynamic parts that are constantly updated: the encoder and adversary. The results in the first iteration dictate strongly how the complete adversarial model will behave in the future. It is very much possible that two minima that look to be equally good in the first iteration result in wildly different performance. The variance from both models having to be trained and ambiguity in what the optimal objective is in a dynamic setting makes it difficult to put the performance of the implemented gradient estimation approaches side by side. To evaluate the performance of the gradient estimators we simply freeze the adversary and look at how well the estimators reduce the performance of the adversary.

Given the datasets we first fitted the adversary to the data then froze the training of the adversary. We trained the encoder for 300 epochs on this frozen adversary. It this setting we can easily compare how well the gradient estimation methods are at training the encoder to minimize the adversary performance.

Results

We ran the experiment on the data with categorical values and a k of 7 searching for the best generalization. We visualize the average case with histograms for both figures in 6.1 we can see that the multi-softmax converges quickly more quickly than the REINFORCE gradient estimator, however, it is also more prone to remain in local minima. As more epochs pass the sampling approach finds better minima and performs better on average. This is in accordance to the results from the literature about biased and unbiased estimators [44].

6.3 Dynamic Adversary

The results on the static adversary shows us that an encoder can be trained to reduce performance of one specific adversary. A larger challenge is to reduce the performance for an unknown adversary. We evaluate this by training an independent classifier to predict the sensitive features as well as possible. This independent classifier is trained and cross validated on the produced data, which gives us an indication of how well an adversary could be predict the sensitive values from the processed data. The independent classifier had the same number of layers as nodes as the ad-

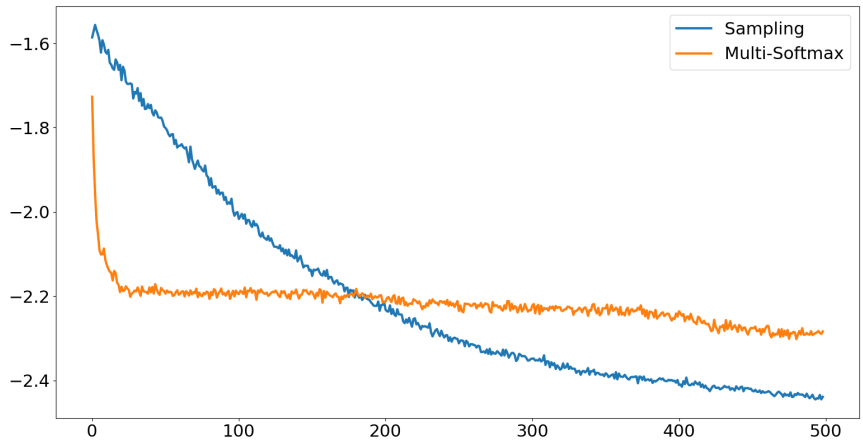


Figure 6.1: *Minimizing adversary performance with both methods. Showing the average over 30 runs for both methods. The y-axis shows the negative entropy score and the x-axis the epochs.*

versary. The adversary needs to be as strong or stronger than the classifier to provide meaningful guarantees. This same classifier was used to compare the fairness of our method to other methods, for both task-dependent as well as task-independent cases.

6.3.1 Alpha Result

We have a parameter α to control the tradeoff between the distance or potential utility and how well the adversary can predict the sensitive attributes. In Figure 6.2 we show that as α increases the distance increases. Furthermore, as the distance increases the performance of the adversary with regard to the sensitive attributes drops. The AUC on the task and the worst-case discrimination, measured as statistical parity (2.2) from the independent classifier predictions, were also reduced as the distance grew. The drop in AUC for task and the worst-case discrimination looks about the same, which coincides to the expectation that as the distance increases the information loss for a task increases on average, described in Section 5.1.1.

The α parameter allows for the choice of trade-off. The α parameter does not work wel for extreme range of values for α. An $\alpha = 1e{-}4$ and $\alpha = 1e4$ would require different learning rates for the networks at a minimum. Careful tuning is required to find an α and corresponding model

that parameterize the trade-off well. In Figure 6.2a we can see small maxima as alpha increases that may have been caused due to the learning rate no longer being a good choice for the specific α value.

