Adversarial mitigation to reduce unwanted biases in machine learning

Vincent Grari



Sorbonne Université

École doctorale Informatique, Télécommunications et Électronique (Paris)

Équipe MLIA, ISIR

Adversarial mitigation to reduce unwanted biases in machine learning

Par Vincent Grari

Thèse de doctorat d'Informatique

Dirigée par Sylvain Lamprier et Marcin Detyniecki

Présentée et soutenue publiquement le 22 juin 2022

Devant un jury composé de :

Nicolas Usunier	Facebook, Paris	Rapporteur
Patrick Loiseau	Inria Grenoble Rhône-Alpes, POLARIS team, LIG	Rapporteur
Catuscia Palamidessi	Lix, Inria, Institute Polytechnique de Paris	Examinateur
Arthur Charpentier	Université du Québec à Montréal (UQAM), Canada	Examinateur
Christophe Marsala	LIP6, Sorbonne Université	Examinateur
Sylvain Lamprier	ISIR, Sorbonne Université	Directeur de thèse
Marcin Detyniecki	AXA, Paris	Directeur de thèse

Adversarial mitigation to reduce unwanted biases in machine learning

Résumé

Ces dernières années, on a assisté à une augmentation spectaculaire de l'intérêt académique et sociétal pour l'apprentissage automatique équitable. En conséquence, des travaux significatifs ont été réalisés pour inclure des contraintes d'équité dans les algorithmes d'apprentissage automatique. Le but principal est de s'assurer que les prédictions des modèles ne dépendent d'aucun attribut sensible comme le genre ou l'origine d'une personne par exemple. Bien que cette notion d'indépendance soit incontestable dans un contexte général (Dwork et al., 2012), elle peut théoriquement être définie de manière totalement différente selon la façon dont on voit l'équité. Par conséquent, de nombreux articles récents abordent ce défi en utilisant leurs "propres" objectifs et notions d'équité. Les objectifs sont catégorisés en deux familles différentes : L'équité individuelle et l'équité de groupe. Cette thèse donne d'une part, une vue d'ensemble des méthodologies appliquées dans ces différentes familles afin d'encourager les bonnes pratiques. Ensuite, nous identifions et complétons les lacunes en présentant de nouvelles métriques et des algorithmes de machine learning équitables qui sont plus appropriés pour des contextes spécifiques.

La méthode d'équité de groupe impose l'égalité sur des distributions et s'oppose donc à l'équité individuelle qui impose l'égalité à un niveau local ou le plus individuel possible. De nombreuses approches ont été employées ces dernières années pour atteindre cet objectif. Nous distinguons deux familles d'atténuation que nous avons nommées, premièrement, *méthodes de retraitement de prédiction* où le modèle de prédiction encourage l'atténuation du biais sur la prédiction elle-même ; et deuxièmement, *méthodes de représentation équitable* où un adversaire atténue le biais sur une représentation intermédiaire latente. Dans le cadre de la première famille qui s'interesse à la sortie des prédictions, nous avons proposé deux nouveaux algorithmes: Le premier basé sur du boosting d'arbes de décision à montré des résultats très compétitifs sur des ensembles de données réels tabulaires. Le second, basé sur le coefficient de corrélation maximal de Hirschfeld-Gebelein-Rényi (HGR) permet l'application de technique d'atténuation de biais pour le cas continue. Des articles récents ont montré que les approches par représentation adverse équitable peuvent donner de meilleurs résultats en termes de précision de prédiction tout en restant équitable dans des scénarios complexes de monde réel. Nous avons étudié les raisons qui peuvent induire cette surperformance et nous avons proposé une extension de notre algorithme basé sur le HGR pour la représentation équitable.

D'autre part, un autre sous-domaine de l'apprentissage automatique équitable est l'équité individuelle. Le concept d'équité individuelle peut être résumé comme suit : "Les personnes similaires doivent être traitées de manière similaire", la notion de similarité étant codée par une distance spécifique sur l'espace d'entrée. Nous avons proposé de nouvelles métriques pour évaluer le niveau

de biais au sens individuel, ainsi que deux nouveaux algorithmes. Le premier s'appuie sur de l'auto-encodage variationnel (VAE) pour garantir un traitement similaire aux personnes similaires, et le second par *intervention*, où nous nous appuyons sur un graphe causal spécifique pour générer des observations contrefactuelles pour chaque individu (Grari et al., 2021a).

Nous considérons ensuite des scénarios pratiques du monde réel. Tout d'abord, nous examinons un contexte où la variable sensible n'est pas présente dans l'ensemble d'entrainement. Nous avons pour cela proposé un modèle d'inférence bayésien basé sur un graphe de connaissances causales afin de récupérer les informations sensibles et d'atténuer ce proxy biaisé dans le prédicteur final. Enfin, nous nous intéressons également à l'atténuation des biais du modèle dans une application operationelle, en particulier dans la construction d'un modèle de tarification d'assurance débiaisé. À cette fin, nous avons créé un cadre général dans lequel un seul modèle de tarification complet est entraîné en générant les composantes de tarification géographiques et automobile nécessaires pour prédire la prime pure tout en atténuant le biais indésirable.

Abstract

The past few years have seen a dramatic rise of academic and societal interest in *fair machine learning*. As a result, significant work has been done to include fairness constraints in the training objective of machine learning algorithms. Its primary purpose is to ensure that model predictions do not depend on any sensitive attribute as gender or race, for example. Although this notion of independence is incontestable in a general context (Dwork et al., 2012), it can theoretically be defined in many different ways depending on how one sees fairness. As a result, many recent papers tackle this challenge by using their "own" objectives and notions of fairness. Objectives can be categorized in two different families: Individual and Group fairness. This thesis gives an overview of the methodologies applied in these different families in order to encourage good practices. Then, we identify and complete gaps by presenting new metrics and new Fair-ML algorithms that are more appropriate for specific contexts.

The group fairness method enforces equality over the general distributions and is therefore opposed to individual fairness for enforcing equality at a local level. Many approaches have been employed these recent years to achieve this objective. We distinguish two mitigation families that we have named, first, *prediction retreatment methods* where a prediction model encourages bias mitigation on the output prediction; and second, *fair representation methods* where an adversarial mitigates the bias on a latent intermediary representation. In the first family, which focuses on output predictions, we proposed two new algorithms: One based on decision tree boosting has shown very competitive results on real tabular data sets. The second one, based on the Hirschfeld-Gebelein-Rényi (HGR) maximum correlation coefficient, allows bias mitigation techniques for the continuous case. Recent papers have shown that the fair adversarial representation can give better results in prediction accuracy while remaining fair in complex real-world scenarios (Adel et al., 2019). We investigated empirically why these methods can give better results and we proposed an extension of our algorithm based on the HGR for fair representation.

On the other hand, another sub-field of fair machine learning is individual fairness. The concept can be summarized as follows: "similar people should be treated similarly", the similarity being encoded in a specific distance between individuals. We have proposed some new metrics to assess the level of bias in the individual sense, along with two new algorithms. The first one relies on Variational Autoencoding (VAE) to ensure similar treatment to similar people, the second one leveraging *intervention*, where we rely on a specific causal graph for generating some counterfactual observations for each individual (Grari et al., 2021a).

We then consider practical, real-world scenarios. First, we examine a context where the sensitive variable cannot be present in the training set. For this spe-

cific purpose, based on a causal knowledge graph, we have created a Bayesian inference model to recover the sensitive information and subsequently mitigate this biased proxy in the final predictor model. Finally, we are also interested in mitigating the biases model in a real-world application, particularly in constructing a debiased insurance pricing model. For this purpose, we created a general framework in which a single whole pricing model is trained by generating geographic and car pricing components needed to predict the pure premium while mitigating the unwanted bias according to the desired metric.

Publications

The work conducted during the Ph.D program has led to the following publications:

Mentioned in this thesis

Journals:

Vincent Grari, Boris Ruf, Sylvain Lamprier, and Marcin Detyniecki. Achieving fairness with decision trees: An adversarial approach. *Data Sci. Eng.*, 5(2):99–110, 2020c. doi: 10.1007/s41019-020-00124-2

Accepted to Machine Learning Journal with minor modifications (send to Editor the 13 April 2022):
Vincent Grari, Sylvain Lamprier, and Marcin Detyniecki. Adversarial learning for counterfactual fairness. *arXiv preprint arXiv:2008.13122*, 2020b

Conferences:

Vincent Grari, Sylvain Lamprier, and Marcin Detyniecki. Fairness-aware neural rényi minimization for continuous features. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 2262–2268. ijcai.org, 2020a. doi: 10.24963/ijcai.2020/313. URL https://doi.org/10.24963/ijcai.2020/313

Vincent Grari, Sylvain Lamprier, and Marcin Detyniecki. Fairness without the sensitive attribute via causal variational autoencoder. *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, (still waiting DOI)*, 2109.04999, 2021c

Vincent Grari, Boris Ruf, Sylvain Lamprier, and Marcin Detyniecki. Fair adversarial gradient tree boosting. In Jianyong Wang, Kyuseok Shim, and Xindong Wu, editors, *2019 IEEE International Conference on Data Mining, ICDM 2019, Beijing, China, November 8-11, 2019*, pages 1060–1065. IEEE, 2019. doi: 10.1109/ICDM.2019.00124. URL https://doi.org/10.1109/ICDM.2019.00124

Vincent Grari, Oualid El Hajouji, Sylvain Lamprier, and Marcin Detyniecki. Learning unbiased representations via rényi minimization. In Nuria Oliver, Fernando Pérez-Cruz, Stefan Kramer, Jesse Read, and José Antonio Lozano, editors, *Machine Learning and Knowledge Discovery in Databases. Research Track - European Conference, ECML PKDD 2021, Bilbao, Spain, September 13-17, 2021, Proceedings, Part II*, volume 12976 of *Lecture Notes in Computer Science*, pages 749–764. Springer, 2021b. doi: 10.1007/978-3-030-86520-7_46. URL https://doi.org/10.

1007/978-3-030-86520-7_46

Vincent Grari, Oualid El Hajouji, Sylvain Lamprier, and Marcin Detyniecki. Enforcing individual fairness via rényi variational inference. In *International Conference on Neural Information Processing*, pages 608–616. Springer, 2021a. doi: 0.1007/978-3-030-92307-5_71. URL https://doi.org/10.1007/978-3-030-92307-5_71

Preprints:

The following paper has been submitted to the ASTIN Journal but at the time of sending this thesis, we were still waiting for the response:

Vincent Grari, Arthur Charpentier, Sylvain Lamprier, and Marcin Detyniecki. A fair pricing model via adversarial learning. *arXiv:2202.12008*, 2022

Other works (joint collaborations)

The following works are not mentioned directly in this manuscrit, but have been conducted in parallel as collaborations on related topics:

Alexandre Corradin, Michel Denuit, Marcin Detyniecki, Vincent Grari, Matteo Sammarco, and Julien Trufin. Joint modeling of claim frequencies and behavioral signals in motor insurance. *ASTIN Bulletin: The Journal of the IAA*, pages 1–22, 2022

Contents

Acronyms

CART	Classification And Regression Tree
CNN	Convolutional Neural Network
ELBO	Evidence Lower Bound
EM	Expectation–Maximization algorithm
Fair-ML	Fair Machine Learning
FR	Fair Representation
FTA	Fairness Through Awareness
GAM	Generalized Additive Model
GAN	Generative Adversarial Network
GDPR	General Data Protection Regulation
GLM	Generalized Linear Model
GTB	Gradient Tree Boosting
HGR	Hirschfeld-Gebelein-Rényi
KDE	Kernel Density Estimation
KL	Kullback-Leibler Divergence
KNN	K-Nearest Neighbors
MI	Mutual Information
MLP	Multilayer Perceptron
MMD	Maximum Mean Discrepancy
MSE	Mean Squared Error
PR	Prediction Retreatment
RDC	Randomized Dependence Coefficient
RKHS	Reproducing Kernel Hilbert Space
VAE	Variational Autoencoder

1

Introduction

Over the past few years, machine learning algorithms have emerged in many different fields of application. However, this development is accompanied by a growing concern about their potential threats, such as their ability to reproduce discrimination against a particular group of people based on sensitive characteristics (e.g., religion, race, gender, etc.). The standard machine learning models only optimize accuracy and are prone to learn all the relevant information for the task whether they are sensitive or not. In particular, algorithms trained on biased data have been shown to be susceptible to learn, perpetuate or even reinforce these biases (Bolukbasi et al., 2016). Many incidents of discrimination have been documented these recent years. For example, an algorithmic model used to generate predictions of criminal recidivism in the United States (COMPAS) discriminated against black defendants (Angwin et al., 2016). Also, discrimination based on gender and race could be demonstrated for targeted and automated online advertising on employment opportunities (Lambrecht and E. Tucker, 2016). The stakes are therefore major for citizens, and we must understand and master them. In this context, the EU introduced the General Data Protection Regulation (GDPR) in May 2018. This legislation represents one of the most important changes in the regulation of data privacy in more than 20 years. It strictly regulates the collection and use of sensitive personal data. With the aim of obtaining non-discriminatory algorithms, it rules in Article 9(1): "Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited." (European Commission, 2016). A naive method to mitigate the underlying bias could be to simply remove sensitive attributes from the training data set. This concept is

known as "fairness through unawareness" (Pedreshi et al., 2008). While this approach may prove viable when using conventional, deterministic algorithms with a manageable quantity of data, it is insufficient for machine learning algorithms trained on "big data". Here, complex correlations in the data may provide unexpected links to sensitive information. For example, the height of an adult could provide a strong indication about gender. Such a situation will be called "*proxy discrimination*" in (Prince and Schwarcz, 2019). For this reason, next to optimizing the performance of a machine learning model, the new challenge for data scientists/actuaries is to determine whether the model output predictions are discriminatory, and how it would be possible to mitigate such unwanted bias. A new research field has emerged to find solutions to this problem: fair machine learning. Recently, there has been a dramatic rise of interest by the academic community. Many questions have been raised, such as: How to define fairness (Hinnefeld et al., 2018; Hardt et al., 2016; Dwork et al., 2012; Kusner et al., 2017; Goyal et al., 2022)? How to mitigate the sensitive bias (Zhang et al., 2018; Grari et al., 2020a; Kamiran and Calders, 2012; Bellamy et al., 2018; du Pin Calmon et al., 2017; Zafar et al., 2017c; Celis et al., 2019; Wadsworth et al., 2018; Louppe et al., 2017; Do et al., 2021; Chen et al., 2019; Kearns et al., 2017; Emelianov et al., 2019)? How to keep a high prediction accuracy while remaining fair in a complex real-world scenario (Grari et al., 2019; Adel et al., 2019)? These questions will be further reviewed in the following parts.

Quantify and Define Fairness Objectives

Research on discrimination in technical systems has at its root a social and ethical concept and significantly precedes the current trend of work on fairness in machine learning (Moor, 1985; Huff and Cooper, 1987; Friedman and Nissenbaum, 1996). However, this literature reflects a subjective human sense of unfairness and does not suggest formalized quantitative measures. The first examples of fairness definitions (or fairness criteria) appeared in the field in 2008 known as "discrimination-aware data mining" (Pedreshi et al., 2008). Since then, quantifying and monitoring fairness from different perspectives has gained significant momentum in the scientific community, resulting in an incredible number of mathematical definitions proposed over the past decade. For example, in 2018, (Verma and Rubin, 2018) have identified more than twenty different notions of fairness, and that number has been growing ever since. This phenomenon is not surprising since every moral aspect of society, which may seem subjective, results in different desired positions for everyone. These various

objectives are separated into two main groups: Group fairness and Individual fairness. While Group fairness aims at treating all the different sensitive groups equally, Individual fairness aims at treating similar people similarly. Unfortunately, these fairness definitions have fundamental incompatibilities, they are partially contradictory and/or partially complementary. They cannot be satisfied simultaneously with each other, except under certain constraints that we will discuss. Therefore, because of the inherent incompatibility in the current standard strategies in fairness, practitioners implementing and/or evaluating fairness must choose only one among them. There is no explicit agreement on the definition to apply for each situation. The right choice of fairness definition depends primarily on the goodwill of practitioners and on the regulation law applied to each specific use case.

We note that, while plenty of measures have been proposed recently to monitor and quantify fairness for discrete variables, only a few are well suitable for continuous ones. This is a critical point since monitoring a fairness objective between the outputs of the regression models and any given continuous sensitive variable can be desirable, for example, between age-sensitive attribute and income output. For this purpose, we propose to address this issue with a new estimation of the Hirschfeld-Gebelein-Renyi (HGR) maximal correlation coefficient by neural network. We will show that it is a suitable measure to assess the level of dependence between outputs of regression models and any given continuous sensitive variable. This measure is more appropriate than the traditional Pearson correlation or Kendall's tau since it captures in a much more consistent way the non-linearity between variables. Also, we will show that it is a more relevant measure than some divergences from probability theory, such as the popular mutual information. Though the mutual information is widely spread in the literature, it is, however, difficult to measure, to be interpretable (e.g., not a normalized measure), and optimize in the continuous case with a finite set, as already shown numerous times in the literature (Mary et al., 2019; Lee, 2021; Bach and Jordan, 2002; Yan et al., 2020a). Also, note that mutual information is not a dependence measure according to Renyi's stipulations (Rényi, 1959). We will provide a theoretical analysis of the consistency of our HGR estimator, along with its nice properties compared to state-of-the-art measures. In addition, we will see that compared to the binary case where the fairness measures are fully reliable, they are, for the majority, only estimations in the continuous case. For this reason, we also contribute to a new measure, *FairQuant*, based on discretization. It requires splitting the set samples with different quantiles with regard to the sensitive attribute.

In terms of individual fairness, metrics measuring the total discrepancies between individuals remain understudied. For example, they do not allow the choice of a fixed

threshold defining when individuals are similar. We consider it essential that practitioners have control over this choice of similarity threshold, as it allows greater control over the objective. To this end, we will propose new metrics to assess individual-level fairness based on this purpose. Furthermore, regarding fairness from a causal perspective, we note that fairness criteria for assessing counterfactual fairness are employed in different ways in state-of-the-art. We will briefly address the difference and discuss some metrics that make the most sense in the desired context.

Achieving Fairness Objectives

In light of the recent popularity of fair machine learning, various developments have been made to increase fairness of the predictor model. Unfortunately, in most cases, improving fairness is at the expense of the primary goal of machine learning: Accuracy. However, sacrificing predictive performance is often not viewed as an acceptable option. Hence, the need to keep the maximum fairness level without degrading too much prediction accuracy has gained a high interest in the research community, leading to the development of new architectures for maximizing this trade-off between accuracy and fairness. Three prominent families of fairness approaches exist in the literature. While pre-processing (Kamiran and Calders, 2012; Bellamy et al., 2018; du Pin Calmon et al., 2017) and post-processing (Hardt et al., 2016; Chen et al., 2019) approaches respectively act on the input or the output of a classically trained predictor, in-processing approaches mitigate the undesired bias directly during the training phase (Zafar et al., 2017c; Zhang et al., 2018; Wadsworth et al., 2018; Louppe et al., 2017). In this thesis, we will focus on in-processing fairness and in particular with adversarial learning, which reveals as the most powerful framework for settings where acting on the training process is an option. The emergence of Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) provided the required underpinning for fair predictors using adversarial debiasing. The predictor model and the adversarial model are optimized simultaneously in a min-max game in order to find a trade-off between prediction accuracy and fairness. However, we will show in this thesis that while adversarial strategies are effective for debiasing in most cases and are considered as the most generalizable approach to different bias inducements (Li et al., 2021), many key application objectives are still missing and should be further studied.

Currently, we observe that many fairness applications focus on tabular data, mainly because it contains sensitive personal information about individuals and causes direct

and indirect discrimination. However, most bias mitigation strategies focus on neural networks and we noticed a lack of work on fair classifiers based on decision trees even though they have proven very efficient for tabular datasets. In an up-to-date comparison of state-of-the-art classification algorithms in tabular data, tree boosting outperforms deep learning (Zhang et al., 2017). For this reason, we propose a novel approach of adversarial gradient tree boosting for increasing fairness during training. To the best of our knowledge, this is the first adversarial learning method for generic classifiers, including decision trees. We empirically compare our proposal and its variants with several state-of-the-art approaches, for different fairness metrics. The results show that our algorithm achieves a higher accuracy while obtaining the same level of fairness, as measured using a set of different common fairness definitions.

The second studied issue is the mitigation of unwanted biases in a continuous case. As mentioned above and as shown in (Mehrabi et al., 2021), the continuous case for regression task and/or continuous sensitive attribute has not received the same amount of attention from the research community. Plenty of fair algorithms have been proposed recently to tackle this challenge for discrete variables, but only a few ideas exist for continuous ones. This is a central point since we demonstrate that the most traditional fair adversarial algorithms (Zhang et al., 2018; Adel et al., 2019) are not suitable for the continuous case. Indeed, we show that these latter algorithms theoretically do not generally optimize the most traditional fairness objective when the sensitive attribute is continuous. The traditional model structure has to be therefore revisited. For this reason, we propose a new adversarial architecture by minimizing our HGR estimation directly with adversarial neural network architecture. This strategy is theoretically optimal for solving the most common fairness objectives. The idea is to predict the target task while minimizing the ability of an adversarial neural network to find the estimated transformations required to predict the HGR coefficient. We empirically assess and compare our approach and demonstrate significant improvements on previously presented work in the field.