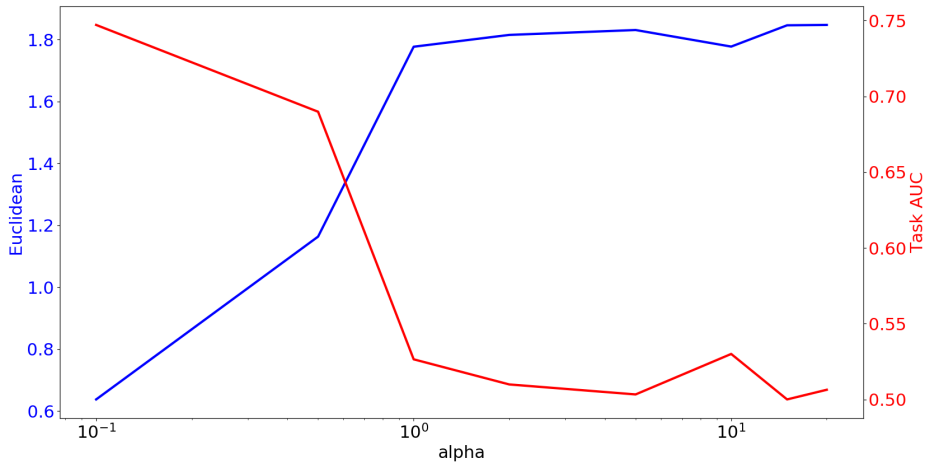
6.4 Comparison

We have shown that we can control the distance, our measure for the potential utility, fairly well in the previous section. We are also interested in how well our method compares to other methods in the fairness literature on the level of discrimination and utility, defined as the AUC on tasks on the data. Methods for fairness are quite varied and comparison is difficult as the objectives and limitations are slightly different for each method. We will be solely evaluating pre-processing methods as our approach falls into that category as well. One issue in comparison is that knowledge of the task, e.g. target labels or the decision model that belong to the task, is commonly assumed. These models perform better on the utility-fairness tradeoff for knowledge with regard to the known tasks. When not assuming knowledge of the task the data produced has to be more generic in its fairness guarantee.

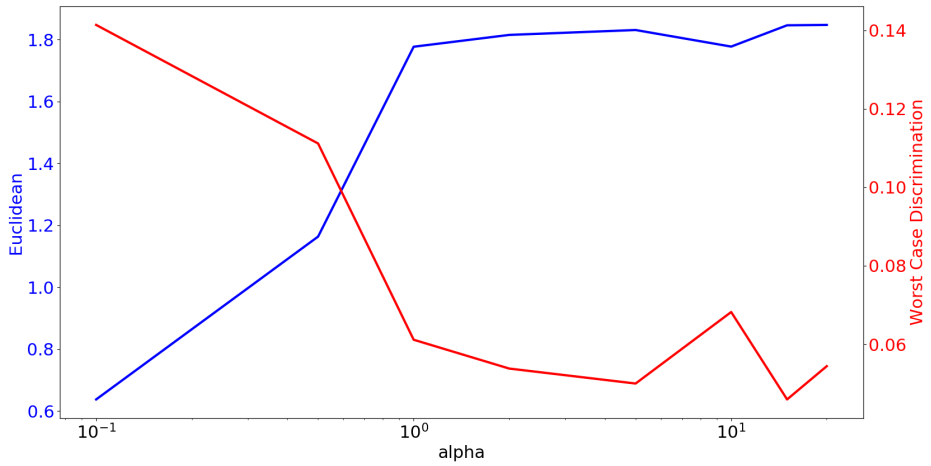
6.4.1 Compared Methods

We compare against the optimized mapping method introduced by Calmon et al. [8]. This method is similar to ours that it seeks to find a mapping for categorical values, however, it does not do attribute generalization, the mappings between categorical values are one-to-one and not one-to-many. Furthermore, it includes the target values in the processing making it task-dependent. We will be comparing this specific method to our generic approach. The method allows for a distortion function to assign weight to changes to certain categories, we set the distortion such that every category would be weighted equally, similarly to how distance is equally weighted for all categories in our method.

The second method we compare to is introduced by Mcnamara et al. [40] which, similar to our approach, uses an adversary independent of the task to train for fairness. We refer to this approach as PFR. PFR does not produce a discrete representation of the data and so does also not perform attribute generalizations. Instead it produces a continuous representation of the data, which results in a loss of comprehensibility of the categorical values. We used the same settings as in the original work [40] and searched over the learning rates that are presented in Table 6.1.



(a) *Trade-off between the AUC for the classifier and the distance introduced by the encoding*



(b) *Trade-off worst case discrimination, as statistical parity, and distance on the task based classifier.*

Figure 6.2: *Showing the trade-off of between the euclidean distance, our measure of potential utility, and the statistical parity that the adversary can identify or that is part of the task. Results are averaged over the adult over 30 repeats. The α parameter can be used to determine an acceptable level of statistical parity and potential utility.*

6.4.2 Evaluation

We plotted the discrimination, task and the distance of the methods in Figure 6.3. We defined discrimination as the most discrimination the independent classifier could cause by predicting the sensitive attributes, this was measured in statistical parity [13].