Another research sub-field family tackles the problem of learning *fair representations*. This approach resembles a mix of pre- and in-processing as the input data is mapped into intermediate learning representations that are unbiased with respect to sensitive source distribution and can therefore generalize to other domains. The literature shows that this approach empirically outperforms models that act directly on output predictions, which we denote *prediction retreatment*. We will propose to study in this thesis the reasons why such an architecture outperforms the state of the art. To the best of our knowledge, this is the first work to compare mitigation at different levels of neural architectures. We argue that acting at intermediary levels of neural

representations allows the best trade-off between expressiveness and generalization for bias mitigation. In addition, we propose to extend our neural network architecture to a fair representation by minimizing the HGR coefficient on an intermediary latent space. The HGR network is trained to discover non-linear transformations between the multidimensional latent representation and the sensitive feature. We empirically compare our different proposals and their variants with several state-of-the-art approaches, for different fairness metrics. Experiments show the great performance of our different approaches.

Third, we note that the vast majority of the state-of-the-art approaches rely on having access to the sensitive information to be mitigated during training. However, in practice, it is often unrealistic to assume that this sensitive information is available or even collected. In Europe, for example, a car insurance company cannot ask a potential client about his/her origin or religion, as this is strictly regulated. Yet, only a few prior works address the issue of mitigating bias in this difficult setting, in particular to meet classical fairness objectives. By leveraging recent developments for approximate inference, we propose a novel approach to fill this gap. To infer a sensitive information proxy, we introduce a new variational auto-encoding-based framework named SRCVAE that relies on knowledge of the underlying causal graph. The bias mitigation is then done in an adversarial fairness approach. Our proposed method empirically achieves significant improvement over existing works in the field. We observe that the generated proxy's latent space correctly recovers sensitive information and that our approach achieves a higher accuracy against competitors while obtaining the same level of fairness on real datasets.

Furthermore, we claim that mitigating undesired biases with a generic fair algorithm can be counterproductive for specific applications. For example, mitigating unwanted biases in insurance pricing with a traditional fair algorithm may be insufficient to maintain adequate accuracy. Indeed, the traditional pricing model is currently built in a two-stage structure that considers many potentially biased components such as car or geographic risks. We will show that this traditional structure has significant limitations in achieving fairness. There is a risk of not acting correctly all along with the components (for e.g. some of them can be unfairly neutralized on the objective predictive task). We extend the use of autoencoders to generate multiple aggregated pricing factors in a fairness context. We propose a general framework in which a single whole pricing model is trained by generating the geographic and car pricing components needed to predict the pricing premium while mitigating the unwanted bias according to a desired fair objective.

In addition, the definition of individual fairness suffers from a subtle concept of

"similar individual". According to its original mathematical definition, individual fairness requires a similarity distance between individuals to assess whether proximate individuals have comparable outcomes. However, this distance is a source of bias, and most current approaches rely on it. To this end, without accessing a similarity distance, we present a new framework method that leverages Variational Autoencoder (VAE). We demonstrate that our algorithm, can enforce individual fairness. In addition, we propose studying the connection with group fairness approaches to explore whether they can enforce fairness in an individual sense.

A last, from a fairness causal perspective, counterfactual fairness enforces the idea to imagine what any individual would look like with a variation of a given attribute of interest, such as a different gender or race for instance. This enables the simulation of counterfactual samples used for training the target fair model, the goal being to produce similar outcomes for every alternate version of any individual. We report that most of the current counterfactual approaches mitigate the bias in this intermediate phase called causal graph generation and then assess its fairness level. These works, which do not focus on the final predictor itself, assume that giving fair generated counterfactual observations as input to a traditional machine learning algorithm is sufficient to maintain the fairness objective. We argue that it is not always the case and the final predictions need to be evaluated to ensure a good fairness level. In addition, we note that no approach addresses the continuous case, the existing approaches may not hold when, for instance, the sensitive attribute is the age or the weight of an individual. To this end, we define a novel approach for counterfactual individual fairness tolerant to continuous features that focuses on the predictor model itself, notably via a dynamic sampling method that targets individualized hard locations of the sensitive space (Grari et al., 2020b).

Document Structure

The thesis is structured as follows. After a brief overview of the vast domain of the *Fair Machine Learning* field, Chapter 2 presents some key elements of technical context from related background that are relevant to this thesis and, in particular, the two main families of methods that we focus on: Group and Individual Fairness. Chapter 3 is devoted to the fairness measures for the different tasks of a predictive model. Then, we focus on how we correct biased models via adversarial learning for neural networks predictors in Chapter 4 and Gradient Tree Boosting in Chapter 5, respectively. In Chapter 6, we study the fairness issue when the sensitive attribute is not accessible,

and we propose a novel approach to fill this gap. In Chapter 7, we briefly describe the actuarial pricing literature with a traditional model, and propose a general model that is better adapted for a real-world fairness context. Chapter 8 is devoted to the study of individual fairness and counterfactual fairness.

Finally, this manuscript ends by summarizing the contributions of this thesis and discussing the perspectives it opens.

2

Unwanted Biases in Machine Learning

2.1 | Fairness in Machine Learning: Background

There are various notions of fairness, that can be related to discrimination. For instance, direct discrimination happens when a person is treated less favorably than another person in a comparable situation, in the sense that the two persons are otherwise similar except on a sensitive attribute. This is also called systematic discrimination or disparate treatment. In contrast, indirect discrimination happens when an "*apparently neutral practice put persons of a protected ground at a particular disadvantage compared with other persons*", as explained in (Zliobaite, 2015). Such discrimination is also known as structural discrimination or disparate outcome.

In many articles about discrimination, the sensitive attribute is the race, religion, gender, or sex (difference with gender in (Torgrimson and Minson, 2005)) of a person. The gender bias as for example a long history. In 1978, the Supreme Court in the U.S. stated that "*the differential was discriminatory in its "treatment of a person in a manner which, but for that person's sex, would be different." The statute, which focuses on fairness to individuals, rather than fairness to classes, precludes treating individuals as simply components of a group such as the sexual class here. Even though it is true that women as a class outlive men, that generalization cannot justify disqualifying an individual to whom it does not apply*". Following that decision, theoretical discussions about fairness of gender-based pricing started. In Europe, on 13 December 2004, the so-called *gender directive* was introduced (Council Directive 2004/113/EC), even if it took almost ten years to provide legal guidelines on the application of the directive to insurance activities. The goal was to enforce the principle of equal treatment between men and women in the

access to and supply of goods and services, including insurance. As a direct consequence, it prohibited the use of gender as a rating variable in insurance pricing. As discussed in (Schmeiser et al., 2014), gender equality in the European Union (EU) has been supposed to be ensured from 21 December 2012. However, even if the sensitive variable is not included in the training data, complex correlations from other features may provide unexpected persistent bias in the prediction outputs (Dwork et al., 2012; Pedreshi et al., 2008). For instance, the color and the model of a car combined with the driver's occupation can lead to unwanted gender bias in the prediction of car insurance price. Nevertheless, it is not obvious to claim that biases come explicitly from features in the training data, the model itself, or even the practitioner because each of them contributes to this result in one way or another. Yet, we will separate them to clarify it.

Biased Data Most sources of bias in machine learning lie in the data itself. For example, it may occur whenever the training data is not representative of the true population. One of them is the *Reporting Biases* that arise when reporting and measuring particular features of interest is skewed (Suresh and Guttag, 2019; Mehrabi et al., 2021). A dramatic example is the involvement of a commercial risk assessment software for criminal risk prediction in the United States, COMPAS, which takes as input a proxy of crimes from prior arrests. However, this proxy reflects a bias of policing and discriminatory laws in the United States that is known to be more likely to arrest more non-Caucasians than Caucasians. The results showed higher false-positive rates for African American offenders than Caucasian offenders. Another one is the *Selection biases* that occur when individuals in the data come from a non-random selection of the full distribution population. It may happen, for example, in insurance pricing, where a lack of geographical diversity in policyholders can be observed (Mathis, 2007). The results can demonstrate biases in favor of specific areas, which is known to be correlated with race (Frees and Huang, 2021).

Biased Predictor Model Some other sources of bias in machine learning come from the model predictor itself. First, even if the data are completely correct, the model can make errors on specific local segments especially when the observations are not majority. Note that monitoring multiple sensitives attributes such as gender and race, for example, increase the probability of having some minority group unbalanced with other and being unfairly treated. In addition, another concept less known in the Fair-ML community is the indirect discrimination from a causal paradox. For example, imagine an insurance context where sensitive attribute gender has originally no dependence on the claim target task. Trivially, one might expect the output of a model classifier to be unbiased since there is no bias between the target and the

sensitive. However, the output can be highly biased towards the sensitive in practice (Grari et al., 2019). This is related to the fact that the training dataset can contain explanatory variables highly correlated to the target and the sensitive attribute (e.g., car's color correlated to aggressiveness and gender). Unfortunately, in trying to recover the aggressiveness information, the predictor classifier indirectly captures the gender information in its prediction. This effect is illustrated below in a synthetic scenario described in subsection 5.3.

Biased Practitioner The practitioner can also bias the predictions in a certain manner. First, the process of 'Feature engineering', that consists in cleaning and transforming the data before the training phase, can be done in subjective ways. Some input variables are transformed to make them more valuable to the predictor model, which has a crucial downstream impact on the output predictions. Second, a less common example in the community is the use of specific "tricks" to modify the output as desired. Data scientists or actuaries may, in cases, slightly modify the output of ML models in a way that suits them. This type of practice can induce implicit biases in the prediction by the practitioner. A well-known example is the so-called "GAM" models because they allow modeling the prediction as desired. Non-life insurance actuaries, for example, use them widely in pricing modeling. Indeed, polynomial regression or splines allow correcting the trend. For example, by giving minimal credibility to the prediction of the number of claims of the elderly, they can reshape the trend that seems most accurate. However, there is no evidence that this is a wise choice. If, for example, several actuaries were given the same modeling task, the predictions might be different. This induces a bias on the part of the practitioners, based on their own mental and personal experiences, that may not apply more generally.

2.1.1 | Group Fairness

Group fairness refers to the notion that specific groups of people (e.g., men and women, Caucasians and non-Caucasians) are potentially subject to prejudice and unfair decisions. It, therefore, aims to ensure equal treatment of groups. Multiple statistical definitions have been proposed recently to address this intention from different perspectives. The most common one in the literature is the statistical independence criterion, *Demographic Parity*, which requires independence between the output prediction and the sensitive attribute. Other statistical criteria address the objective of the predictive task error measurement. One of them is the statistical separation criterion, *Equalized Odds* (Hardt et al., 2016), which requires the same independence as demographic parity and an additional condition based on the target feature task. This

metric is widely used in binary settings for comparing the different elements of confusion matrices between the sensitive groups. However, as we will discuss, the use of this latter criterion in a continuous setting is less meaningful and more challenging. For this reason, we have addressed a new statistical criterion based on residual independence, *Residuals Parity*, by requiring that the difference between the prediction and the target is independent to the sensitive attribute. We will also discuss the tension when trying to achieve several of these objectives simultaneously. Significant work has been done in enforcing these different fairness criteria. Three distinct strategy groups exist.

Algorithms that belong to the "pre-processing" family ensure that the input data is fair. This can be achieved by suppressing the sensitive attributes, changing class labels of the data set, and reweighting or resampling the data (Kamiran and Calders, 2012; Bellamy et al., 2018; du Pin Calmon et al., 2017).

The second group of mitigation algorithms follows a "post-processing" approach. In this case, only the output of a trained classifier is modified. For instance, a Bayes optimal equalized odds predictor can be used to change output labels with respect to an equalized odds objective (Hardt et al., 2016). A different paper presents a weighted estimator for demographic disparity which uses soft classification based on proxy model outputs (Chen et al., 2019). The advantage of post-processing algorithms is that fair classifiers are derived without the necessity of retraining the original model which may be time-consuming or difficult to implement in production environments. However, this approach may have a negative effect on accuracy or could compromise any generalization acquired by the original classifier (Donini et al., 2017).

The final type of mitigation strategy corresponds to the "in-processing" algorithms. Here, undesired bias is directly mitigated during the training phase. A straightforward approach to achieve this goal is to integrate a fairness penalty directly in the loss function. One such algorithm integrates a decision boundary covariance constraint for logistic regression or linear SVM (Zafar et al., 2017c). Then, the emergence of adversarial learning by employing two models that play against each other has led to many biases mitigation architectures. In this field, a neural network predictor is trained to predict the label Y, while simultaneously minimizing the ability of an adversarial model to identify the sensitive attributes from predictions (Zhang et al., 2018; Wadsworth et al., 2018; Louppe et al., 2017; Adel et al., 2019).

Prediction Retreatment We refer by *Prediction Retreatment* all the approaches that act on the output prediction itself. One of the most known approaches is the simple adversarial algorithm: To ensure independence between the output and the sensitive attribute, (Zhang et al., 2018) feeds the prediction output as input to an adversary net-

work. The objective of such an adversarial model is to predict the sensitive attribute, and update the predictor weights to fool the adversary.

Fair Representation On the other hand, several research sub-fields in the mix of in-processing and post-processing family tackle the problem of learning *fair representations*. Domain adaptation (Daume III and Marcu, 2006; Blitzer et al., 2006) and domain generalization (Muandet et al., 2013; Li et al., 2017) consist in learning representations that are unbiased with respect to a source distribution, and can therefore generalize to other domains. Some of the works in these fields involve the use of adversarial methods (Ganin and Lempitsky, 2014; Ganin et al., 2016; Tzeng et al., 2017), close to our work. Several strategies mitigate bias towards a sensitive attribute through representation. (Zemel et al., 2013) relies on a discriminative clustering model to learn a multinomial representation that removes information regarding a binary sensitive attribute. In a different approach, (Alvi et al., 2018) learn an unbiased representation by minimizing a confusion loss. Invariant representations can also be learnt using Variational Auto-Encoders (Kingma and Welling, 2013), by adding a mutual information penalty term (Moyer et al., 2018). (Adel et al., 2019) learn a fair representation by inputting it to an adversary network, which is prevented from predicting the sensitive attribute (see Chapter 4.3). Other works minimize the mutual information between the representation and the sensitive attribute: (Kim et al., 2019b) rely on adversarial training with a discriminator detecting the bias, while (Ragonesi et al., 2020) rely on a neural estimation of mutual information (Belghazi et al., 2018).

While group fairness is the predominant sub-field in the Fair-ML community and has proven effective in treating groups equally, it suffers from a significant ethical weakness: individuals can be unfairly treated.

2.1.2 | Individual Fairness

The idea behind individual fairness, as defined by (Dwork et al., 2012), is that "similar people should be treated similarly". This implies the existence of a similarity distance between individuals, referred to as $d_{\mathcal{X}}$, which generally comes from expert knowledge about the domain at hand but can also be learnt from data (Mukherjee et al., 2020; Ilvento, 2019), with either a human feedback or the assumption of access to embedded features satisfying a factor model. Some approaches assume access to this distance and propose to enforce individual fairness via regularization (Yurochkin and Sun, 2020) or distributionally robust optimization (Yurochkin et al., 2019). The process involves, at each iteration, finding similar individuals with the most disparate

treatments. Other methods consist in enforcing individual fairness without access to $d_{\mathcal{X}}$ (Gillen et al., 2018; Jung et al., 2019), but with access to an oracle that identifies fairness violations across pairs of individuals.
A slightly different notion of individual fairness is counterfactual fairness (Kocaoglu et al., 2018) where a decision is fair for an individual if it coincides with the one that would have been taken in a counterfactual world in which the values of its sensitive attributes were different. The objective is therefore to produce similar outcomes for every alternate version of any individual. For example, if I am a man who lives in Paris, have a black Renault car and am an engineer, the question is: what would I have as characteristics if I were a woman? Would I still have the same car, the same color ? Once assigned, the idea is to ensure similar outcomes for factual and counterfactual versions of individuals. Usually, counterfactuals are obtained by learning a causal inference framework involving unobserved confounders. Some approaches leverage the causal framework of Pearl (Russell et al., 2017). Other works have addressed this objective from GANs models (Kocaoglu et al., 2018; Xu et al., 2019). Other methods propose to use variational autoencoder methods (Louizos et al., 2017; Kim et al., 2021) to disentangle the exogenous uncertainty into different latent variables. However, most of these approaches enforce the fairness objective on the counterfactual generation and do not focus on the final predictor itself. They assume that giving a fair generated counterfactual observation as input to a traditional machine learning algorithm is sufficient to maintain the fairness objective. We argue that a two-step method that focusing separately on the generation of counterfactuals and on a fair prediction model is more appropriate. We propose to develop this general method by dealing with binary and continuous settings. We will compare empirically these two approaches in chapter 8.2.

2.2 | Problem Statement

Throughout this document, we consider a supervised machine learning predictive model h_{w_h} with parameters w_h for regression or classification problems, uses as a predictor model. We consider a variable Y that we want to predict for every input X, that is either quantitative or categorical and a collection of possible features that were collected. Among the features, S will denote sensitive attributes, for which we aim at ensuring fairness of the model. The training data consists of n examples $(x_i, s_i, y_i)_{i=1}^n$, where $x_i \in \mathbb{R}^d$ is the feature vector with d predictors of the i-th observation, $s_i \in \Omega_S$ the value of its sensitive attribute and $y_i \in \Omega_Y$ its label to be predicted. According

to the setting, the domain Ω_S of the sensitive attribute S can be either a discrete or a continuous set.

Although fairness is inherently an ethical and social concept, it is essential to define some clear mathematical objectives for machine learning predictions. In the following, we outline the most popular formal definitions used in recent research.

2.2.1 | The Objectives in Group Fairness

The Group Fairness family focus on the equality over the general distributions. In this part, we list three common objectives that we will use all along the thesis. Note that other objectives that we do not focus on are reported in (Verma and Rubin, 2018).

1) Demographic Parity The most common objective in fair machine learning is *Demographic parity* (Dwork et al., 2012) (or *Independence criteria*). Based on this definition, a model is considered fair if the output prediction $\widehat{Y}$ from features X is independent of the sensitive attribute S: $\widehat{Y} \perp\!\!\!\perp S$.

Definition 1. *A machine learning algorithm achieves* Demographic Parity *if the associated prediction $\widehat{Y}$ is independent of the sensitive attribute S* [1]*:*

$$\mathbb{P}(\widehat{Y}|S) = \mathbb{P}(\widehat{Y})$$

2) Equalized Odds The second most common objective in fair machine learning is *Equalized Odds* (Hardt et al., 2016) (or *strict separation criteria*). Based on this definition, a model is considered fair if the output prediction $\widehat{Y}$ from features X is independent of the sensitive attribute S given the outcome true value Y: $\widehat{Y} \perp\!\!\!\perp S|Y$.

Definition 2. *A machine learning algorithm achieves Equalized Odds if the associated prediction $\widehat{Y}$ is conditionally independent of the sensitive attribute S given Y:*

$$\mathbb{P}(\widehat{Y}|S, Y) = \mathbb{P}(\widehat{Y}|Y)$$

To illustrate Equalized Odds, consider a model developed to predict the presence of patient tumors in medical records with gender as a sensitive feature. Let's imagine that a traditional model (e.g, a logistic regression) would predict more tumors on male than female patients. In satisfying demographic parity on this model, the (positive) base rate would be the same for both males and females. This means that the

[1] For the binary case, it is equivalent to $\mathbb{E}(\widehat{Y}|S) = \mathbb{E}(\widehat{Y})$

model would increase the predictive error by detecting fewer tumors for males and more for females to enforce fairness. However, in a medical context, an error is more dramatic, especially for false negatives. In the Equalized Odds objective, the false-positive and false-negative rates will be the same for males and females. This fairness criterion seems, therefore, more appropriate for this medical application. Note that it would also be possible to focus solely on false negatives by assigning only the $Y = 1$ condition in the equalized odds definition. This particular case is named as *equalized opportunity* (Hardt et al., 2016).

3) Equalized Residuals We contributed to a new objective in fair machine learning denoted as *Equalized Residuals* (Grari et al., 2020a), in which a model is considered fair if the residuals $\widehat{Y} - Y$ from features X is independent of the sensitive attribute S: $\widehat{Y} - Y \perp\!\!\!\perp S$. While for the binary case, equalized odds appear to be sufficient since it allows the control of false-positive and false-negative rates, this fairness criterion is, better adapted for the continuous case.

Definition 3. *A machine learning algorithm achieves Equalized Residuals if the associated residuals* $(\widehat{Y} - Y)$ *is independent of the sensitive attribute S:*

$$P(\widehat{Y} - Y|S) = P(\widehat{Y} - Y)$$

To illustrate it, let's imagine a car insurance pricing scenario where young people have higher claims than older people. A classical pricing model would charge young people a higher premium. However, in the case of demographic parity, the average price would be the same across all ages. This means that older people would generally pay more more than their real cost, and younger people less. Equalized residuals will ensure that the residuals between the predictions and the real claim cost are preserved, independently from the sensitive variable age. This ensures that for all ages, the overall error does not deviate too much. The choice between the two is then subjective. Some will find it fairer to give the same price by choosing demographic parity, and others will find it abnormal to charge more elders than they should. They may consider it fairer to assign the same error.