We show the discrimination, distance and task AUC over the data before a transformation occurs as the grey plots in the Figures 6.3. The optimized mapping method, shown as the green plots, was used to determine the level of utility. This is because the optimized mapping computes the discrimination from the data as a statistical distance between protected groups statically in the data [8]. We minimized the statistical distance, however, the discrimination for the independent classifier was at least the levels shown in the Figures in 6.3.

PFR allows for the categorical values to be transformed to a continuous representation. This may lose on some comprehensibility of the data but allows for a better tradeoff for the utility and fairness as presented in the Figures in 6.3. PFR produces the best fairness utility result but this came at the cost of not producing data with categorical values.

Our approach performs slightly worse than PFR, with the added benefit that the processed data retains the categorical values. Our approach with $k = 1$, which does not generalize but instead remaps values one-to-one similarly to the optimized mapping method, performs slightly better than our approach with $k = 2$ on the utility-discrimination plot. Our approach with $k = 2$ performs far worse on the distance-discrimination plot. This indicates that on average it will retain less information which will also cause worse performance for task AUC. Doing an attribute generalization with larger k will result in more distance but allows the truthfulness to be more easily achieved because a larger group of values can be used to obscure the original value. When comparing to the optimized mapping for the Adult data our remapping approach performs better on distance as can be seen in 6.3b.

Lastly, we can see in Figure 6.3c that the variance of the results of our approach is fairly large for the Compas data. This is to be expected given that we are working with two dynamic models, which may cause a large range of results to be produced over processing runs. The variance is a lot less pronounced for the Adult data compared to other methods. The reduction of variance in these GAN-like is being addressed in current research [9] and may need to be applied for certain datasets.