Incompatibilities Objectives

The objectives mentioned above present fundamental incompatibilities. It is not usually possible for a predictor model to hold these objectives simultaneously (Guidotti et al., 2018; Hardt et al., 2016). We list below the different possibilities:

a) Demographic Parity AND Equalized Odds: Many papers recently claim (Barocas et al., 2017; Alves et al., 2021; Castelnovo et al., 2022) that when the Y target is binary, the independence and the separation criterion are at odds except in two specific cases. The first case is when the output prediction $\widehat{Y}$ is independent of the target variable Y. It creates tension with the predictive objective where the output predictions cannot accurately predict the target. The second statement is when the target variable Y must be independent of the sensitive attribute S. However, this specific case is only dependent on the original data set and can not be modified by the practitioners.

These two statements seen above are only compatibility requirements: at least one of them must be satisfied to achieve both objectives. However, note that satisfying one of them does not necessarily imply achieving both fairness criteria. We still have to enforce it. We cannot state, for example, that if $S \perp\!\!\!\perp Y$ then it necessarily implies demographic parity and equalized odds simultaneously. For example, let's consider $S \perp\!\!\!\perp Y$ and X not independent of S, the predictor model may not satisfy just demographic parity in particular cases (as demonstrated in Proposition 1). Please note that we empirically experiment this specific case with the red car example (see the subsection 5.3).

Proposition 1. *Assume Y and S binaries or continuous, S is independent of Y, X is not independent of S. Then, the demographic parity may not be satisfied.*

Proof. Assume $X \in [X_a, X_b]$ where $X_a \perp\!\!\!\perp S$ and $X_b \not\perp\!\!\!\perp S$ and suppose that the predictor h_{w_h} is a regression model where the activation function h is either sigmoid for a binary setting or identity for regression setting.

$$P(\widehat{Y} \mid S) = P(h_{w_h}(X) \mid S) = P(h_{w_h}(X_a, X_b) \mid S) \tag{2.1}$$
$$= P(h(w_{h_a} X_a + w_{h_b} X_b + b)) \mid S) \tag{2.2}$$

If $X_b \not\perp\!\!\!\perp Y \mid X_a, w_{h_b} \neq 0$ then: $P(\widehat{Y} \mid S) \neq P(S)$

Please note that Demographic Parity may be satisfies when the correlated variable X to S are independent to Y. For example, in our case if $X_b \perp\!\!\!\perp Y$ the weight w_{h_b} should be zero:

$$P(\widehat{Y} \mid S) = P(h_{w_h}(X) \mid S) = P(h_{w_h}(X_a, X_b) \mid S) \tag{2.3}$$
$$= P(h(w_{h_a} X_a + b)) \mid S) = P(S) \tag{2.4}$$

□

Finally, note that taking the output predictions as a constant corresponding to the majority classes or the mean (in the continuous case) as $\widehat{Y} = E(Y)$ achieves both objectives, but the performance accuracy cannot be greater than the variance.

b) Demographic Parity AND Equalized Residuals: Independence ($\widehat{Y} \perp\!\!\!\perp S$) and residual independence ($\widehat{Y} - Y \perp\!\!\!\perp S$) are incompatible except in one case. Both the target and the prediction are independent of the sensitive attribute S. In this case, it is not sufficient to obtain the output prediction $\widehat{Y}$ independent of the target variable Y. Moreover, the residuals may also suffer from a dependence on S related to the initial dependence between Y and S. Please note that using the above example with a constant model is not possible, as the error between the sensitive groups may vary.

In practice The different compatibility requirements are therefore strong and not necessary feasible in practice, this is is a fundamental problem since if multiple notions are required simultaneously by the practitioners, the machine learning model should make a trade-off to satisfy some objectives notions at the expense of others. Therefore, in the current standard strategies in fairness, practitioners implementing and/or monitoring fairness must choose only one among them.

2.2.2 | The Objectives in Individual Fairness

It is possible for a ML model that satisfies group fairness to be manifestly unfair to subgroups of protected groups and individuals (Dwork et al., 2012). For example, a person may be refused a position only because she belongs to a privileged group, regardless of her merit within the group. To cope with this issue two sub-fields in the individual domain area exist in the literature: Fairness Through Awareness and a causal perspective, counterfactual fairness, that we will describe below.

Fairness Through Awareness (FTA)

In this specific sub-field the objective is to force a predictor model to have similar output predictions to similar individuals. This objective involves comparing individuals rather than focusing on groups of people who share specific characteristics. Let's denote the corresponding information features x_1 and x_2 of two individuals. The objective is defined as follows (Dwork et al., 2012):

Definition 4. *A machine learning algorithm, with an associated predictor h with parameters w_h, achieves Individual Fairness with respect to a distance metric $d_{\mathcal{X}}$ on the input space $\mathcal{X}$ if*

h_{w_h} is K-lipschitz for a certain K:

$$\forall\, x_1, x_2 \in \mathcal{X}, |h_{w_h}(x_1) - h_{w_h}(x_2)| \leq K d_{\mathcal{X}}(x_1, x_2)$$

Individual fairness is therefore dependent on the choice of the input space distance $d_{\mathcal{X}}$. The distance $d_{\mathcal{X}}$ can come from expert knowledge, or can be learnt from data as outlined in (Mukherjee et al., 2020). This choice affect the fairness objective.

In particular, the distance $d_{\mathcal{X}}$ can be learned with the sensitive subspace method of (Yurochkin et al., 2019) which is defined as follows:

$$d_{\mathcal{X}}(x_1, x_2) = (x_1 - x_2)^T (I - P_{ran(A)})(x_1 - x_2) \tag{2.5}$$

where $P_{ran(A)}$ is the projection matrix onto the span of $A = [a_1, .., a_k]$ which is referred to as the sensitive subspace. The sensitive subspace can be learnt by fitting a model to predict S with X as variable: either a softmax regression model for a discrete sensitive variable, or an appropriate generalized linear model for a continuous sensitive variable. The vectors $[a_1, .., a_k]$ can then be defined as the weights of the fitted model. Defined that way, the distance declares pairs of points that differ mainly in their sensitive attributes as similar.

Counterfactual Individual Fairness

The Counterfactual fairness has been recently introduced for quantifying fairness at the most individual sense (Kusner et al., 2017). Rather than globally considering equity over the entire population, the idea is to imagine what any individual would look like with a variation of a given attribute of interest, such as a different gender or *race* for instance. This approach considers additional knowledge, it requires strong hypothesis about the structure of the world, in the form of a causal model. One of the most known example in the literature is to leverage the previous work in (Pearl et al., 2009), which provides general theory for modeling causal relationships between variables. Inferring causal effects in the causal model is facilitated by do-operator by simulating the physical intervention that forces some variable X to take a certain value x, formally denoted by do(X = x) or do(x).

Definition 5. *Counterfactual demographic parity (Kusner et al., 2017): A predictive function h_{w_h} is considered counterfactually fair for a causal world G, if for any $x \in X$ and $\forall y \in Y, \forall (s_0, s_1) \in \Omega_S$ with $s_0 \neq s_1$: $p(\hat{Y}_{S \leftarrow s_0} = y | X = x, S = s_0) = p(\hat{Y}_{S \leftarrow s_1} =$*

$y|X = x, S = s_0)$, *where* $\hat{Y}_{S \leftarrow s_1} = h_\theta(X_{S \leftarrow s_1})$ *is the outcome of the predictive function* h_{w_h} *for any transformation* $X_{S \leftarrow s_1}$ *of input* X*, resulting from setting* s_1 *as its sensitive attribute value, according to the causal graph* G*.*

Following definition 5, an algorithm is considered counterfactually fair in term of *independence* if the predictions are equal for each individual in the factual causal world and in any counterfactual world. It therefore compares the predictions of the same individual with an alternate version of him/herself. Similar extension can be done to adapt a *separation* objective for the Counterfactual framework (Pfohl et al., 2019). Learning transformations $\hat{X}_{S \leftarrow s}$ for a given causal graph is at the heart of Counterfactual Fairness, as described in the section 8.2

2.3 | Discussion

After having sketched some key elements of fairness, this chapter presented two families of approaches: group fairness and individual fairness. Practitioners implementing and/or evaluating fairness in technical systems must first decide the main concept to apply. This choice is determined by the goodwill of the practitioner or by the regulation law. We note that the law is currently not very clear on this subject. Only the objectives of fairness through unawareness are mandatory in some areas, as in insurance pricing with the gender feature. However, the situation is likely to change in the future, and criteria other than fairness unawareness will probably be considered. Once the choice of an objective has been made, it is necessary to choose the statistical measure to be used in order to quantify and mitigate the desired bias. The following points will be discussed in the next two chapters.

3

Measuring Fairness

In this chapter, we discuss how to mathematically quantify the different fairness objectives that were outlined in the previous chapter. Because the different measures differ depending on the types of sensitive attributes S and the objective task, we propose to organize this chapter in three different parts: Binary, Continuous, and Frequency settings. Note that we contributed to new estimation measures on Continuous Statistical Dependence and Frequency Statistical dependence.

The contributions of this chapter are threefold:

- We propose a new estimation of the the χ^2 divergence and the HGR maximal correlation along with a theoretical analysis of its consistency ;
- We propose a new measure for continuous setting based on the discretization of the sensitive attribute.
- We provide new measures for Fairness Through Unawareness based on a similarity individual threshold and we propose a Total Effect measure for counterfactual fairness based on the predictor model.

3.1 | Group Fairness in Binary Setting

This section focuses on a binary scenario where the targeted sensitive attribute and the actual value of the outcome are both binary ($(S, Y) \in [0, 1]$)

1) Demographic Parity The use of *Demographic Parity* was originally introduced in this context of binary scenarios (Dwork et al., 2012), where the underlying idea is that each demographic group has the same chance for a positive outcome.

Definition 6. *A classifier is considered fair according to the demographic parity principle if*

$$\mathbb{P}(\widehat{Y} = 1|S = 0) = \mathbb{P}(\widehat{Y} = 1|S = 1)$$

There are multiple ways to assess this objective. The p-rule assessment ensures the ratio of the positive rate for the unprivileged group is no less than a fixed threshold p. The classifier is considered totally fair when this ratio satisfies a 100%-rule. Conversely, a 0%-rule indicates a completely unfair model.

$$\textit{p-rule: } \min\left\{\frac{\mathbb{P}(\widehat{Y} = 1|S = 1)}{\mathbb{P}(\widehat{Y} = 1|S = 0)}, \frac{\mathbb{P}(\widehat{Y} = 1|S = 0)}{\mathbb{P}(\widehat{Y} = 1|S = 1)}\right\} \tag{3.1}$$

The second metric available for Demographic Parity is the disparate impact (DI) assessment (Feldman et al., 2015). It considers the absolute difference of outcome distributions for subpopulations with different sensitive attribute values. The smaller the difference, the fairer the model.

$$\textit{DI: } |\mathbb{P}(\widehat{Y} = 1|S = 1) - \mathbb{P}(\widehat{Y} = 1|S = 0)| \tag{3.2}$$

Note that potential differences between demographic groups are not taken into account in this notion. Indeed, in this binary context, only the *weak independence* is required[1]: $\mathbb{E}(\widehat{Y}|S) = \mathbb{E}(\widehat{Y})$.

2) Equalized Odds Equalized Odds can be measured with the disparate mistreatment (DM) (Zafar et al., 2017c). It computes the absolute difference between the false positive rate (FPR) and the false-negative rate (FNR) for both demographics.

$$D_{FPR} : |\mathbb{P}(\widehat{Y} = 1|Y = 0, S = 1) - \mathbb{P}(\widehat{Y} = 1|Y = 0, S = 0)| \tag{3.3}$$

$$D_{FNR} : |\mathbb{P}(\widehat{Y} = 0|Y = 1, S = 1) - \mathbb{P}(\widehat{Y} = 0|Y = 1, S = 0)| \tag{3.4}$$

The notion of fairness here is that chances of being correctly (or incorrectly) classified as positive should be the same across groups. The closer the values of D_{FPR} and D_{FNR} to 0, the lower the degree of disparate mistreatment of the classifier. Therefore the classifier is considered fair if across both demographics $S = 0$ and $S = 1$, for the outcome $Y = 1$ the predictor $\widehat{Y}$ has equal *true* positive rates, and for $Y = 0$ the predictor $\widehat{Y}$ has equal *true*-negative rates (Hardt et al., 2016). The implicit notion here is that a perfect classifier is necessarily fair. The European Commission warns that this practice can lead to taking as much information as possible for prediction, which contradicts the data minimization requirement of the General Data Protection Regulation. Also, giving too much importance to the target Y can be counterproductive if this variable is itself a source of bias (Castelnovo et al., 2022).

[1] $\mathbb{E}(\widehat{Y}) = 1 * \mathbb{P}(\widehat{Y} = 1) + 0 * \mathbb{P}(\widehat{Y} = 0) = \mathbb{P}(\widehat{Y} = 1)$ (same with $\mathbb{E}(\widehat{Y} \mid S) = \mathbb{P}(\widehat{Y} = 1 \mid S)$)

3.2 | Group Fairness in Continuous Statistical Dependence

We consider now a continuous scenario where the targeted sensitive attribute and the actual value of the outcome are both continuous ($(S, Y) \in \mathbb{R}^2$). In order to assess these fairness definitions in the continuous case, it is essential to look at the concepts and measures of statistical dependence. Simple ways of measuring dependence are Pearson's correlation, Kendall's tau, or Spearman's rank correlation. Those types of measures have already been used in fairness, with the example of mitigating the conditional covariance for categorical variables (Zafar et al., 2017c). However, the major problem with these measures is that they only capture a limited class of association patterns, like linear or monotonically increasing functions. For example, a random variable with standard normal distribution and its cosine (hence, non-linear) transformation are not correlated in the sense of Pearson.

3.2.1 | Dependence via Information Theory

One of the most fundamental quantity for measuring the dependence between two random variables is the Mutual information. Let U and V be random variables, the mutual information is defined as:

$$I(U, V) = \int_{\mathbb{R}} \int_{\mathbb{R}} P_{U,V}(j, j') * log(Q_{U,V}(j, j'))djdj' \tag{3.5}$$

with :

$$Q_{U,V}(j, j') = \frac{P_{U,V}(j, j')}{\sqrt{P_U(j)}\sqrt{P_V(j')}} \tag{3.6}$$

Where $P_{U,V}$ is the joint distribution of U and V, P_U and P_V are the corresponding marginal distributions. Many recent works focused on the approximation of the mutual information. Its estimation can be done with a k-NN-based non-parametric estimator (Kraskov et al., 2004) or by neural estimation with MINE (Belghazi et al., 2018).

Recently, the usage of an other f-divergence, the χ^2 divergence, has been of high interest in the fairness research field (Fukuchi and Sakuma, 2015; Hashimoto et al., 2018) and, more specifically, in the continuous setting where the mutual information have more difficulty as shown in (Mary et al., 2019). The χ^2 divergence between the joint distribution and the product of its marginals is defined as follows:

$$\chi^2(P_{U,V}, P_U \otimes P_V) = \int_{\mathbb{R}} \int_{\mathbb{R}} Q_{U,V}(j, j')^2 djdj' - 1 \tag{3.7}$$

The strategy proposed by (Mary et al., 2019) is to estimate the χ^2 divergence via Kernel Density Estimation (KDE). This implies the difficult choice of the kernel density function to be used. All estimations in (Mary et al., 2019) have been done with a Gaussian kernel by setting the bandwidth with the classic Silverman's rule (Läuter, 1988), which can be hard to generalize for all data. For this reason, we propose a new approach to avoid approximating the density with KDE. We apply an approach similar to MINE (Belghazi et al., 2018). For this, we use the following "dual representation" of χ^2.

Theorem 3.2.1. *The χ^2 divergence admits the following representation (Broniatowski and Leorato, 2006):*

$$\chi^2(P,Q) = \sup_{f:\Omega\to\mathbb{R}} E_P(f) - E_Q(f + \frac{1}{4}f^2) \tag{3.8}$$

where the supremum is taken over all functions f such that the expectations are finite.

Algorithm 1 χ^2 Neural Estimation

Input: Distributions $P_{U,V}$ and P_V, Neural Network f_θ, Batchsize b, Learning rate α
repeat
Draw b samples from the joint distribution:
$(u_1, v_1), ..., (u_b, v_b) \sim P_{UV}$
Draw b samples from the V marginal distribution:
$\bar{v}_1, ..., \bar{v}_b \sim P_V$
Evaluate the lower bound:
$J(\theta) \leftarrow \frac{1}{b}\sum_{i=1}^{b} f_\theta(u_i, v_i) - \frac{1}{b}\sum_{i=1}^{b}(f_\theta(u_i, \bar{v}_i) + \frac{1}{4}f_\theta(u_i, \bar{v}_i)^2)$
Update the network parameters by gradient ascent:
$\theta \leftarrow \theta + \alpha\nabla J(\theta)$
until convergence

Like MINE (Belghazi et al., 2018) we use a set of functions $\mathcal{F}_\Theta = \{f_\theta\}_{\theta\in\Theta}$, defined by a given neural network f with parameters $\theta \in \Theta$, as the class of functions that can be considered in 3.8 for the approximation χ^2_θ of χ^2. We get the following objective to optimize for neural estimation of χ^2:

$$\chi^2_\Theta(P_{UV}, P_U \otimes P_V) = \max_{\theta\in\Theta} E_{P_{UV}}(f_\theta) - E_{P_U\otimes P_V}(f_\theta + \frac{1}{4}f_\theta^2) \tag{3.9}$$

This is done by algorithm 1, which takes distributions P_{UV} and P_V as input, from which it draws mini-batches of $P_{U,V}$ and $P_U \otimes P_V$ (in practice, samples are drawn from training data (U, V) rather than from true distributions). Then, mini-batches are used

to estimate and optimize the difference of expectations from 3.9 via stochastic gradient ascent. Note that, since the use of neural networks for χ^2 estimation restricts the possible functions to a given compact set $\mathcal{F}_\Theta$ defined by the neural architecture used, χ^2_Θ is only a lower-bound of χ^2: $\chi^2(P_{UV}, P_U \otimes P_V) \geq \chi^2_\Theta(P_{UV}, P_U \otimes P_V)$. However, it enables to obtain very efficient estimations of χ^2 for various distributions of variables, as shown in figure 3.1 for Multivariate Gaussians where the x-axis represents different covariances between U and V. In this figure, we observe that, as expected, our estimated χ^2_θ divergence is below the real χ^2 for this data, while presenting a very similar shape.

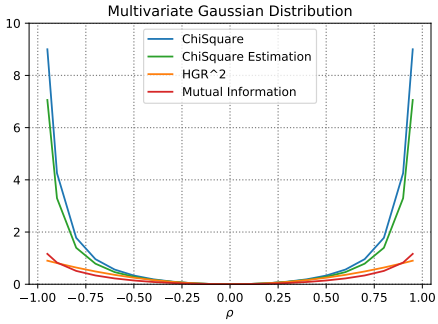


Figure 3.1: χ^2 estimation for bivariate Gaussians with a ρ correlation coefficient

Although the f-divergence and mutual information is widely spread in the literature, it is, however, difficult to measure, to be interpretable (e.g., not a normalized measure), and optimize in the continuous case with a finite sample, as already shown numerous times in the litterature (Lee, 2021; Bach and Jordan, 2002; Yan et al., 2020a). To cope with these issues, we will focus on statistical correlation measures in the following.

3.2.2 | Correlation Dependence

To overcome linearity and the f-divergence limitations, (Scarsini, 1984) first introduced a series of axioms that a *measure of concordance* δ between two random variables should satisfy. Among them, $\delta(U, V) = 0$ if and only if U and V are independent, and $\delta(U, V) = 1$ if and only if U and V are co-monotonic, meaning that there is some deterministic monotone relationship between U and V (i.e., there are f and g such that $V = f(U)$ and $U = g(V)$). (Rényi, 1959) suggested to consider the supremum of $r(f(U), g(V))$, where r denotes Pearson's correlation, for all functions f and g such that the correlation can be computed. Such measure was considered earlier in (Hirschfeld, 1935) and (Gebelein, 1941), defined as:

$$HGR(U,V) = \max_{f,g} \{r(f(U), g(V))\},$$

An alternative expression is obtained by considering standardized transformations for both variables where $\mathcal{S}(U) = \{f : \mathcal{U} \to \mathbb{R} : \mathbb{E}[f(U)] = 0 \text{ and } \mathbb{E}[f(U)^2] = 1\}$ for any variable U.

$$HGR(U,V) = \max_{f \in \mathcal{S}(U), g \in \mathcal{S}(V)} \{\mathbb{E}[f(U)g(V)]\}. \tag{3.10}$$

This measure also appeared earlier in (Barrett and Lampard, 1955) and (Lancaster, 1958), while introducing what is called *nonlinear canonical analysis*, where we write the joint density of the pair (U,V) as:

$$P_{UV}(u,v) = P_U(u)P_V(v)\left[1 + \sum_{i=1}^{\infty} \alpha_i h_{U,i}(u) h_{V,i}(v)\right]$$

for some decreasing α_i's in $[0,1]$ and for some series of orthonormal centered functions $h_{U,i}$'s and $h_{V,i}$'s, called canonical components. Then one can prove that under mild technical conditions,

$$HGR(U,V) = \max_{f \in \mathcal{S}(U), g \in \mathcal{S}(V)} \{\mathbb{E}[f(U)g(V)]\} = \mathbb{E}[h_{U,1}(U)h_{V,1}(V)]$$

The HGR coefficient is equal to 0 if, and only if, the two random variables are independent. If they are strictly dependent the value is 1. For instance, if (U,V) is a Gaussian vector with correlation r, then h_i's are Hermite's polynomial functions, $h_1(u) = u$, and $HGR(U,V) = |r|$ (the value of the maximal correlation in the Gaussian case was actually established in (Gebelein, 1941)). (Jensen and Mayer, 1977) extended (Rényi, 1959)'s approach by considering some association measure that depend non only on the first canonical correlation, but all of them. Several papers, such as (Buja, 1990), discussed the estimation of maximal correlation, or such as kernel based techniques in (Dauxois and Nkiet, 1998), where U and V are no longer univariate random variables but can take values in more general Hilbert spaces, and more recently (Tjøstheim et al., 2022).

The spaces for the functions f and g are infinite-dimensional. This property is the reason why the HGR coefficient proved difficult to compute. Over the last few years, many other non-linear dependence measures have been introduced like the Kernel Canonical Correlation Analysis (KCCA) (Hardoon and Shawe-Taylor, 2009), the Distance or Brownian Correlation (dCor) (Székely et al., 2009), the Hilbert-Schmidt Independence Criterion (HSIC and CHSIC) (Gretton et al., 2005; Póczos et al., 2012).

Comparing those non-linear dependence measures (Lopez-Paz et al., 2013), the HGR coefficient seems to be an interesting choice: it is a normalized measure which is capable of correctly measuring linear and non-linear relationships, it can handle multi-dimensional random variables and it is invariant with respect to changes in marginal distributions.

Note that correlation measure HGR can be extended to a conditional version [3] as:

$$HGR(U,V|Z) = \max_{f\in\mathcal{S}_{(U|Z)},g\in\mathcal{S}_{(V|Z)}} \Big\{\mathbb{E}[f(U)g(V)|Z]\Big\}$$

Several approaches rely on Witsenhausen's linear algebra characterization (see (Witsenhausen, 1975) to compute the HGR coefficient. For discrete features, this characterization can be combined with Monte-Carlo estimation of probabilities (Baharlouei et al., 2019), or with kernel density estimation (KDE) (Mary et al., 2019) to compute the HGR coefficient. Note that this latter approach can be extended to the continuous case by discretizing the density support. However, they make a strong assumption by basing their approach on a Gaussian Kernel Distribution Estimator (KDE). This makes it difficult to generalize on all different kinds of data sets. Another way to approximate this coefficient, Randomized Dependence Coefficient (RDC) (Lopez-Paz et al., 2013), is to require that f and g belong to reproducing kernel Hilbert spaces (RKHS) and take the largest canonical correlation between two sets of copula random projections. However, it has been shown to have difficulty to recover optimal transformations on empirical scenarios in which the relationship is highly unsteady (Mary et al., 2019). We will propose a new estimation of the HGR maximal correlation to overcome these limitations.

3.2.3 | HGR Estimation by Neural Network

Recently we proposed a new approach (Grari et al., 2020a) to estimate the HGR by deep neural network (HGR_{NN}). The main idea is to use two inter-connected neural networks to approximate the optimal transformation functions f and g from Eq. 3.10. The HGR_NN estimator is computed by considering the expectation of the products of standardized outputs of both networks ($\hat{f}_{w_f}$ and $\hat{g}_{w_g}$). The respective parameters w_f and w_g are updated by gradient ascent on the objective function to maximize: $J(w_f, w_g) = E[\hat{f}_{w_f}(U)\hat{g}_{w_g}(V)]$. This estimation has the advantage of being estimated by backpropagation. Algorithm 2 depicts the optimization process of our estimation of the HGR. Until convergence, it samples instantiations of (U, V) from $P_{U,V}$ (or from

[2] $r(U,V) := \frac{Cov(U;V)}{\sigma_U\sigma_V}$, where $Cov(U;V)$, σ_U and σ_V are respectively the covariance between U and V,

Algorithm 2 HGR Estimation by Neural Network

Input: Distributions $P_{U,V}$, Neural Networks f_{ω_f} and g_{ω_g},
Batchsize b, Learning rates α_f, α_g
repeat
Draw b samples from the joint distribution:
$(u_1, v_1), ..., (u_b, v_b) \sim P_{UV}$
Calculate the average and variance of the transformation predictions:
$m_f \leftarrow \frac{1}{b}\sum_{i=1}^{b} f_{\omega_f}(u_i)$; $\sigma_f^2 \leftarrow \frac{1}{b}\sum_{i=1}^{b}(f_{\omega_f}(u_i) - m_f)^2$
$m_g \leftarrow \frac{1}{b}\sum_{i=1}^{b} g_{\omega_g}(v_i)$; $\sigma_g^2 \leftarrow \frac{1}{b}\sum_{i=1}^{b}(g_{\omega_g}(v_i) - m_g)^2$
Standardize w.r.t. the minibatch:
$\forall i : \hat{f}_{\omega_f}(u_i) \leftarrow \frac{f_{\omega_f}(u_i) - m_f}{\sqrt{\sigma_f^2 + \epsilon}}; \hat{g}_{\omega_g}(v_i) \leftarrow \frac{g_{\omega_g}(v_i) - m_g}{\sqrt{\sigma_g^2 + \epsilon}}$
Maximize the following objective function J by gradient ascent:
$J(\omega_f, \omega_g) = \frac{1}{b}\sum_{i=1}^{b} \hat{f}_{\omega_f}(u_i) * \hat{g}_{\omega_g}(v_i)$
$\omega_f \leftarrow \omega_f + \alpha_f \frac{\partial J(\omega_f, \omega_g)}{\partial \omega_f}$; $\omega_g \leftarrow \omega_g + \alpha_g \frac{\partial J(\omega_f, \omega_g)}{\partial \omega_g}$
until convergence

a training set of data) to form mini-batches. At each iteration, it computes expectation and variance estimators of f_{ω_f} and g_{ω_g} on the current batch. These estimators are used to standardize the outputs of both neural networks on the batch. Finally, it updates the parameters of both networks by gradient ascent on the objective function to maximize $J(\omega_f, \omega_g)$. Note that the gradients are computed by back-propagating not only through the output values of ω_f and ω_g but also through means and variances of the batch, to ensure convergence. At the end, the $HGR_{NN}(U, V)$ estimator can be computed by considering the expectation of the products of standardized outputs of both networks.

This neural estimator $HGR_{NN}(U, V)$ is a lower-bound of $HGR(U, V)$ (at least on the training data set). However, as experimentally shown in figure 3.2, our estimator gives very accurate approximations in various settings. For these experiments, we produced artificial data (U, V) with non-linear dependencies. Four data sets were generated by instantiating U with samples drawn from an uniform distribution $\mathcal{U}(-10; 10)$ and defining V according to a non-linear transformation of U: $V = F(U) + \epsilon$, with F a given association pattern and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ a random noise added to V. Each sub-figure in Fig.3.2 corresponds to a data set generated according to the association pattern plotted in the small box in its left corner (500 pairs (U, V) were generated for each data set). Note that for each of the scenarios, the linear corre-

the standard deviation of U and the standard deviation of V, respectively.

[3]Where $\mathcal{S}_{(U|Z)} = \{f : \mathcal{U} \to \mathbb{R} : \mathbb{E}[f(U)|Z] = 0 \text{ and } \mathbb{E}[f(U)^2|Z] = 1\}$

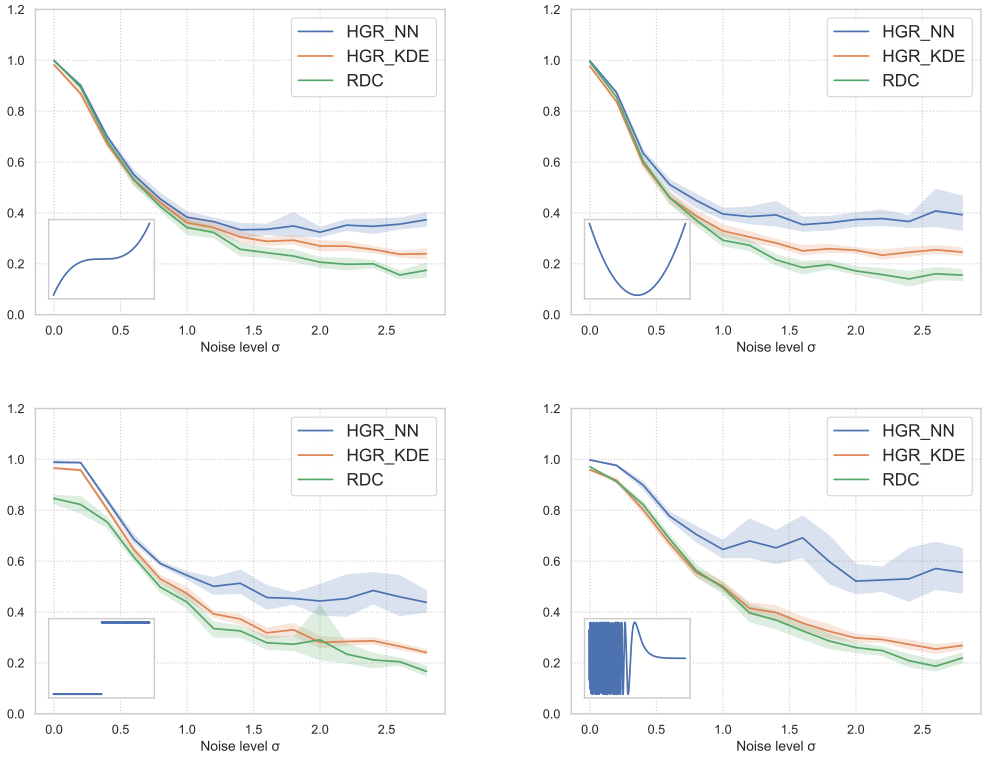


Figure 3.2: HGR estimation in various settings

lation between U and V is 0, but the HGR coefficient is theoretically equal to 1 when no noise is added to the transformation (when $\sigma^2 = 0$). The aim is to assess the ability of the HGR estimators to recover the HGR value, despite some complex association patterns and some noise in the data. We compare the HGR estimation values with other approaches described above, denoted HGR_KDE for its KDE estimation and HGR_RDC for the random copula approach. For our estimator HGR_NN, we consider some neural networks f and g of three layers, each including ten units with tanh activation function and Xavier initialization.

It shows that that, when no noise is added to the data, HGR_KDE and HGR_RDC have difficulties to recover the optimal transformations on the two last scenarios in which the relationship is either not continuous or highly unsteady. Thanks to the higher freedom provided by the use of neural networks, HGR_NN succeeds in retrieving a HGR of 1 for these settings. When noise is added to the data, the true HGR coefficient could be lower than 1. We thus assess the ability of the estimators to approach the HGR value that would be induced by optimal transformations f and g on the data. Note that our approach cannot exceed its value, due to the use of a

restricted set of neural transformation functions. From the figure, we observe that the curve of HGR_NN is always the highest (thus the closest to the optimal value from the data), and that the difference between our approach and the others increases with noise. HGR_NN appears more robust to noise. Additional experiments on the power of dependence of our estimator have also shown that our estimator is usually more efficient than its competitors for discerning dependent from independent samples in various settings.

Theoretical Properties: In the following we study the consistency of our HGR_{NN} estimator. We consider a family of continuous neural networks $F_\Theta = (f_{w_f}, g_{w_g})$ parametrized by a compact domain $\Theta \subset \mathbb{R}^k$ with $\Theta = (w_f, w_g)$ where our HGR approximation is denoted as $HGR_\Theta(U, V)$. All the proofs can be found in the appendix in section A.1.

Definition 7. *(Strong consistency) The estimator* $\widehat{HGR(U,V)}_n$ *is strongly consistent if for all* $\epsilon > 0$*, there exists a positive integer* N *and a choice of statistics network such that:*

$$|HGR(U,V) - \widehat{HGR(U,V)}_n| \leq \epsilon, a.s. \tag{3.11}$$

As explained in MINE (Belghazi et al., 2018), the question of consistency is divided into two problems: a deterministic approximation problem related to the choice of the statistics network, and an estimation problem related to the use of empirical measures.

The first lemma addresses the approximation problem using universal approximation theorems for neural networks (Hornik et al., 1989):

Lemma 3.2.2. *(approximation) Let* $\eta > 0$*. There exists a family of continuous neural networks* F_Θ *parametrized by a compact domain* $\Theta \subset \mathbb{R}^k$*, such that*

$$|HGR(U,V) - HGR_\Theta(U,V)| \leq \eta. \tag{3.12}$$

The second lemma addresses the estimation problem, making use of classical consistency theorems for extremum estimators (Geer and van de Geer, 2000). It states the almost sure convergence of HGR_NN to the associated theoretical neural HGR measure as the number of samples goes to infinity:

Lemma 3.2.3. *(estimation) Let* $\eta > 0$*, and* F_Θ *a family of continuous neural networks parametrized by a compact domain* $\Theta \subset \mathbb{R}^k$*. There exists an* $N \in \mathbb{N}$ *such that:*

$$\forall n \geq N, |\widehat{HGR(U,V)}_n - HGR_\Theta(U,V)| \leq \eta, a.s. \tag{3.13}$$

It is implied here that, from rank N, all sample variances are positive in the definition of $\widehat{HGR(U,V)}_n$, which makes the latter well-defined.

We deduce from these two lemmas the following result:

Theorem 3.2.4. *$\widehat{HGR(U,V)}_n$ is strongly consistent.*

3.2.4 | Demographic Parity in the Continuous Case

Compared to the most common discrete binary setting, where the demographic parity can be reduced to ensure *weak independance:* $\mathbb{E}[\widehat{Y}|S] = \mathbb{E}[\widehat{Y}]$ (Agarwal et al., 2018) and implies $\sup_f r(f(X),S) = 0$, **it does not generally imply demographic parity** when S is continuous. On the other hand, the minimization of the HGR dependence ensures *strong Independence* on distributions: $\mathbb{P}[\widehat{Y}|S] = \mathbb{P}[\widehat{Y}]$ (Grari et al., 2020a) and therefore satisfies the demographic parity objective as below:

Definition 8. *A machine learning algorithm achieves* Demographic Parity *if the associated prediction $\widehat{Y}$ and the sensitive attribute S satisfies:*

$$HGR(\widehat{Y},S) = 0. \tag{3.14}$$

Compared to the binary case where fairness measures as p-rule or the DI metrics are fully reliable, they are not sufficient in the continuous case. Also, all the continuous measures discussed so far (HGR_NN,HGR_KDE, HGR_RDC,χ^2_NN and MINE) are only estimations and are therefore not totally reliable. For this reason, we also introduce a metric based on discretization of the sensitive attribute. This metric, denoted as *FairQuant* metric (Grari et al., 2020a) splits the set samples X in K quantiles denoted as X_k (50 in our experiments) with regards to the sensitive attribute, in order to obtain sample groups of the same size. It computes the mean absolute difference between the global average and the means computed in each quantile:

Definition 9. *We define K as the number of quantiles, m_k as the mean of the predictions $h_w(X_k)$ in the k-th quantile set X_k, and m its mean on the full sample X. The FairQuant is defined as below:*

$$FairQuant = \frac{1}{K}\sum_{k=1}^{K}|m_k - m| \tag{3.15}$$

3.2.5 | Equalized Residuals in the Continuous Case

Following the same reasoning, we compare the model's residuals with the sensitive attribute. We also consider the HGR measure in this context since the weak independence by expectation is not sufficient from residuals observations.

Proposition 1. *A machine learning algorithm achieves equalized residuals if the associated residuals $\widehat{Y} - Y$ and the sensitive attribute S satisfies:*

$$HGR(\widehat{Y} - Y, S) = 0. \tag{3.16}$$

Therefore, the two metrics we will use for this purpose are $HGR(\widehat{Y} - Y, S)$ and the Fairquant on residuals. The latter takes m_k as the mean of the residuals over the k-th set of quantiles and m its mean on the full sample X.

3.2.6 | Group Fairness in Frequency Statistical Dependence

In this setting, we assume that the outcome target Y can be represented as a number of events occurring in a fixed interval of time. In this particular frequency setting, we notice a lack of work for assessing the level of fairness.

1) Demographic Parity For the demographic parity objective, the weak independence by expectation is not sufficient in this context to ensure the definition of 2.1. Indeed, the nature of the continuous output of a number of events requires the notion of strong independence in the distribution. We propose in this context to apply the same notion as seen above in the continuous proposes section 3.2 by assessing the HGR neural network estimator and the Fairquant Def. 3.15.

2) Equalized Odds For the equalized odds objective, we propose to assess the level of independence on each number of events Y. For this purpose, the notion of demographic parity is required for each subset of value Y.

$$HGR_{EO} = \frac{1}{\#\Omega_Y} \sum_{y \in \Omega_Y} HGR(h_{w_h}(X), S)$$

$$FairQuant_{EO} = \frac{1}{\#\Omega_Y} \sum_{y \in \Omega_Y} \frac{1}{K} \sum_{i=1}^{K} |m_{i,y} - m|$$

with $m_{i,y}$ as the mean of the predictions $h_w(X_k)$ in the k-th quantile set X_k conditioned on $Y = y$, and m its mean on the full sample X. For computational reasons, Ω_Y denotes the set of 'valid' entries. In the context of counts, the standard goodness

of fit test is Pearson's chi-squared test, unfortunately, the latter is not valid if the expected frequencies are too small (as in (Bol'shev and Mirvaliev, 1979) and there is no general agreement on the minimum expected frequency allowed). In practice, if $y \in \{0,1,2,3,4,5,6\}$ with (say) less than 1% for counts (strictly) exceeding 2, it is rather common to consider $\Omega_Y = \{0,1,2^+\}$, where 2^+ denotes the case where counts exceed (strictly) 1.

3.3 | Measuring Individual Fairness

3.3.1 | Fairness Through Awareness

One way to assess individual fairness is to estimate the consistency metric (Mukherjee et al., 2020; Yurochkin et al., 2019), which measures the difference in output prediction between individuals who share the same characteristics except for a specific selected demographic characteristic. For example, in a classification task, (Yurochkin et al., 2019) measures how often output prediction classes change only because of differences in a demographic feature. The implicit idea is to observe if the predictor is invariant under certain sensitive perturbations to the input. We argue that this measure is difficult to use in practice since it strongly depends on which variable is selected. Finding a strongly sensitive dependant feature is not always feasible in practice. For example, consider the individual fairness objective on the Adult data set (Dua and Graff, 2017) that aims to predict income above \$50,000 with gender as a binary sensitive attribute. In this case, (Mukherjee et al., 2020; Yurochkin et al., 2019) compare the predictions following changes in the explanatory variable *Marital Status* which corresponds to information such as Married, Divorced, etc... However, basing all our effort on this sort of proxy is not feasible in practice as there is no real evidence that this variable is the more appropriate. Note that we cannot consider the demographic feature as the sensitive attribute since it is not considered as input to the model in our problem setting.

In order to overcome this limitation, we propose as new contributions, two individual fairness metrics that are dependent on $d_{\mathcal{X}}$ (e.g., obtained via equation 2.5). For $\alpha \in [0,1]$, we denote as q_α the quantile of level α of the set $\{d_{\mathcal{X}}(x_i,x_j), 0 \leq i < j \leq n\}$ and $\tilde{q}_\alpha$ the quantile of level α of the set $\{||x_i - x_j||, 0 \leq i < j \leq n\}$

Definition 10. *We define Mean Region Discrepancy of level α (MRD_α) as:*

$$MRD_\alpha = \frac{\sum_{i<j} |h(x_i) - h(x_j)| \mathbb{1}_{\{d_{\mathcal{X}}(x_i,x_j) \leq q_\alpha\}}}{\sum_{i<j} \mathbb{1}_{\{d_{\mathcal{X}}(x_i,x_j) \leq q_\alpha\}}}$$

Definition 11. *We define Mean Double Region Discrepancy of levels α, β ($MDRD_{\alpha,\beta}$) as:*

$$MDRD_{\alpha,\beta} = \frac{\sum_{i<j} |h(x_i) - h(x_j)| \mathbb{1}_{\{d_{\mathcal{X}}(x_i,x_j) \leq q_\alpha\}} \mathbb{1}_{\{||x_i - x_j|| \geq \tilde{q}_\beta\}}}{\sum_{i<j} \mathbb{1}_{\{d_{\mathcal{X}}(x_i,x_j) \leq q_\alpha\}} \mathbb{1}_{\{||x_i - x_j|| \geq \tilde{q}_\beta\}}}$$

Definition 10 directly encodes the idea that "similar people should be treated similarly": we select pairs of data point that are similar to a predefined level α and measure the mean discrepancies of outputs for these pairs. The smaller the MRD_α, the fairer the algorithm at the individual level. Definition 11 considers an additional $\mathbb{1}_{\{||x_i - x_j|| \geq \tilde{q}_\beta\}}$ factor that eliminates pairs that are already similar in an euclidean sense. Indeed, the predictor h, assuming an adequate choice of activation functions, can be considered as a lipschitz function with respect to euclidean distances, which guarantees closeness of outputs for close couples (in the euclidean sense). By weeding out those pairs, we ensure that we measure discrepancies for relevant data points.