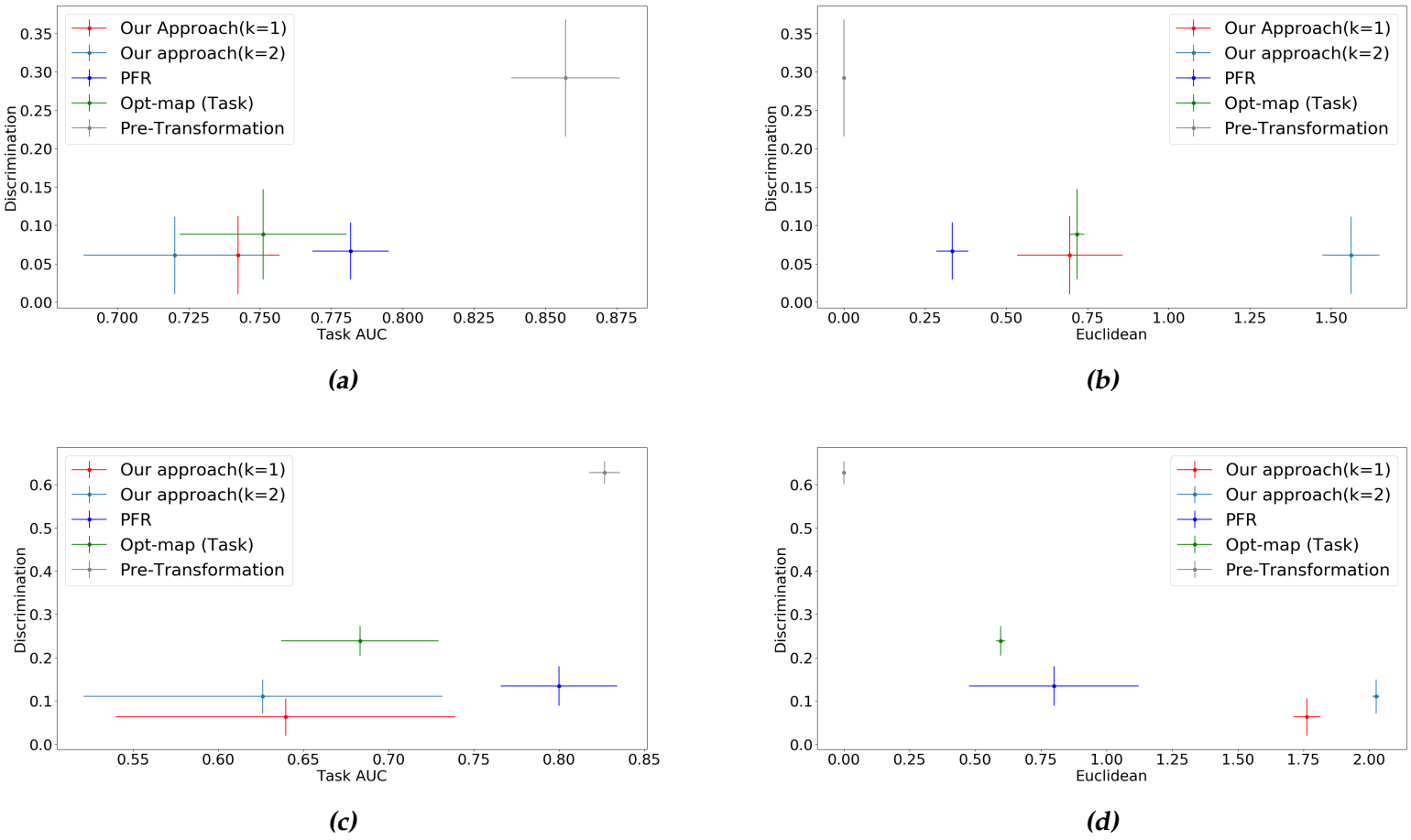


Figure 6.3: *We plot the performance of the methods for several parameters settings. Figure 6.3a visualizes the AUC of the classifier on the task that the methods achieved at specific levels of discrimination, defined as the statistical parity. Figure 6.3b visualizes the distance they moved from the original data against the discrimination and for the same AUC values as Figure 6.3a. The dots represent the mean computed from 30 runs and the lines represent the standard deviation. Top row is for the Adult dataset and the bottom row is for the Compas dataset.*

Chapter 7

Conclusions

The fairness literature addresses a fundamental issue in machine learning: the bias in the data. Data always presents a limited view of reality and it is therefore difficult to discover what the bias and its effects are. We know that certain groups of people are sensitive to discrimination and have been actively discriminated. We can prevent this discrimination from occurring to a degree by removing information that distinguishes these groups from the data.

We implemented a method that allows us to process the data for task-independent fairness. Our method seeks to minimize the information regarding sensitive attributes while also keeping as much other information in the data as possible. Doing this while keeping the values in the original domain is not an easy task. We used a GAN-like method for this approach. What separates our approach from all other GAN methods in fairness is that we keep the representation comprehensible instead of allowing continuous representations, which can either be potentially misleading or remove too little sensitive information. To provide a comprehensible categorical representation we introduced a gradient estimator that can more consistently find good results with regard to sensitive information removal.

We examined the effect of our tradeoff parameter and found that we can control fairly well the desired level of utility and fairness. We compared our method to other approaches in the literature. We found the results of our method and methods present in the literature to be comparable. Furthermore, the differences in performance that do exist all fall within reason due to slight differences in the objectives of the methods, like whether or not comprehensibility or truthfulness were regarded.

7.1 Future Work

Most of the fairness literature use binary sensitive values to evaluate the models, this includes our work as well. The simplicity of the binary case makes it easier to perform generic processing of the data. However, in real-life situations the number of sensitive groups and values are numerous. For fairness to become more widely practiced it will be necessary to allow for multivariate combinations of non-binary sensitive values to be addressed. It remains to be seen whether this will be possible to do generically or would require extensive knowledge of the task and the risk of discrimination for each group.

Our approach can be extended with regard to the number of values to generalize to, we used a static k for all categories, which was able to achieve comparable results as other methods. Generalization does not have to be limited by a single size. Certain categories or even unique records may need to be generalized more strongly than others. We mention in Chapter 5 that is is possible to do with offline tuning, if choosing the size per category at least. Setting the size per individual record forms a very large search space. Investigation into whether this is possible to make part of the network and thereby choosing k in a dynamic manner could potentially greatly improve the performance of our method.

We made our method specific for tabular datasets commonly used in the fairness literature. We used neural networks for this purpose. It would be interesting to see if it is possible to apply this method to problems in other types of data where neural networks are known to perform fairly well, for example fairness in vision-based tasks.

Bibliography

[1] Eeoc, the u.s. uniform guidelines on employee selection procedures. `https://www.eeoc.gov/policy/docs/qanda_clarify_procedures.html`, march 1979.

[2] Supreme court of the united states. ricci v. destefano. 557 u.s. 557, 174, 2009.

[3] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. propublica, may 23, 2016, 2016.

[4] S. Barocas and A. D. Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

[5] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth. Fairness in criminal justice risk assessments: the state of the art. *arXiv preprint arXiv:1703.09207*, 2017.