3.3.1.1 | Counterfactual Fairness

There are multiple ways of measuring counterfactual fairness in the literature. We note two different approaches: Some measure separation at a group level, others focus at individual level.

First, there is a statistical parity version at a group level specific for counterfactual fairness. It measures the separation of the generated outcomes from the causal intervention. We denote this metric *TCE* for Total Causal Effect, which is defined as follows:

$$TCE = \mathbb{P}(Y_{A \leftarrow a_1}) - P(Y_{A \leftarrow a_0}) \tag{3.17}$$

This metric is widely used in the literature to evaluate the performance of counterfactual models (often denoted as *Total Effect* in the literature). As discussed in the subsection 2.1, the majority of the state of the art algorithms compute the counterfactual fairness penalization directly on the outcome generation from causal graph (inference phase). They compute afterward a final predictor model on the factual and counterfactual observations as (Kim et al., 2021; Xu et al., 2019). However, depending on the final predictor model, the predictions must be evaluated to ensure fairness. For this purpose, we introduce the Total Predictions Effect (TPE), which refers to the statistical parity on the output prediction from intervention. The metric is defined for a predictor h_{w_h} as follows:

$$TPE = \mathbb{P}(h_{w_h}(X_{A \leftarrow a_1})) - P(h_{w_h}(X_{A \leftarrow a_0})) \tag{3.18}$$

Other approaches (Kim et al., 2021; Xu et al., 2019) evaluate the total effect on specific subgroups, denoted o, generally not dependent on the sensitive. We define them separately using the same procedure, Counterfactual Causal effect: $CPE = \mathbb{P}(Y_{A\leftarrow a_1}|o) - P(Y_{A\leftarrow a_0}|o)$ and Counterfactual Prediction Effect: $CPE = \mathbb{P}(h_{w_h}(X_{A\leftarrow a_1})|o) - P(h_{w_h}(X_{A\leftarrow a_0})|o)$.

Finally we present the counterfactual metric that measure fairness at individual level.

Definition 12. *Counterfactual demographic parity (Kusner et al., 2017)*

The CF measure is defined, for the m_{test} individuals from the data set, as:

$$CF = \frac{1}{m}\sum_{i}^{m}\mathbb{E}_{(x',a')\sim C(i)}[\Delta(h_{w_h}(x_i,a_i),h_{w_h}(x',a'))] \tag{3.19}$$

where Δ is a cost function between two predictions (e.g., the logit paring cost for the binary case and a simple squared difference for the continuous setting). $C(i)$ is the set of counterfactual samples for the i-th individual of the data set. This corresponds to counterfactuals sampled with an inference process.

3.4 | Conclusion

We have shown in this chapter that even when the objective is decided, there are many ways to quantify it. This makes deployment in practice complicated. Why choosing this particular metric with this or that threshold to accept that the model is fair? These are questions that various regulators are currently asking.

Also, some sub-areas, such as the continuous case, have not yet been studied enough. We have proposed new ways to measure the X^2 divergence and the HGR maximum correlation with neural networks that shown to be more suitable in the continuous setting. We will now focus on reducing unwanted biases to improve the level of fairness.

4

Ensuring Group Fairness for Neural Network Predictors

In this chapter, we focus on in-processing fairness and in particular with adversarial learning, which reveals as the most powerful framework for settings where acting on the training process is an option. Since this method relies on penalization during training, we will first describe the general literature on this subject. Then, we will show how penalization is achieved in the fairness framework for the two fairness objectives (demographic parity and those referred to error parity) and this for three following prediction tasks: binary, continuous and frequency task.

Parts of the work presented in this chapter are the subject of two papers: *Fairness-Aware Neural Rényi Minimization for Continuous Features*, published at the IJCAI 2020 conference (Grari et al., 2020a) and *Learning Unbiased Representations via Rényi Minimization*, published at the ECML-PKDD 2021 conference (Grari et al., 2021b).

The contributions are fourfold:

- We have identified a main issue for applying fairness with traditional state-of-the-art adversarial approaches: They are theoretically not able to optimize the most classical fairness criterion such as demographic parity in the continuous case;
- We propose new adversarial approaches based on the minimization of the HGR coefficient. The structures allow mitigation either on the output prediction itself or on latent spaces. The adversaries approximates the HGR by finding non-linear transformations of the data;
- We proposed the first work to compare attenuation at different levels of neural architectures. We argue that operating at intermediate levels of neural repre-

sentations offers the best trade-off between expressiveness and generalization for bias mitigation.

- We demonstrate empirically that our neural HGR-based approach is very competitive for fairness learning on artificial and real-world popular data sets.

The remainder of this chapter proceeds as follows. First, Section 4.1 briefly recaps the principle of adversarial learning. Next, Section 4.2 outlines the biases mitigation algorithms based on the output predictions and Section 4.3 on latent representation. Then in section 4.4 we propose a summary of the different methods. Finally, section 5.3 presents different experimental results of our approaches.

4.1 | Penalization via Adversarial Learning

In this section, we will describe briefly certain types of penalization during the training of the predictor model. One of the most known penalization procedure for a classical machine learning model $h_{w_h}(X)$ is as follow:

$$\underset{w_h}{\operatorname{argmin}}\big\{\mathcal{L}(h_{w_h}(X), Y) + \lambda p(w_h)\big\}$$

for some penalty function p, usually taking into account the complexity of the model, in order to avoid overfitting.

Adversarial machine learning (Goodfellow et al., 2018) are machine learning techniques that aim at robustifying the predictive mode, by attempts to fool models by supplying deceptive input. In order to improve the robustness of the model, it can be natural to consider an adversarial approach such as:

$$\underset{w_h}{\operatorname{argmin}}\Big\{\max_{m,|m|<\epsilon}\big\{\mathcal{L}(h_{w_h}(X), Y) + \lambda \ell(h_{w_h}(X), h_{w_h}(X+m))\big\}\Big\}$$

where we consider the worst-case impact of a small perturbation of the data. The first term represents the classical loss function $\mathcal{L}$ which is minimized in order to improve the accuracy of the predictions. The second term is a penalization that ensures that the prediction between an altered version (the highest perturbation m on X) and the real version will be as similar as possible. A classical example in the context of pictures labeling is the *ostrich* example in (Szegedy et al., 2014), where all pictures (yellow bus, dog, pyramid, insect, etc), slightly perturbed with some noisy picture, still, look as before for a human eye, but are all labeled as an ostrich.

Some other types of adversarial learning which are more related to our work rely on two separate machine learning models trained simultaneously to play against each

other. The most popular one is the Generative Adversarial Networks (GANs) (Goodfellow et al., 2014). The objective is, given a training set, to generate new data following the underlying distribution. The two models are defined as the generator G that captures the data distribution $X \in \mathbb{R}^d$ from a d_z-dimensional random noise $Z \in \mathbb{R}^{d_z}$, and a discriminative model D that estimates the probability that a sample came from the real training data X rather than G.

$$\underset{G}{\operatorname{argmin}}\{ \max_{D} \{\mathbb{E}_{X\sim p_{data}}[\log(D(X)] + \mathbb{E}_{Z\sim p_z}[\log(1 - D(G(Z))]\}\}$$

This framework corresponds to a minimax two-player game. The first term represents the expectation over the real data X from the training distribution p_{data}. This term is only used for the discriminator and allows to assign the larger label to the real data. The second term is expectation over noise. It inputs the noise Z with distribution p_z (e.g., a multivariate normal distribution) to the generator to obtain $G(Z)$. This latter generator's output represents the generated observations and is fed as input to the discriminator for obtaining $D(G(Z))$. While the discriminator attempts to minimize $D(G(Z))$, the generator attempts to fool the discriminator. The generator discovers new samples that the discriminator is not able to distinguish from real data.

4.2 | Adversarial Prediction Retreatment

4.2.1 | Improving Demographic Parity

The fair state-of-the-art algorithms for achieving the demographic parity objective are generally constructed with a penalization term that can be plugged in the following generic optimization problem as below:

$$\arg\min_{w_h} \left\{ \mathcal{L}(h_{w_h}(X), Y) + \lambda p(h_{w_h}(X), S) \right\} \tag{4.1}$$

where $\mathcal{L}$ is the predictor loss function (the mean squared error for regression or log-loss for the binary classification task for example) between the output $h_{w_h}(X) \in \mathbb{R}$ and the corresponding target Y, with h_{w_h} the prediction model which can be for example be GLM or a deep neural network with parameters w_h, and $p(h_{w_h}(X), S)$ the penalization term which evaluates the correlation loss between two variables. The aim is thus to find a mapping $h_{w_h}(X)$ which both minimizes the deviation with the expected target Y and does not imply too much dependency with the sensitive S. The hyperparameter λ controls for impact of the correlation loss in the optimization. The correlation loss p can correspond to a Pearson coefficient, a Mutual Information

Neural Estimation (MINE, (Belghazi et al., 2018)), or our HGR neural estimators that we will discuss below. Our proposed approach to estimate the HGR coefficient with neural networks as shown in the chapter 3 can allow us to mitigate the bias by adversarial learning during the process. We propose two adversarial algorithms that we will describe bellow and we will compare them. We first describe our theoretical contribution for showing the interest of the Rényi adversarial method comparing to the simple adversarial methods which are currently the most used algorithms in the state-of-the-art literature.

4.2.1.1 | Adversarial Simple Architecture

Some approaches (Zhang et al., 2018) assess the level of dependency by considering the ability to reconstruct the sensitive attribute S from the output prediction $h_{w_h}(X)$. By feeding the output prediction as input to an adversary f_{w_f} that takes the form of a GLM or deep neural network with the objective to predict S, it allows to measure the level of dependence during the training. The goal is to obtain a predictor model h_{w_h} whose outputs do not allow the adversarial function to reconstruct the value of the sensitive attribute. If this objective is achieved, the data bias in favor of some demographics disappeared from the output prediction. The predictor with weights w_h has fooled the adversary. The optimization problem is as below:

$$\underset{w_h}{\operatorname{argmin}}\Big\{ \max_{w_f} \Big\{ \mathcal{L}_{\mathcal{Y}}(h_{w_h}(X), Y) - \lambda \mathcal{L}_{\mathcal{S}}(f_{w_f}(h_{w_h}(X)), S) \Big\} \Big\} \tag{4.2}$$

where $\mathcal{L}_{\mathcal{Y}}$ is the predictor loss function between the output $h_{w_h}(X) \in \mathbb{R}$ and the corresponding target Y and $\mathcal{L}_{\mathcal{S}}$ is the adversary loss function between the adversary output $f_{w_f}(h_{w_h}(X)) \in \mathbb{R}$ and the corresponding sensitive attribute S. The hyperparameter λ controls the impact of the dependence loss in the optimization. The prediction $h_{w_h}(X)$ is the input given to the adversarial f_{w_f}. Figure 4.1 gives the architecture of this adversarial learning algorithm. It depicts the predictor function h_{w_h}, which outputs the prediction from X, the adversarial predictor f_{w_f} which seeks at defining the most accurate prediction of S from the output of the predictor function h_{w_h}. Left arrows represent gradients back-propagation. The learning is done via stochastic gradient, alternating steps of adversarial maximization, and global loss minimization. At the end of each iteration, the algorithm updates the parameters of the prediction parameters h_{w_h} by one step of gradient descent. Concerning the adversarial, the back-propagation of the parameters w_f is carried by multiple steps of gradient ascent. This

allows to optimize an accurate estimation of the sensitive attribute at each step, leading to a greatly more stable learning process.

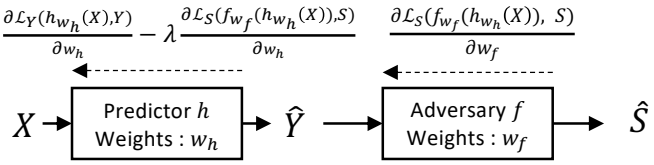


Figure 4.1: The Adversarial simple architecture

Note that for an infinite λ, the second term is the only one to be optimized. Maximizing the negative gradient on the parameter w_f allows minimizing the loss function between the adversary's prediction and the sensitive attribute.

Then, by minimizing this term in a second step via the w_h parameters, it allows removing all the sensitive biases. Also, note that, if there exists (w_h^*, w_f^*) such that $w_f^* = \arg\max_{w_f}\{\mathbb{P}_{w_h^*, w_f}(S|h^*_{w_h}(X))\}$ on the training set, $\mathbb{P}_{w_h^*, w_f^*}(S|h^*_{w_h}(X)) = \widehat{\mathbb{P}}(S)$ and $\mathbb{P}_{w_h^*}(Y|X) = \widehat{\mathbb{P}}(Y|X)$, with $\widehat{\mathbb{P}}(S)$ and $\widehat{\mathbb{P}}(Y|X)$ the corresponding distributions on the training set, (w_h^*, w_f^*) is a global optimum of our min-max problem eq. (4.2). In that case, we have both a perfect classifier in training and a completely fair model since the best possible adversary is not able to predict S more accurately than the estimated prior distribution. While such a perfect setting does not always exist in the data, it shows that the model is able to identify a solution when it reaches one. If a perfect solution does not exist in the data, the optimum of the minimax problem is a trade-off between prediction accuracy and fairness, controlled by the hyperparameter λ.

Theoretical Properties in Continuous Setting

Consider a continuous setting where the sensitive attribute S is a continuous one-dimensional random variable with a regression problem as follows:

$$\inf_{f:\mathbb{R}\to\mathbb{R}} E((S - f(\widehat{Y}))^2) \tag{4.3}$$

The variable that minimizes the quadratic risk is $E(S|\widehat{Y})$. Thus, prediction retreatment algorithms with predictive adversaries (Zhang et al., 2018) (i.e., $\mathcal{L}_S$ as mean square error loss), which consider such optimization problems for mitigating biases, achieve the global fairness optimum when $E(S|\widehat{Y}) = E(S)$. This does **not generally imply demographic parity** when S is continuous. On the other hand, adversarial

approaches based on the HGR_NN (Grari et al., 2020a) achieve the optimum when $HGR(\widehat{Y}, S) = 0$, which is equivalent to demographic parity: $P(\widehat{Y}|S) = P(\widehat{Y})$.

To illustrate this, we consider the maximization problem $\sup_{f:\mathbb{R}\to\mathbb{R}} \rho(f(\widehat{Y}), S)$, which corresponds to the situation where the neural network g is linear in the HGR neural estimator. We have the following result:

Theorem 4.2.1. *If* $E(S|\widehat{Y}) = E(S)$, *then* $\sup_f \rho(f(\widehat{Y}), S) = 0$. *Else,* $f^* \in \arg\max_f \rho(f(\widehat{Y}), S)$ *iff there exists* $a, b \in \mathbb{R}$, *with* $a > 0$, *such that:*

$$f^*(\widehat{Y}) = aE(S|\widehat{Y}) + b \tag{4.4}$$

In other words, the simpler version of the HGR_NN, with g linear, finds the optimal function in terms of regression risk, up to a linear transformation that can be found by simple linear regression. The simplified HGR estimation module therefore captures the exact same non-linear dependencies as the predictive adversary in related work (Zhang et al., 2018). Thanks to the function g, in cases where S cannot be expressed as a function of $\widehat{Y}$ only, the HGR neural network can capture more dependencies than a predictive NN (or equivalently a simplified HGR neural network).

Specific example to understand the difference: Let us consider the following example below where:

$$Y \sim \mathcal{N}(\mu, \sigma^2) \qquad \widehat{Y} = \arctan(S^2) + U\pi \tag{4.5}$$

where $U \perp Y$ and U follows a Bernoulli distribution with $p = \frac{1}{2}$. In this setting, we have $S^2 = \tan(\widehat{Y})$, $HGR(\widehat{Y}, S) = 1$ and due to the hidden variable U, neither $\widehat{Y}$ nor S can be expressed as a function of the other. In that case, the simplified maximal correlation, $\rho(E(S|\widehat{Y}), S)$, has the following bounds, with $\alpha = \frac{\mu}{\sigma}$: $\sqrt{1 - e^{-\frac{\alpha^2}{2}}} \leq \rho(E(S|\widehat{Y}), S) \leq \sqrt{1 - e^{-\frac{\alpha^2}{2}}(1+\alpha^2)^{-\frac{3}{2}}}$. In the degenerate case $\alpha = 0$, we have $E(S|\widehat{Y}) = 0$: the predictive neural network cannot find any dependence. For non-zero values of α, the distribution of S is no longer centered around the axis of symmetry of the square function, so that the prediction becomes possible. However, as shown in the inequality above, the simplified maximal correlation is less than 1, and close to 0 when $\mu \ll \sigma$.

In Figure 4.2, we illustrate the bounds (proof in appendix in section B.1), $\rho(E(S|\widehat{Y}), S)$ being estimated by Monte-Carlo. First, we note that the upper bound is close to $\rho(E(S|\widehat{Y}), S)$, whereas the lower bound $\sqrt{1 - e^{-\frac{\alpha^2}{2}}}$ is not as precise. For non-zero values of α, $\rho(E(S|\widehat{Y}), S)$ is positive, so that a predictive neural network can capture some non-linear dependencies between S and $\widehat{Y}$.

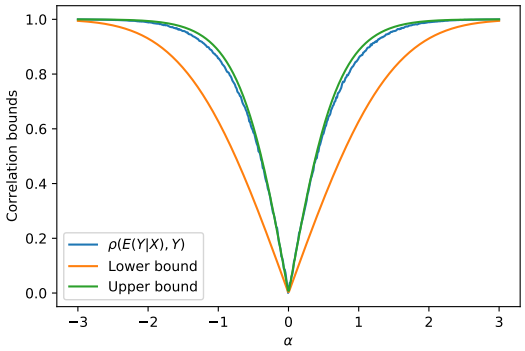


Figure 4.2: Simplified HGR w.r.t α

This is due to the fact that, for $\alpha \neq 0$, the square function is bijective when restricted to some open interval containing the mean of S, whereas when $\alpha = 0$, such an interval cannot be found. When this interval is large and the standard deviation of S is not too large (which corresponds to high values of $|\alpha|$), $\rho(E(S|\widehat{Y}), S)$ approaches 1 and the S prediction error approaches 0. In the opposite case, $\rho(E(S|\widehat{Y}), S)$ is close to 0 and a predictive neural network cannot capture dependencies.

Therefore, as shown by the example, the bilateral approach of the HGR, as opposed to the unilateral approach of predictive models, can capture more dependencies in complex regression scenarios. In adversarial bias mitigation settings, predictive adversaries might not be able to properly detect bias. Adversarial approaches based on the HGR_NN are better fitted for bias mitigation in such continuous complex settings.

4.2.1.2 | Adversarial *HGR* Architecture

We propose a novel adversarial approach based on our *HGR* neural network estimation (Grari et al., 2020a). It uses an adversarial network that takes the form of two inter-connected neural networks for approximating the optimal transformations functions f and g for approximating the sensitive dependence by HGR.

$$\arg\min_{w_h} \left\{ \max_{w_f, w_g} \left\{ \mathcal{L}(h_{w_h}(X), Y) + \lambda \mathbb{E}_{(X,S)\sim\mathcal{D}}(\widehat{f}_{w_f}(h_{w_h}(X))\widehat{g}_{w_g}(S) \right\} \right\} \tag{4.6}$$

where $\mathcal{L}$ is the predictor loss function between the output $h_{w_h}(X) \in \mathbb{R}$ and the corresponding target Y. The second term, which corresponds to the expectation of the products of standardized outputs of both networks ($\hat{f}_{w_f}$ and $\hat{g}_{w_g}$), represents the HGR estimation between the output variable $h_{w_h}(X)$ and the sensitive attribute S. The hyperparameter λ controls the impact of the dependence loss in the optimization.

The prediction $h_{w_h}(X)$ is the input given to the adversarial f_{w_f} and the sensitive S is given as input to the adversarial g_{w_g}. In that case, we only capture for each gradient iteration the estimated HGR between the prediction and the sensitive attribute. Figure 4.3 gives the full architecture of the adversarial learning algorithm using the neural HGR estimator for demographic parity. It depicts the prediction function h_{w_h}, which outputs $\widehat{Y}$ from X, and the two neural networks f_{w_f} and g_{w_g}, which seeks at defining the more strongly correlated transformations of $\widehat{Y}$ and S. The algorithm 3 depicts our Fair Rényi algorithm for the Demographic Parity task. The algorithm takes as input a training set from which it samples batches of size b at each iteration. At each iteration, it first standardizes the output scores of networks f_{w_f} and g_{w_g} to ensure 0 mean and a variance of 1 on the batch. Then it computes the objective function to maximize the estimated HGR score and the global predictor objective. Finally, at the end of each iteration, the algorithm updates the parameters of the HGR adversary w_f and w_g by multiple steps of gradient ascent and the regression parameters w_h by one step of gradient descent.

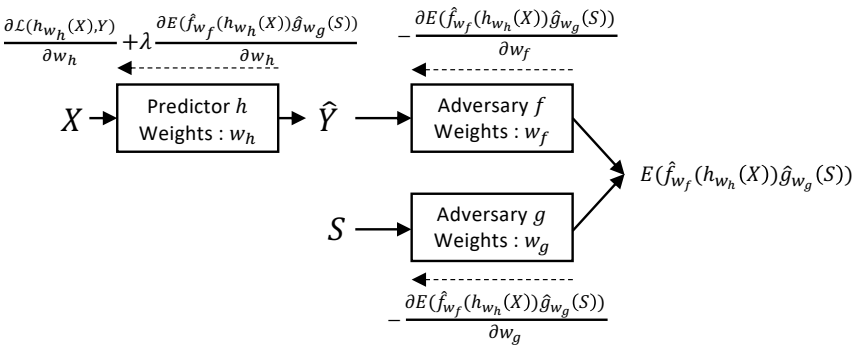


Figure 4.3: The Rényi adversarial algorithm for demographic parity.

Algorithm 3 Fair Rényi Algorithm for Demographic Parity

Input: Training set $\mathcal{T}$, Loss function $\mathcal{L}$, Batchsize b,
Neural Networks h_{ω_h}, f_{ω_f} and g_{ω_g},
Learning rates α_f, α_g and α_h, Fairness control λ

Repeat

Draw b samples $(x_1, s_1, y_1), ..., (x_b, s_b, y_b)$ from $\mathcal{T}$

Calculate the mean and variance of the transformations:

$m_f \leftarrow \frac{1}{b}\sum_{i=1}^{b} f_{\omega_f}(h_{\omega_h}(x_i))$; $m_g \leftarrow \frac{1}{b}\sum_{i=1}^{b} g_{\omega_g}(s_i)$

$\sigma_f^2 \leftarrow \frac{1}{b}\sum_{i=1}^{b}(f_{\omega_f}(h_{\omega_h}(x_i)) - m_f)^2$

$\sigma_g^2 \leftarrow \frac{1}{b}\sum_{i=1}^{b}(g_{\omega_g}(s_i) - m_g)^2$

Standardize the transformations:

$\forall i : \hat{f}_{\omega_f}(h_{\omega_h}(x_i)) \leftarrow \frac{f_{\omega_f}(h_{\omega_h}(x_i)) - m_f}{\sqrt{\sigma_f^2 + \epsilon}}$

$\forall i : \hat{g}_{\omega_g}(s_i) \leftarrow \frac{g_{\omega_g}(s_i) - m_g}{\sqrt{\sigma_g^2 + \epsilon}}$

Compute the objectives:

$J(\omega_f, \omega_g) = \frac{1}{b}\sum_{i=1}^{b} \hat{f}_{\omega_f}(h_{\omega_h}(x_i)) * \hat{g}_{\omega_g}(s_i)$

$L(\omega_h, \omega_f, \omega_g) = \frac{1}{b}\sum_{i=1}^{b} \mathcal{L}(h_{\omega_h}(x_i), y_i) + \lambda J(\omega_f, \omega_g)$

Update the adversary by gradient ascent:

$\omega_f \leftarrow \omega_f + \alpha_f \frac{\partial J(\omega_f, \omega_g)}{\partial \omega_f}$; $\omega_g \leftarrow \omega_g + \alpha_g \frac{\partial J(\omega_f, \omega_g)}{\partial \omega_g}$

Update the predictor model h_{ω_h} by gradient descent:

$\omega_h \leftarrow \omega_h - \alpha_h(\frac{\partial L(\omega_h, \omega_f, \omega_g)}{\partial \omega_h})$

4.2.2 | Improving Equalized Odds

The fair in-processing algorithms for achieving the equalized odds objective are generally constructed with a penalization term that can be plugged in the following generic optimization problem as below:

$$\arg\min_{w_h} \Big\{ \mathcal{L}(h_{w_h}(X), Y) + \lambda p(h_{w_h}(X), S, Y) \Big\} \tag{4.7}$$

where $p(h_{w_h}, Y, S)$ is the penalization term which evaluates the correlation loss between the output prediction and the sensitive attribute given the expected outcome Y. The aim is thus to find a mapping $h_{w_h}(X)$ which both minimizes the deviation with the expected target Y and does not imply too much dependency with the sensitive S given Y.

4.2.2.1 | Adversarial Simple Architecture

Following the idea of adversarial simple architecture for demographic parity, (Zhang et al., 2018) proposes to concatenate the label Y to the output prediction to form the

input vector of the adversary $(h_{w_h}(X), Y)$, so that the adversary function f_{w_f} could be able to output different conditional probabilities depending on the label Y_i of i.

$$\underset{w_h}{\operatorname{argmin}}\Big\{ \max_{w_f} \big\{\mathcal{L}_{\mathcal{Y}}(h_{w_h}(X), Y) - \lambda\big\{\mathcal{L}_{\mathcal{S}}(f_{w_f}(h_{w_h}(X), Y), S)\big\}\big\}$$

Figure 4.4 gives the full architecture of this adversarial learning algorithm for equalized odds. It depicts the predictor function h_{w_h}, which outputs the prediction from X, the adversarial predictor f_{w_f} which seek at defining the most accurate prediction of S from the predictor function h_{w_h} and the targeted variable Y. Left arrows represent gradients back-propagation. Here again, the learning is done via stochastic gradient, alternating steps of adversarial maximization, and global loss minimization.

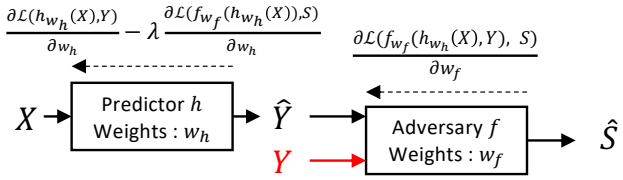


Figure 4.4: The Fair adversarial simple algorithm for equalized odds.

However, as demographic parity, this algorithm can only be considered in binary cases since it does not generally imply equalized odds when S is continuous.

4.2.2.2 | Adversarial HGR Architecture

Whether the sensitive variable is continuous or binary, we propose to extend the idea of our *HGR* adversarial algorithm for equalized odds for binary outcomes Y (Grari et al., 2021c). For the decomposition of disparate mistreatment, we divide the mitigation based on the two different values of Y. Identification and mitigation of the specific non-linear dependence for these two subgroups leads to the same false-positive and the same false-negative rates for each demographic. The optimization is written as follows:

$$\begin{aligned}\arg\min_{w_h} \Big\{ \max_{w_{f_0}, w_{g_0}, w_{f_1}, w_{g_1}} \Big\{ & \mathcal{L}(h_{w_h}(X), Y) \\ & + \lambda_0 \mathbb{E}_{(X,S)\sim\mathcal{D}_0}(\widehat{f}_{w_{f_0}}(h_{w_h}(X))\widehat{g}_{w_{g_0}}(S)) \\ & + \lambda_1 \mathbb{E}_{(X,S)\sim\mathcal{D}_1}(\widehat{f}_{w_{f_1}}(h_{w_h}(X))\widehat{g}_{w_{g_1}}(S)) \Big\}\Big\}\end{aligned}$$

where $\mathcal{D}_0$ corresponds to the distribution of pair (X, S) conditional on $Y = 0$ and $\mathcal{D}_1$ to the distribution conditional on $Y = 1$. The hyperparameters λ_0 and λ_1 control the

impact of the dependence loss for the false positive and the false negative objective respectively.

For instance, the first penalization (controlled by λ_0) enforces the independence between the output prediction $h_{w_h}(X)$ and the sensitive S only for the cases where $Y = 0$. It enforces naturally the mitigation of the difference of false positive rate between demographics since at optimum for w_h^* with $(X, S) \sim \mathcal{D}_0$, $HGR(h_{w_h^*}(X), S) = 0$ and implies theoretically: $h_{w_h^*}(X, S) \perp\!\!\!\perp S|Y = 0$. This idea can be generalized for non-binary outcomes Y as:

$$\arg\min_{w_h} \left\{ \max_{w_{f_0}, w_{g_0}, \ldots, w_{f_K}, w_{g_K}} \left\{ \mathcal{L}(h_{w_h}(X), Y) + \frac{1}{K+1} \sum_{y \in \Omega_Y} \lambda_y \mathbb{E}_{(X,S)\sim\mathcal{D}_y}(\widehat{f}_{w_{f_y}}(h_{w_h}(X))\widehat{g}_{w_{g_y}}(S)) \right\} \right\}$$

where $\mathcal{D}_y$ corresponds to the distribution of pair (X, S) conditional on $Y = y$ and $K = \#\Omega_Y - 1$. The hyperparameters λ_y control the impact of the dependence loss for the objectives. Each penalization enforces the independence between the output prediction and the sensitive S only for the cases where $Y = y$.

This latter objective can be conducted on frequency tasks where the outcome target Y can be represented as a number of events occurring in a fixed interval of time.

4.2.3 | Improving Equalized Residuals

The adversarial approaches structures for the equalized residual objective are quite similar to demographic parity. The idea is that instead of comparing the dependence between the output prediction and the sensible attribute, we compare the observed residuals $R = \widehat{Y} - Y$ against the latter. The penalty compares at each step the dependence term $HGR(\widehat{Y} - Y, S)$ (more information in appendix in section B.2).

4.3 | Extension to Fair Representation

This recent decade, deep learning models have shown very competitive results by learning representations that capture relevant information for the learning task. However, the representation learnt by the deep model may contain some bias from the training data. This bias can be intrinsic to the training data, and may therefore induce a generalisation problem due to a distribution shift between training and testing data. For instance, the color bias in the colored MNIST data set (Kim et al., 2019b) can make models focus on the color of a digit rather than its shape for the classification task. The bias can also go beyond training data, so that inadequate representations can perpetuate or even reinforce some society biases (Bolukbasi et al., 2016) (e.g. gender or age).

The fair representation algorithms are constructed by mitigating the underlying bias on an intermediary latent space Z. We have carried out a specific analysis for comparing this approach to the prediction retreatment methods. We show in the subsection 4.5.3 that mitigation at intermediate levels of encoding can induce the best trade-offs expressiveness/generalization.

4.3.1 | Improving Demographic Parity

The fair representation algorithms for achieving the demographic parity objective are generally constructed with a penalization term that can be plugged in the following generic optimization problem as below:

$$\underset{w_h, w_\psi}{\arg\min} \Big\{ \mathcal{L}(h_{\omega_h}(\phi_{\omega_\phi}(X)), Y) + \lambda p(\phi_{\omega_\phi}(X)), S) \Big\} \tag{4.8}$$

where $\mathcal{L}$ is the predictor loss function between the output prediction $h_{\omega_h}(\phi_{\omega_\phi}(X)) \in \mathbb{R}$ and the corresponding target Y, with h_{ω_h} the predictor neural network with parameters ω_h and $Z = \phi_{\omega_\phi}(X)$ the latent fair representation with ϕ_{ω_ϕ} the encoder neural network, with parameters ω_ϕ.

The second term, $p(\phi_{\omega_\phi}(X), S)$ is the penalization term which evaluates the correlation loss between the latent space and the sensitive feature. Notice that in comparison with the prediction retreatment the correlation loss is evaluated with only multidimentional space since the latent space Z is generally multidimensional, which is not necessarily the case for prediction retreatment where the output feature can be 1-dimensional. Note that by mitigating the underlying bias in the latent space Z, it enforces the demographic parity task. Due to L-Lipschitzness of neural network architectures, we know that $HGR(Z, S) \geq HGR(h_{\omega_h}(Z), S)$. Acting on Z leads to remove bias from Z even for components ignored by the predictor h_{w_h} in train. We argue that this allows to gain in stability at test time, which induces a greater variance of sensitive dependence of the output $\hat{Y}$ (see subsection 4.5.3).

Simple adversarial algorithm extension: Following the idea of (Zhang et al., 2018), (Adel et al., 2019) proposes to extend this methodology in a fair representation way by using a penalization function p that takes the form of a deep neural network. This latter function has the objective of predicting the sensitive feature S by taking the latent space Z as input. Note that the predictor h_{w_h} and the adversarial ϕ_{w_ϕ} models are, as in prediction retreatment, optimized simultaneously in a min-max game. The level of dependence here is assessed by how we can be able to reconstruct S from Z.

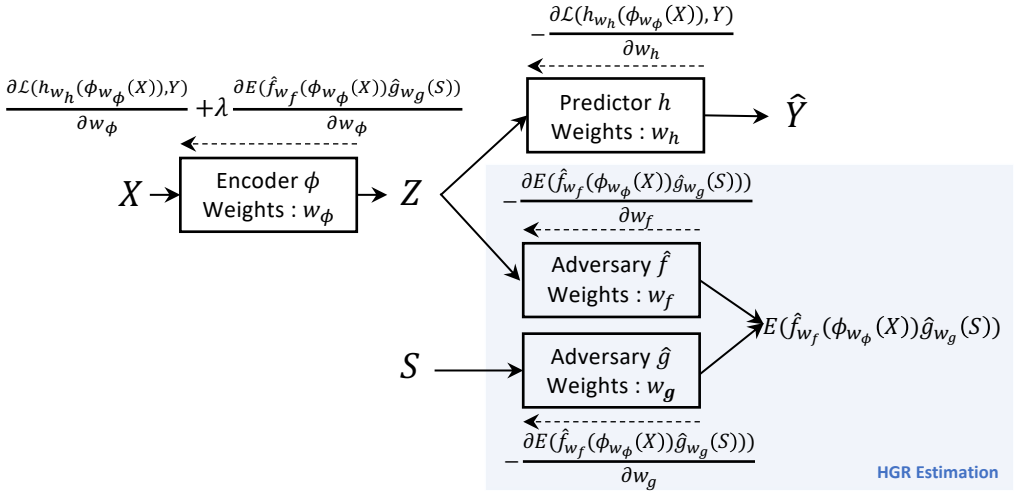


Figure 4.5: Learning Unbiased Representations via Rényi Minimization

Rényi adversarial architecture extension: We propose a method that enforces Demographic Parity via Renyi minimization for a latent variable. The objective is to find a latent representation Z which both minimizes the deviation between the target Y and the output prediction $\widehat{Y}$, provided by a function $h_{w_h}(Z)$, and does not imply too much dependence with the sensitive S. As explained above in section 2.2, our HGR estimation by deep neural network (Grari et al., 2020a) is a good candidate for standing as the adversary $HGR(Z,S)$ to plug in the global objective (4.1). Notice, we can consider the latent representation Z or even the sensitive attribute S as multi-dimensional. This can therefore provide a rich representation of the latent space or even take into account several sensitive features at the same time (for e.g. gender and age or the 3 channels of an image, see 4.5). Please note that our HGR NN estimation can handle multi-dimensional cases for both U and V. The mitigation procedure follows the optimization problem:

$$\min_{w_h, w_\phi} \max_{w_f, w_g} \mathcal{L}(h_{\omega_h}(\phi_{\omega_\phi}(X)), Y) + \lambda E(\widehat{f}_{w_f}(\phi_{\omega_\phi}(X))\widehat{g}_{w_g}(S)) \tag{4.9}$$

where $\mathcal{L}$ is the predictor loss function between the output prediction $h_{\omega_h}(\phi_{\omega_\phi}(X)) \in \mathbb{R}$ and the corresponding target Y, with h_{ω_h} the predictor neural network with parameters ω_h and $Z = \phi_{\omega_\phi}(X)$ the latent fair representation with ϕ_{ω_ϕ} the encoder neural network, with parameters ω_ϕ. The second term, which corresponds to the expectation of the products of standardized outputs of both networks ($\hat{f}_{w_f}$ and $\hat{g}_{w_g}$), represents the HGR estimation between the latent variable Z and the sensitive attribute S.

The hyperparameter λ controls the impact of the correlation loss in the optimization.

Figure 4.5 gives the full architecture of our adversarial learning algorithm using the neural HGR estimator between the latent variable and the sensitive attribute. It depicts the encoder function ϕ_{w_ϕ}, which outputs a latent variable Z from X, the two neural networks f_{w_f} and g_{w_g}, which seek at defining the most strongly correlated transformations of Z and S and the neural network h_{ω_h} which outputs the prediction $\widehat{Y}$ from the latent variable Z. Left arrows represent gradients back-propagation. The learning is done via stochastic gradient, alternating steps of adversarial maximization and global loss minimization. In algorithm 4, we present the pseudo-code for our Rényi Fair Representation algorithm.

Algorithm 4 Rényi Fair Representation

Input: Training set $\mathcal{T}$, Loss function $\mathcal{L}$, Batchsize b, Epochs for HGR n_{HGR}
Neural Networks ϕ_{w_ϕ},h_{w_h}, f_{w_f} and g_{w_g},
Learning rates α_f, α_g, α_ϕ and α_h. Fairness control λ

Repeat
Draw b samples $(x_1, s_1, y_1), ..., (x_b, s_b, y_b)$ from $\mathcal{T}$
Compute the predictor objective:
$L_Y(w_h, \phi_{w_\phi}) = \frac{1}{b}\sum_{i=1}^{b} \mathcal{L}(h_{w_h}(\phi_{w_\phi}(x_i)), y_i)$
Update the predictor model h_{w_h} by gradient descent:
$w_h \leftarrow w_h - \alpha_h(\frac{\partial L_Y}{\partial w_h})$
Repeat n_{HGR} times
Calculate the mean and variance of the transformations:
$m_f \leftarrow \frac{1}{b}\sum_{i=1}^{b} f_{w_f}(\phi_{w_\phi}(x_i))$; $m_g \leftarrow \frac{1}{b}\sum_{i=1}^{b} g_{w_g}(s_i)$
$\sigma_f^2 \leftarrow \frac{1}{b}\sum_{i=1}^{b}(f_{w_f}(\phi_{w_\phi}(x_i)) - m_f)^2$
$\sigma_g^2 \leftarrow \frac{1}{b}\sum_{i=1}^{b}(g_{w_g}(s_i) - m_g)^2$
Standardize the transformations:
$\forall i : \hat{f}_{w_f}(\phi_{w_\phi}(x_i)) \leftarrow \frac{f_{w_f}(\phi_{w_\phi}(x_i)) - m_f}{\sqrt{\sigma_f^2 + \epsilon}}$
$\forall i : \hat{g}_{w_g}(s_i) \leftarrow \frac{g_{w_g}(s_i) - m_g}{\sqrt{\sigma_g^2 + \epsilon}}$
Compute the objectives:
$J(w_f, w_g, w_\phi) = \frac{1}{b}\sum_{i=1}^{b} \hat{f}_{w_f}(\phi_{w_\phi}(x_i)) * \hat{g}_{w_g}(s_i)$
$L_E(w_h, w_\phi, w_f, w_g) = \frac{1}{b}\sum_{i=1}^{b} \mathcal{L}(h_{w_h}(\phi_{w_\phi}(x_i)), y_i) + \lambda J(w_f, w_g, w_\phi)$
Update the adversary by gradient ascent:
$w_f \leftarrow w_f + \alpha_f \frac{\partial J}{\partial w_f}$; $w_g \leftarrow w_g + \alpha_g \frac{\partial J}{\partial w_g}$
Update the encoder model ϕ_{w_ϕ} by gradient descent:
$w_\phi \leftarrow w_\phi - \alpha_\phi(\frac{\partial L_E}{\partial w_\phi})$

4.4 | Summary of the Different Methods

In order to fully understand our contributions against the state of the art, we illustrate in Figure 4.6 the various fair adversarial neural network architectures existing in the literature (including two of our contributions in the middle). We have named by ourselves these different sub-domains (Prediction retreatment, etc.). We distinguish two fair adversarial families:

- Fair Representation: The mitigation is carried on an intermediary latent variable Z. The multidimensional latent variable is fed to the adversary and to the predictor.

- Prediction Retreatment: The mitigation is carried on the prediction itself. The prediction is fed to the adversary.

For these two families, we distinguish 3 subfamilies:

- Simple Adversarial: The adversary tries to predict the sensitive attribute. The bias is mitigated by fooling this adversary.

- Rényi Adversarial: The adversary tries to find adequate non-linear transformations for the estimation of the HGR coefficient. The bias is mitigated via the minimization of this estimation.

- F-divergence Adversarial: The Mutual Information Neural Estimator (Belghazi et al., 2018) or our chi-square divergence neural estimation is used as adversary. The bias is mitigated via the minimization of the dual-representation of the f-divergence (Donsker-Varadhan representation for mutual information and representation of Theorem 3.8 for χ^2 divergence).

4.5 | Empirical Results

4.5.1 | Synthetic Scenario

In order to test the efficiency of our different algorithms, we set up three synthetic scenarios.