[6] J. A. Berkovec, G. B. Canner, S. A. Gabriel, and T. H. Hannan. Race, redlining, and residential mortgage loan performance. *The Journal of Real Estate Finance and Economics*, 9(3):263–294, 1994.

[7] A. Beutel, J. Chen, Z. Zhao, and E. H. Chi. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*, 2017.

[8] F. Calmon, D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, and K. R. Varshney. Optimized pre-processing for discrimination prevention. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3992–4001. Curran Associates, Inc., 2017.

[9] T. Chavdarova, G. Gidel, F. Fleuret, and S. Lacoste-Julien. Reducing noise in gan training with variance reduced extragradient. *arXiv preprint arXiv:1904.08598*, 2019.

[10] A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[11] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806. ACM, 2017.

[12] W. Dieterich, C. Mendoza, and T. Brennan. Compas risk scales: Demonstrating accuracy equity and predictive parity. 2016.

[13] 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*, pages 214–226. ACM, 2012.

[14] H. Edwards and A. Storkey. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*, 2015.

[15] R. S. Engel and J. M. Calnon. Comparing benchmark methodologies for police-citizen contacts: Traffic stop data collection for the pennsylvania state police. *Police Quarterly*, 7(1):97–125, 2004.

[16] 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*, pages 259–268. ACM, 2015.

[17] A. W. Flores, K. Bechtel, and C. T. Lowenkamp. False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *Fed. Probation*, 80:38, 2016.

[18] 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*, pages 329–338. ACM, 2019.

[19] B. C. Fung, K. Wang, A. W.-C. Fu, and S. Y. Philip. *Introduction to privacy-preserving data publishing: Concepts and techniques*. Chapman and Hall/CRC, 2010.

[20] P. W. Glynn. Likelihood ratio gradient estimation for stochastic systems. *Communications of the ACM*, 33(10):75–84, 1990.

[21] S. Goel, J. M. Rao, R. Shroff, et al. Precinct or prejudice? understanding racial disparities in new york city's stop-and-frisk policy. *The Annals of Applied Statistics*, 10(1):365–394, 2016.

[22] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[23] B. Goodman and S. Flaxman. European union regulations on algorithmic decision-making and a" right to explanation". *arXiv preprint arXiv:1606.08813*, 2016.

[24] S. Gu, S. Levine, I. Sutskever, and A. Mnih. Muprop: Unbiased backpropagation for stochastic neural networks. *arXiv preprint arXiv:1511.05176*, 2015.

[25] P. Hacker. Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under eu law. *Common Market Law Review*, 55(4):1143–1185, 2018.

[26] M. Hardt, E. Price, N. Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[27] E. Jang, S. Gu, and B. Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[28] C. Jernigan and B. F. Mistree. Gaydar: Facebook friendships expose sexual orientation. *First Monday*, 14(10), 2009.

[29] J. E. Johndrow and K. Lum. An algorithm for removing sensitive information: application to race-independent recidivism prediction. *arXiv preprint arXiv:1703.04957*, 2017.

[30] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 35–50. Springer, 2012.

[31] 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.

[32] M. Kosinski, D. Stillwell, and T. Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, page 201218772, 2013.

[33] A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Physical review E*, 69(6):066138, 2004.

[34] J. D. Lawson and Y. Lim. The geometric mean, matrices, metrics, and more. *The American Mathematical Monthly*, 108(9):797–812, 2001.

[35] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

[36] D. J. MacKay and D. J. Mac Kay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[37] C. J. Maddison, A. Mnih, and Y. W. Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

[38] C. J. Maddison, D. Tarlow, and T. Minka. A* sampling. In *Advances in Neural Information Processing Systems*, pages 3086–3094, 2014.

[39] D. Madras, E. Creager, T. Pitassi, and R. Zemel. Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309*, 2018.

[40] D. McNamara, C. S. Ong, and R. C. Williamson. Provably fair representations. *arXiv preprint arXiv:1710.04394*, 2017.

[41] A. Mnih and K. Gregor. Neural variational inference and learning in belief networks. *arXiv preprint arXiv:1402.0030*, 2014.

[42] S. Ruggieri. Using t-closeness anonymity to control for non-discrimination. *Trans. Data Privacy*, 7(2):99–129, 2014.

[43] C. Simoiu, S. Corbett-Davies, S. Goel, et al. The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics*, 11(3):1193–1216, 2017.

[44] G. Tucker, A. Mnih, C. J. Maddison, J. Lawson, and J. Sohl-Dickstein. Rebar: Low-variance, unbiased gradient estimates for discrete latent variable models. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 2627–2636. Curran Associates, Inc., 2017.

[45] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

[46] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi. Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259*, 2015.

[47] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013.