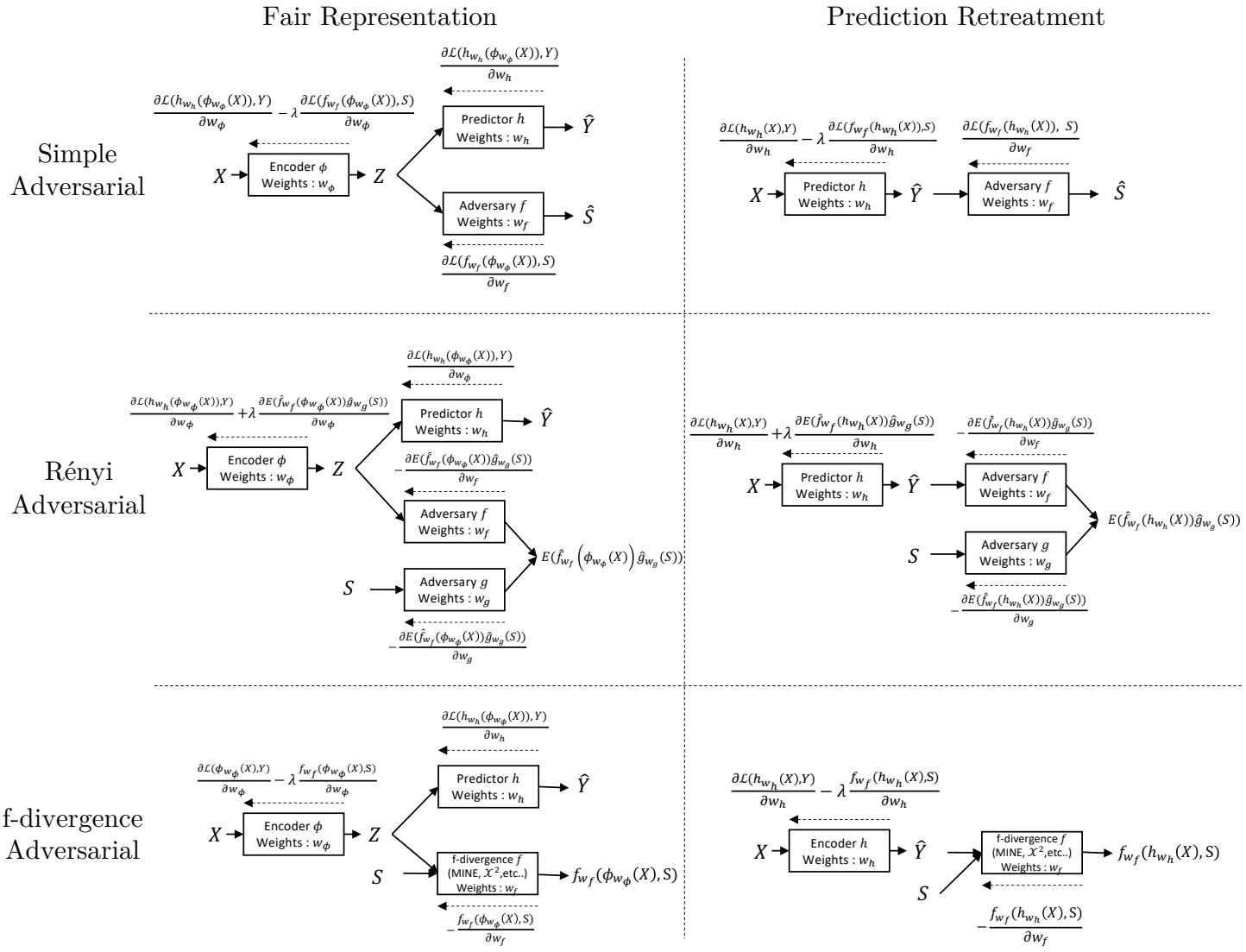


Figure 4.6: Fair adversarial architectures for Demographic Parity

Scenario 1

The subject of this first scenario is a pricing algorithm for a fictional household insurance policy. The goal of this exercise is to achieve demographic parity by producing a fair predictor which estimates the average cost without incorporating any bias against the policyholder's age. We want to compare our proposed algorithm (*Rényi adversar-*

ial on prediction retreatment) with a classical neural network called *Standard NN*). We create three explicit variables: Age of the policyholder, total surface and age of the building. We consider the policyholder's age as sensitive attribute and we construct the average cost variable Y with the last two variables only (without the sensitive variable). To evaluate this, we create the target variable Y with an exponential function which takes into account the two explicit variables mentioned above. The surface variable is a polynomial transformation of age. This transformation is chosen such that no linear correlation exists between surface and age (Pearson correlation = 0.00). On the other hand, it is expected that the HGR coefficient will be non-zero for the Standard NN (estimated to 62%). We report below details on the distributions used in this synthetic scenario:

$$
\begin{aligned}
Age \sim & \mathcal{N}(40,5); \epsilon_1 \sim \mathcal{N}(0,1); \epsilon_2 \sim \mathcal{N}(0,1) \\
Surface = & -0.25 * (-Age + 40)^2 + 150 + 5 * \epsilon_1 \\
BldgAge \sim & \mathcal{N}(30,10) \\
Y = & 100 + 0.0005 * e^{(0.06 * Surface + 0.1 * BldgAge + 0.2 * \epsilon_2)}
\end{aligned}
$$

In order to solve this problem and, thus, to minimize the non-linear dependence between the age and the predictions we execute different scenario and use specific hyperparameters λ for each of them. For each scenario, we repeat five experiments by randomly sampling two subsets, 70% for the training set and 30% for the test set. The choice of this value depends on the main goal, resulting in a trade-off between accuracy and fairness. In figure 4.7, we see clearly that higher values of λ produce fairer predictions, while a specific hyperparameter λ near 0 allows to only focus on optimizing the predictor loss. We note a MSE error gap of 700 between $\lambda = 0$ and $\lambda = 125$. Choosing λ between 25 and 50 appears to be an interesting choice for this scenario.

In Figure 4.8, the blue curves represent the predictions of the Standard NN. The quadratic link between the prediction and the sensitive attribute age can be easily observed. As expected, increasing λ leads to predictions almost stable, around a price of about 226 euros.

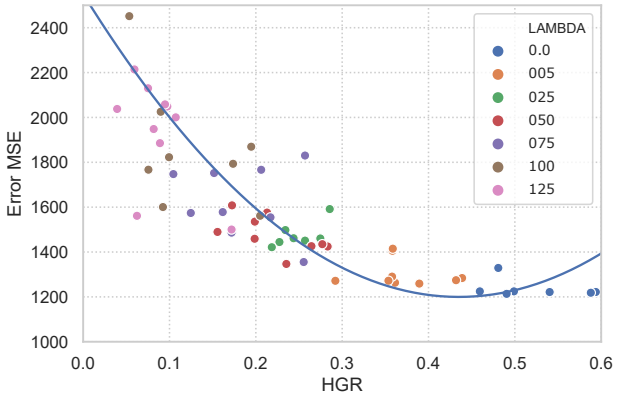


Figure 4.7: Impact of hyperparameter λ. Higher values of λ produce fairer predictions, while λ near 0 corresponds to only optimizing the regression predictor.

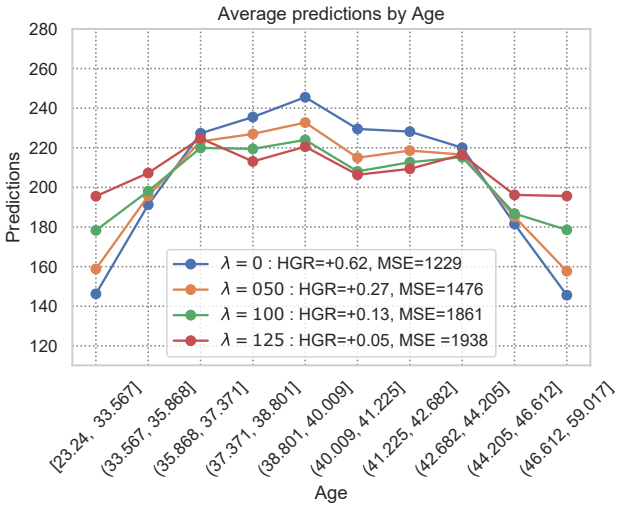


Figure 4.8: Average predictions by age. Higher value of lambda tends to decrease the nonlinear quadratic link between the output predictions and the sensitive feature.

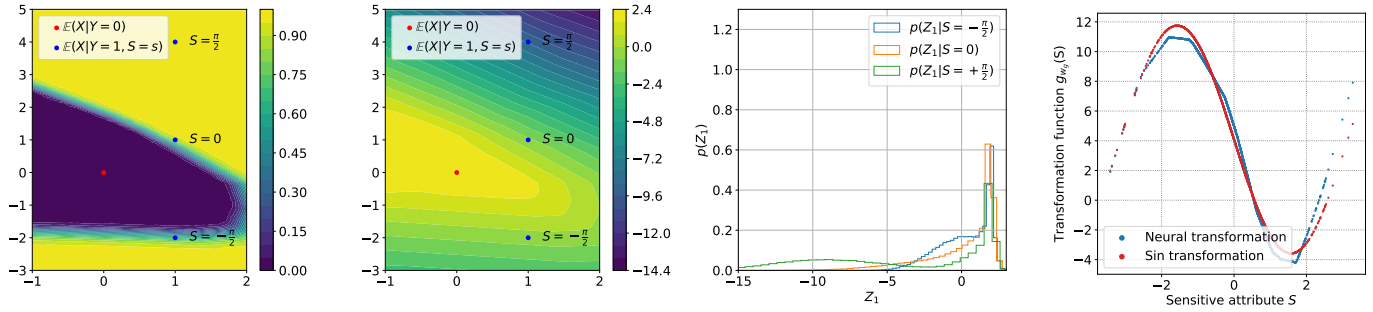
Scenario 2

Inspired by (Louppe et al., 2017), we consider the following toy scenario in a binary target and continuous standard gaussian sensitive attribute setting:

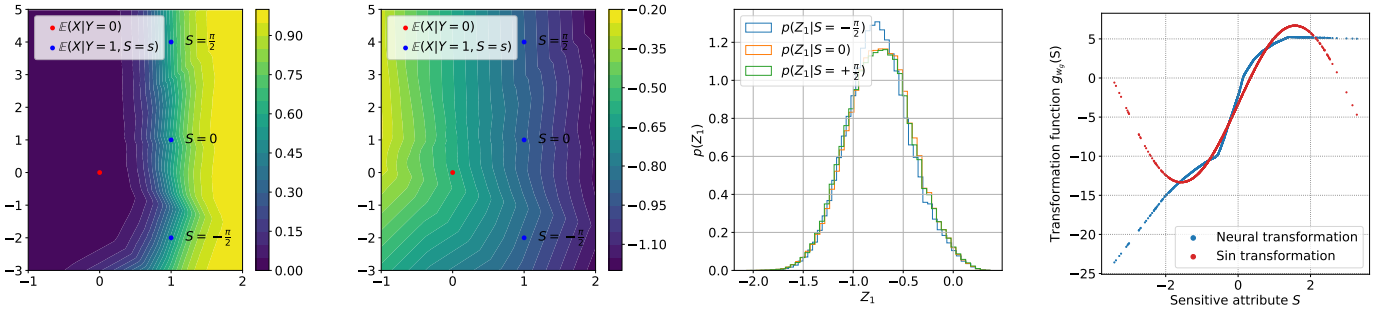
$$X|S=s \sim \mathcal{N}\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & -\frac{1}{2}\\ -\frac{1}{2} & 1\end{pmatrix}\right] \qquad \text{when } Y=0, \tag{4.10a}$$

$$X|S=s \sim \mathcal{N}\left[\begin{pmatrix}1\\1+3\sin s\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right] \qquad \text{when } Y=1 \tag{4.10b}$$

Our goal is to learn a representation Z of the input data that is no longer biased



(a) Biased model: $\lambda = 0$; $HGR(Z,S) = 52\%$; $HGR(\widehat{Y},S) = 30\%$; $Acc = 79\%$



(b) Unbiased model: $\lambda = 13$; $HGR(Z,S) = 5\%$; $HGR(\widehat{Y},S) = 4\%$; $Acc = 68\%$

Figure 4.9: Toy example. (Left) Decision surface in the (X_1, X_2) plane. The figure (a) shows the decision surface for a biased model focused on a prediction loss. $\widehat{Y}$ values are highly correlated with S, samples with S around $\frac{\pi}{2}$ and $Y = 1$ being easier to classify than those with S between $-\frac{\pi}{2}$ and 0. The figure (b) shows decision surfaces for our fair model. These are vertical, meaning that only X_1 influences the classification, and therefore $\widehat{Y}$ is no longer biased w.r.t S. (Middle left) Z_1-slices in the (X_1, X_2) plane. The comparison between the figure below and above highlights the fact that adversarial training allows to create an unbiased representation Z. (Middle right) Conditional probability densities of Z_1 at $S = -\frac{\pi}{2}, 0, \frac{\pi}{2}$. With $\lambda = 0$, the densities are dependent on S, whereas they are not anymore with adversarial training. (Right) In blue, the function modeled by the neural network g in the HGR Neural Network. In red, the closest linear transformation of $\sin(S)$ to $g(S)$.

w.r.t S, while still accurately predicting the target value Y. Figure 4.9 compares the results of both a biased model (a) with a hyperparameter $\lambda = 0$ and an unbiased model (b) with $\lambda = 13$ applied on the toy scenario data. In the context of the Rényi Minimization method, it is interesting to observe the maximal correlation functions learnt by the adversary. When $\lambda = 0$, the adversary with sensitive attribute input models the sin function up to a linear transformation, which also maximizes the correlation with the input data as shown in (4.10b). In that case, the representation Z still carries the bias of X w.r.t S, in the same sin shape. When $\lambda = 13$, the neural network g is unable to find the sin function, which seems to indicate that the representation Z

does not carry the bias w.r.t S anymore. This is confirmed by the low HGR coefficient between Z and S, the Z_1-slices as well as the conditional densities of Z_1 at different values of S. Not only does the adversarial induce an unbiased representation, it also leads to an almost completely unbiased target $\widehat{Y}$, as shown by the vertical decision surfaces and the 4% HGR between $\widehat{Y}$ and S. This is at the cost of a slight loss of accuracy, with an 11% decrease.

Scenario 3

Before considering real-world experiments, we follow the MNIST experimental setup defined by (Kim et al., 2019b), which considers a digit classification task with a color bias planted into the MNIST data set (LeCun et al., 2010; Kim et al., 2019a). In the training set, ten distinct colors are assigned to each class. More precisely, for a given training image, a color is sampled from the isotropic normal distribution with the corresponding class mean color, and a variance parameter σ^2. For a given test image, a mean color is randomly chosen from one of the ten mean colors, without considering the test label, and a color is sampled from the corresponding normal distribution (with variance σ^2). Seven transformations of the data set are designed with this protocol, with seven values of σ^2 equally spaced between 0.02 and 0.05. A lower value of σ^2 implies a higher color bias in the training set, making the classification task on the testing set more difficult, since the model can base its predictions on colors rather than shape. The sensitive feature, color, is encoded as a vector with 3 continuous coordinates. We compare in this experiment algorithms based on prediction retreatment with the simple adversarial (Zhang et al., 2018) and our Rényi adversarial (Grari et al., 2020a). We also add in this comparison the algorithms based on fair representation with the mutual information (Ragonesi et al., 2020), the adversarial simple (Kim et al., 2019b) and our Rényi adversarial (Grari et al., 2021b). For each algorithm and for each data set, we obtain the best hyperparameters by grid search in five-fold cross validation.

Results, in terms of accuracy, can be found in Table 4.1. Notice, the state-of-the-art obtains different results than reported in (Ragonesi et al., 2020) because we consider a continuous sensitive feature and not a 24-bit binary encoding. Our adversarial algorithm via fair representation achieves the best accuracy on the test set for the seven scenarios. The most important gap is for the smallest sigma where the generalisation is the most difficult. The larger number of degrees of freedom carried by the two functions f and g made it possible to capture more unbiased information than the other algorithms on the multidimensional variables Z and S.

	Color variance						
Training	$\sigma = 0.020$	$\sigma = 0.025$	$\sigma = 0.030$	$\sigma = 0.035$	$\sigma = 0.040$	$\sigma = 0.045$	$\sigma = 0.050$
ERM ($\lambda = 0.0$)	0.476 ± 0.005	0.542 ± 0.004	0.664 ± 0.001	0.720 ± 0.010	0.785 ± 0.003	0.838 ± 0.002	0.870 ± 0.001
MI FR (Ragonesi et al., 2020)	0.592 ± 0.018	0.678 ± 0.015	0.737 ± 0.028	0.795 ± 0.012	0.814 ± 0.019	0.837 ± 0.004	0.877 ± 0.010
Simple PR (Zhang et al., 2018)	0.584 ± 0.034	0.625 ± 0.033	0.709 ± 0.027	0.733 ± 0.020	0.807 ± 0.013	0.803 ± 0.027	0.831 ± 0.027
Simple FR (Kim et al., 2019b)	0.645 ± 0.015	0.720 ± 0.014	0.787 ± 0.018	0.827 ± 0.012	0.869 ± 0.023	0.882 ± 0.019	0.900 ± 0.012
Rényi PR (Grari et al., 2020a)	0.571 ± 0.014	0.655 ± 0.022	0.721 ± 0.030	0.779 ± 0.011	0.823 ± 0.013	0.833 ± 0.026	0.879 ± 0.010
Rényi FR (Grari et al., 2021b)	$\mathbf{0.730} \pm 0.008$	$\mathbf{0.762} \pm 0.021$	$\mathbf{0.808} \pm 0.011$	$\mathbf{0.838} \pm 0.010$	$\mathbf{0.878} \pm 0.011$	$\mathbf{0.883} \pm 0.012$	$\mathbf{0.910} \pm 0.007$

Table 4.1: MNIST with continuous color intensity

4.5.2 | Real-World Experiments

Data sets

Our experiments on real-world data are performed on five data sets. First, we experiment with three data sets where the sensitive attribute and the target are both continuous:

- The US Census demographic data set is an extraction of the 2015 American Community Survey 5-year estimates. It contains 37 information features about 74,000 American census tracts. Our goal is to predict the percentage of children below the poverty line. We consider gender as a sensitive attribute encoded as the percentage of the women in the census tract.
- The Motor Insurance data set originates from a pricing game organized by The French Institute of Actuaries in 2015 (The Institute of Actuaries of France, 2015). The data set contains a total of 15 attributes for 36,311 observations. The task is to predict the average claim cost of third-party material claims per policy. As the sensitive attribute we use the driver's age.
- The Crime data set is obtained from the UCI Machine Learning Repository (Dua and Graff, 2017). This data set includes a total of 128 attributes for 1,994 instances from communities in the US. The task is to predict the number of violent crimes per population for US communities. As the sensitive attribute we use the race information with the ratio of an ethnic group per population.

We experiment with two data sets with a binary classification task where the sensitive features are continuous:

- Compas: The COMPAS data set (Angwin et al., 2016) contains 13 attributes of about 7,000 convicted criminals with class labels that state whether or not the individual reoffended within 2 years of their most recent crime. Here, we use age as sensitive attribute.

- Default: The Default data set (Yeh and Lien, 2009) contains 23 features about 30,000 Taiwanese credit card users with class labels which state whether an individual will default on payments. As sensitive attribute, we use age.

Fairness Algorithms

As a baseline, we use a classic, "unfair" deep neural network, Standard NN. We compare our different approaches with state-of-the-art algorithms. We compare fair prediction retreatment methods based on f-divergences (lower right in Figure 4.6) with the mutual information (Grari et al., 2020a), the χ^2 divergence estimated by KDE (Mary et al., 2019) [2] and by neural network (Grari et al., 2020a). In addition, still in the fair prediction retreatment category, we compare the simple adversarial algorithm (Zhang et al., 2018) (upper right in Figure 4.6) and our Rényi adversarial (Grari et al., 2021b) (middle right in Figure 4.6). Also, we include methods based on fair representation with the mutual information (Ragonesi et al., 2020) (lower left in Figure 4.6), simple adversarial (Adel et al., 2019) (upper left in Figure 4.6) and our Rényi algorithm (Grari et al., 2021b) (middle left in Figure 4.6).

Metrics and Experimental Conditions

In a same setting, the hyperparameter λ of the different approaches allows us to balance the relative importance of accuracy and fairness. For each algorithm and for each data set, we attempted to obtain comparable results by giving similar predictive performance of the models. We obtain the best hyperparameters by grid search in five-fold cross validation (specific to each of them), 80% for the training set and 20% for the test set. Depending on the task, we parameterized the number of layers between 3 and 5 and between 8 and 32 for the number of units. For all the different fair representation algorithms, we assign the latent space with only one hidden layer with 64 units. Mean normalization was applied to all the outcome true values in the continuous cases. We used Tanh activation functions, Dropout and Xavier initialization. The considered regression loss is MSE. Notice, we applied a mean normalization to the different outcome true value. Finally, we report the average of the mean squared error (MSE), the accuracy (ACC) and the mean of the fairness metrics HGR_NN (Grari et al., 2020a), HGR_KDE (Mary et al., 2019), HGR_RDC (Lopez-Paz et al., 2013) and MINE (Belghazi et al., 2018) on the test set. Since none of these fairness measures are fully reliable (they are only estimations which are used by the compared models), we also use the *FairQuant* metric (Grari et al., 2020a), based on the quantization of the test samples in 50 quantiles w.r.t. to the sensitive attribute. As discussed in section 3.2.4, this metric corresponds to the mean absolute difference

between the global average prediction and the mean prediction of each quantile.

Results and Discussion

As expected, the baseline, Standard NN, is the best predictor but also the most biased one. It achieves the lowest prediction errors and ranks amongst the highest and thus worst values for all fairness measures throughout all data sets and tasks.

For demographic parity, the results of our experiments can be found in Table 4.2. While being better in terms of predictive performance, our Rényi Fair Representation algorithm (*Rényi FR*) achieves on four data sets (except on the Crime data set) the best level of fairness assessed by HGR estimation, MINE, and FairQuant. On the Crime data set, the approach by χ^2_{KDE} (Mary et al., 2019) gets slightly better results but with a very high volatility. Note, the simple adversarial on Fair Representation (Adel et al., 2019) obtains (except on the Crime data set) better results than simple adversarial on fair prediction retreatment (Zhang et al., 2018).

For equalized residuals, Table 4.3, our approach based on the Rényi minimization achieves (except on the Crime data set, too) the best result with the lowest values for the metric FairQuant. The approach based on χ^2_{KDE} (Mary et al., 2019) performs slightly worse. For the mutual information estimation approach, except on the UC Census data set, it achieves worse results in fairness and accuracy. Globally, our neural approach based on the Rényi minimization, appears to be very competitive in every setting.

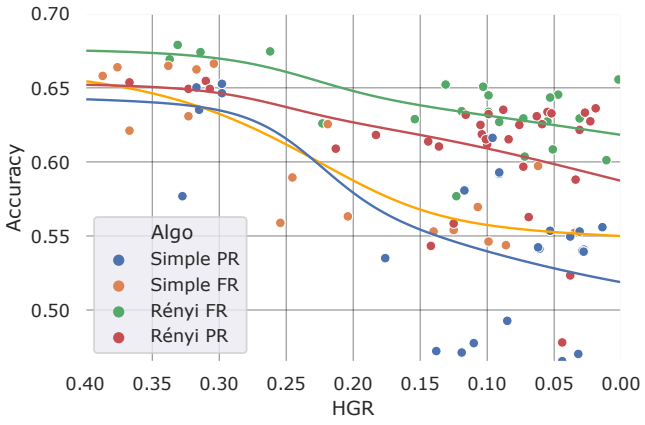


Figure 4.10: Impact of hyperparameter λ (COMPAS data set): Higher values of λ produce fairer predictions.

[2] https://github.com/criteo-research/continuous-fairness

		MSE	HGR_NN	HGR_KDE	HGR_RDC	MINE	FairQuant
US Census	Standard NN	0.274 ± 0.003	0.212 ± 0.094	0.181± 0.00	0.217 ± 0.004	0.023 ± 0.018	0.059 ± 0.00
	Rényi PR (Grari et al., 2020a)	0.526 ± 0.042	0.057 ± 0.011	0.046 ± 0.030	0.042 ± 0.038	**0.001** ± 0.001	0.008 ± 0.015
	χ^2_{KDE} PR (Mary et al., 2019)	0.541 ± 0.015	0.075 ± 0.013	0.061 ± 0.006	0.078 ± 0.013	0.002 ± 0.001	0.019 ± 0.004
	χ^2_{NN} PR (Grari et al., 2020a)	0.535 ± 0.039	0.069 ± 0.037	0.048 ± 0.027	0.044 ± 0.032	0.002 ± 0.001	**0.008** ± 0.013
	MI PR (Grari et al., 2020a)	0.537 ± 0.046	0.058 ± 0.042	0.048 ± 0.029	0.045 ± 0.037	**0.001** ± 0.001	0.012 ± 0.016
	Simple FR (Adel et al., 2019)	0.552 ± 0.032	0.100 ± 0.028	0.138 ± 0.042	0.146 ± 0.031	0.003 ± 0.003	0.035 ± 0.011
	Simple PR (Zhang et al., 2018)	0.727 ± 0.264	0.097 ± 0.038	0.135 ± 0.036	0.165 ± 0.028	0.009 ± 0.005	0.022 ± 0.019
	Rényi FR (Grari et al., 2021b)	**0.523** ± 0.035	**0.054** ± 0.015	**0.044** ± 0.032	**0.041** ± 0.031	**0.001** ± 0.001	**0.007** ± 0.002
Motor	Standard NN	0.945 ± 0.011	0.201 ± 0.094	0.175 ± 0.0	0.200 ± 0.034	0.188 ± 0.005	0.008 ± 0.011
	Rényi PR (Grari et al., 2020a)	0.971 ± 0.004	0.072 ± 0.029	0.058 ± 0.052	**0.066** ± 0.009	**0.000** ± 0.000	0.006 ± 0.02
	χ^2_{KDE} PR (Mary et al., 2019)	0.979 ± 0.119	0.077 ± 0.023	0.059 ± 0.014	0.067 ± 0.028	0.001 ± 0.001	0.006 ± 0.002
	χ^2_{NN} PR (Grari et al., 2020a)	0.975 ± 0.003	0.076 ± 0.016	0.067 ± 0.0005	0.067 ± 0.015	0.001 ± 0.001	0.005 ± 0.0002
	MI PR (Grari et al., 2020a)	0.982 ± 0.003	0.078 ± 0.013	0.068 ± 0.004	0.069 ± 0.009	**0.000** ± 0.000	**0.004** ± 0.001
	Simple FR (Adel et al., 2019)	0.979 ± 0.003	0.101 ± 0.04	0.09 ± 0.03	0.101 ± 0.04	0.002 ± 0.002	0.009 ± 0.004
	Simple PR (Zhang et al., 2018)	0.998 ± 0.004	0.076 ± 0.034	0.091 ± 0.024	0.129 ± 0.08	0.001 ± 0.001	**0.004** ± 0.001
	Rényi FR (Grari et al., 2021b)	**0.962** ± 0.002	**0.070** ± 0.011	**0.055** ± 0.005	0.067 ± 0.006	**0.000** ± 0.000	**0.004** ± 0.001
Crime	Standard NN	0.384 ± 0.012	0.732 ± 0.013	0.525 ± 0.013	0.731 ± 0.009	0.315 ± 0.021	0.353 ± 0.006
	Rényi PR (Grari et al., 2020a)	0.781 ± 0.016	**0.356** ± 0.063	0.097 ± 0.022	**0.171** ± 0.03	**0.009** ± 0.008	**0.039**± 0.008
	χ^2_{KDE} PR (Mary et al., 2019)	**0.778** ± 0.103	0.371 ± 0.116	0.115 ± 0.046	0.177 ± 0.054	0.024 ± 0.015	0.064 ± 0.023
	χ^2_{NN} PR (Grari et al., 2020a)	0.785 ± 0.028	0.385 ± 0.068	0.106 ± 0.024	0.184 ± 0.020	0.020 ± 0.031	0.123 ± 0.012
	MI PR (Grari et al., 2020a)	0.782 ± 0.034	0.395 ± 0.097	0.110 ± 0.022	0.201 ± 0.021	0.032 ± 0.029	0.136 ± 0.012
	Simple FR (Adel et al., 2019)	0.836 ± 0.005	0.384 ± 0.037	0.170 ± 0.027	0.371 ± 0.035	0.058 ± 0.027	0.057 ± 0.007
	Simple PR (Zhang et al., 2018)	0.787 ± 0.134	0.377 ± 0.085	0.153 ± 0.056	0.313 ± 0.087	0.037 ± 0.022	0.063 ± 0.046
	Rényi FR (Grari et al., 2021b)	0.783 ± 0.031	0.369 ± 0.074	**0.087** ± 0.031	0.173 ± 0.044	0.011 ± 0.006	0.043 ± 0.012
		ACC	HGR_NN	HGR_KDE	HGR_RDC	MINE	FairQuant
COMPAS	Standard NN	68.7% ± 0.243	0.363 ± 0.005	0.326 ± 0.003	0.325 ± 0.008	0.046 ± 0.028	0.140 ± 0.001
	Rényi PR (Grari et al., 2020a)	59.7% ± 2.943	0.147 ± 0.000	0.121 ± 0.002	0.101 ± 0.007	0.004 ± 0.001	0.018 ± 0.018
	MI PR (Grari et al., 2020a)	54.4% ± 7.921	0.134 ± 0.145	0.123 ± 0.111	0.141 ± 0.098	0.014 ± 0.023	0.038 ± 0.050
	Simple FR (Adel et al., 2019)	55.4% ± 0.603	0.118 ± 0.022	0.091 ± 0.012	0.097 ± 0.034	0.006 ± 0.007	0.013 ± 0.016
	Simple PR (Zhang et al., 2018)	51.0% ± 3.550	0.116 ± 0.000	0.081 ± 0.003	0.086 ± 0.010	0.002 ± 0.003	**0.010** ± 0.005
	Rényi FR (Grari et al., 2021b)	**60.2**% ± 3.076	**0.063** ± 0.024	**0.068** ± 0.018	**0.067** ± 0.014	**0.001** ± 0.002	0.011 ± 0.018
Default	Standard NN	82.1% ± 0.172	0.112 ± 0.013	0.067 ± 0.010	0.089 ± 0.014	0.002 ± 0.001	0.015 ± 0.002
	Rényi PR (Grari et al., 2020a)	79.9% ± 2.100	0.082 ± 0.015	0.075 ± 0.019	0.072 ± 0.010	0.001 ± 0.001	0.007 ± 0.007
	MI PR (Grari et al., 2020a)	80.1% ± 2.184	0.093 ± 0.020	0.057 ± 0.002	0.066 ± 0.012	0.001 ± 0.001	0.008 ± 0.001
	Simple FR (Adel et al., 2019)	79.2% ± 1.207	0.054 ± 0.025	0.048 ± 0.015	0.064 ± 0.009	0.001 ± 0.001	0.005 ± 0.002
	Simple PR (Zhang et al., 2018)	77.9% ± 9.822	0.052 ± 0.017	**0.044** ± 0.013	0.056 ± 0.004	**0.000** ± 0.000	**0.004** ± 0.000
	Rényi FR (Grari et al., 2021b)	**80.8**% ± 0.286	**0.041** ± 0.008	**0.044** ± 0.006	**0.047** ± 0.002	0.001 ± 0.002	0.005 ± 0.001

Table 4.2: Results for Demographic Parity Best performance among fair algorithms in bold.

		Equalized Residuals				
		MSE	HGR_NN	HGR_KDE	HGR_RDC	FairQuant
US Census	Standard NN	0.274 ± 0.003	0.157 ± 0.006	0.098 ± 0.002	0.122 ± 0.002	0.008 ± 0.001
	Rényi PR (Grari et al., 2020a)	**0.334** ± 0.021	**0.068** ± 0.019	0.053 ± 0.04	**0.055** ± 0.046	**0.003** ± 0.002
	χ^2_{KDE} PR (Mary et al., 2019)	0.408 ± 0.004	0.092 ± 0.017	**0.049** ± 0.003	0.063 ± 0.005	0.009 ± 0.001
	χ^2_{NN} PR (Grari et al., 2020a)	0.384 ± 0.012	0.084 ± 0.021	0.054 ± 0.042	0.057 ± 0.022	0.006 ± 0.004
	MI PR (Grari et al., 2020a)	0.406 ± 0.021	0.083 ± 0.017	0.055 ± 0.017	0.082 ± 0.015	0.008 ± 0.006
Motor	Standard NN	0.945 ± 0.015	0.145 ± 0.005	0.102 ± 0.038	0.123 ± 0.041	0.075 ± 0.006
	Rényi PR (Grari et al., 2020a)	**0.991** ± 0.021	**0.102** ± 0.007	0.082 ± 0.008	**0.092** ± 0.009	**0.011** ± 0.015
	χ^2_{KDE} PR (Mary et al., 2019)	1.019 ± 0.01	0.111 ± 0.007	**0.079** ± 0.005	0.098 ± 0.005	0.015 ± 0.011
	χ^2_{NN} PR (Grari et al., 2020a)	1.011 ± 0.012	0.114 ± 0.011	0.081 ± 0.008	0.077 ± 0.007	0.014 ± 0.0010
	MI PR (Grari et al., 2020a)	1.024 ± 0.017	0.121 ± 0.022	0.091 ± 0.007	0.092 ± 0.005	0.031 ± 0.009
Crime	Standard NN	0.384 ± 0.024	0.472 ± 0.036	0.244 ± 0.01	0.440 ± 0.011	0.047 ± 0.004
	Rényi PR (Grari et al., 2020a)	0.583 ± 0.044	0.382 ± 0.089	**0.151** ± 0.017	0.222 ± 0.045	**0.028** ± 0.006
	χ^2_{KDE} PR (Mary et al., 2019)	**0.579** ± 0.074	**0.381** ± 0.097	0.152 ± 0.035	**0.221** ± 0.068	0.048 ± 0.035
	χ^2_{NN} PR (Grari et al., 2020a)	0.581 ± 0.069	0.353 ± 0.092	0.142 ± 0.029	0.211 ± 0.038	0.042 ± 0.028
	MI PR (Grari et al., 2020a)	0.583 ± 0.054	0.413 ± 0.15	0.161 ± 0.027	0.232 ± 0.018	0.052 ± 0.013

Table 4.3: Results for Equalized Residuals in terms of accuracy (MSE) and fairness metrics.

Impact of mitigation weight In Figure 4.10, we plot the performance of different scenarios by displaying the HGR against the Accuracy with different values of the

hyperparameter λ. This plot was obtained on the COMPAS data set with 4 algorithms: our two Rényi approaches, the Simple Fair Representation (Adel et al., 2019) and the Simple Prediction Retreatment (Zhang et al., 2018). The different curves is obtained by Nadaraya-Watson kernel regression (Bierens, 1988) between the Accuracy of the model and the HGR. Varying the hyperparameter λ allows to control the fairness/accuracy trade-off. Here, we clearly observe for all algorithms that the Accuracy, or predictive performance, decreases when fairness increases. Higher values of λ produce fairer predictions w.r.t the HGR, while near 0 values of the hyperparameter λ result in the optimization of the predictor loss with no fairness consideration (dots in the upper left corner of the graph). We note that, for all levels of predictive performance, our two Rényi approaches outperforms the state of the art algorithms. Also, we note that algorithms based on fair representation outperforms prediction retreament (except some for simple adversarial where the level of fairness is very low close to the most biased model). We will conduct an experiment below to analyze this results.

4.5.3 | Fair Representation Compared to Prediction Retreatment

In order to further analyze the benefits of mitigation in neural representations compared to prediction retreatments, we propose to consider various architectures of predictors h and encoders ϕ, with adversarial HGR mitigation being applied on the output of the encoder as depicted in figure 4.6. To get comparable results between settings, we consider a constant full architecture (encoder + predictor), composed of 5 layers with 4 hidden layers with 32 units each.

In figure 4.11, we compare on the COMPAS dataset 5 different settings where mitigation is applied on a different layer of this full architecture: *LayerX* corresponds to a setting where mitigation is applied on the output of layer X (encoder of X layers, predictor of 5-X layers). *Layer5* thus corresponds to our Rényi prediction retreatment approach , *Rényi PR* (no predictor function, the encoder function h directly outputs the prediction). *Layer3* is the standard setting used for our approach in the previous results reported above (*Rényi FR*). As in Figure 4.10, plotted results correspond to fairness-accuracy trade-offs obtained with different values of λ. We notice that applying mitigation too early in the architecture (Layer1) leads to very poor results. This can be explained by the fact that for this simple encoding setting, the encoder expressiveness is to weak to effectively remove non-linear dependencies w.r.t. the sensitive attribute, without removing too much useful information for prediction. At the contrary, when mitigation is applied late in the architecture (*Layer4* and *Layer5*)

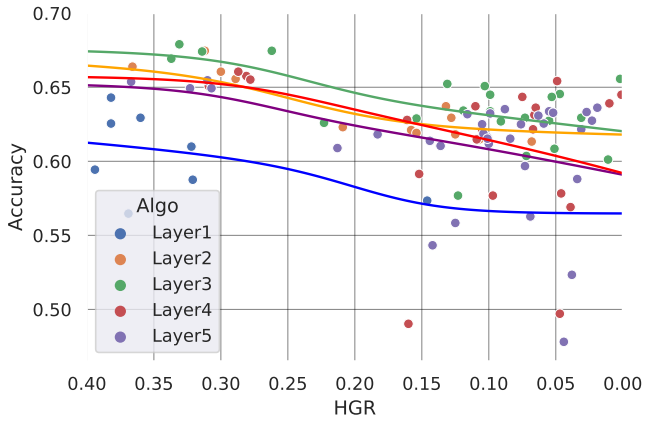


Figure 4.11: Impact of hyperparameter λ (COMPAS data set) for various encoders ϕ and predictors h.

we observe generalization limits of the approach. While results on the training set are similar to those of *Layer3*, these settings lead to predictions at test time that are more dependent on the sensitive attribute. Due to L-Lipschitzness of neural network architectures, we know that $HGR(Z,S) \geq HGR(h_{w_h}(Z),S)$. Acting on Z leads to remove bias from Z even for components ignored by the predictor ϕ in train. However, we argue that this allows to gain in stability at test time, when such components can be activated for new inputs, compared to late approaches, such as *Layer4* or *Layer5*, which induce a greater variance of sensitive dependence of the output $\hat{Y}$. Mitigation at intermediate levels, such as *Layer3*, appears to correspond to the best trade-off expressiveness/generalization.

4.6 | Conclusion

In this chapter we have identified a main issue for applying fairness for any continuous sensitive features: The traditional state-of-the-art adversarial algorithm are theoretically not able to optimize the most classical fairness objective as demographic parity. To address this issue we leverage our HGR maximal correlation, which has shown to be very efficient in capturing non-linear dependencies and for debiasing a predictor model with adversarial learning. We have theoretically showed the interest of using this fairness metric as a penalization term compared to the simple adversarial methods and in particular for the continuous case. We investigated empirically why fair adversarial representation methods can give better results. For this purpose,

we have compared mitigation at different levels of neural architectures. We argue that acting at intermediary levels of neural representations allows the best trade-off between expressiveness and generalisation for bias mitigation. For further investigation, we will apply this architecture for information bottleneck purposes (e.g., for data privacy), which might be improved with an HGR_NN penalization as suggested in (Asoodeh et al., 2015).

5

Ensuring Group Fairness for Gradient Tree Boosting Predictors

This chapter focuses on fairness for another family of ML methods: decision tree predictors. So far, we have mainly focused on linear regression or neural networks, which constitute the bulk of the Fair-ML research community. However, ensemble methods combining several decision tree classifiers have proven very efficient for various applications. In practice for tabular data sets, actuaries and data scientists prefer the use of gradient tree boosting over neural networks due to its generally higher accuracy rates (Zhang et al., 2017). Our purpose in this chapter is the development of fair classifiers based on decision trees. We propose a novel approach to combine the strength of gradient tree boosting with an adversarial fairness constraint.

Most of the work presented in this chapter was the subject of the paper: *Fair Adversarial Gradient Tree Boosting*, published at the ICDM 2019 conference (Grari et al., 2019).

The contributions are threefold:

- To the best of our knowledge, we propose the first adversarial learning method for generic classifiers, including decision trees;
- We apply adversarial learning for fair classification on decisions trees;
- We empirically compare our proposal and its variants with several state-of-the-art approaches, for two different fairness metrics. Experiments show the great performance of our approach.

The remainder of this chapter proceeds as follows. First, Section 5.1 briefly recaps the principle of classical gradient tree boosting. Next, section 5.2 outlines a novel

algorithm which combines gradient tree boosting with adversarial debiasing. Finally, section 5.3 presents experimental results of our approach.

This present work will focus on the group fairness objective in a binary setting. The targeted sensitive attribute and the actual value of the outcome are both binary ($(S, Y) \in [0, 1]$)). Please note that this work could be extended in the continuous case using the techniques in the previous chapter.

5.1 | Gradient Tree Boosting

In order to establish the basis for our approach and also to introduce our notation, we first summarize the principle of classical gradient tree boosting. The "Gradient Boosting Machine" (GBM) constitutes a prediction model for regression and classification problems based on an ensemble technique where multiple weak learners are combined to produce a strong learner (Friedman, 2001). Often, such weak learners are decision trees, generally of the type Classification And Regression Tree (CART). In this case, the algorithm is called gradient tree boosting (GTB). The weak learners are built sequentially. Eventually, a strong classifier is obtained as a weighted sum of the weak learners. The classical gradient method is used to optimize the model for any differentiable loss function.

The objective of the GBM is to find a good estimate of the function F which approximately minimizes the empirical loss function:

$$\min_{F} \sum_{i=1}^{n} \mathcal{L}(y_i, F(x_i)) \tag{5.1}$$

where the loss function $\mathcal{L}(y_i, F(x_i))$ measures the i-th prediction compared to the true label. In the classical version of the GBM, the prediction corresponding to a feature vector x is given by an additive model of the form

$$F_M(x_i) = \sum_{m=0}^{M} \gamma_m h_m(x_i) \tag{5.2}$$

where M is the total number of iterations, and $h_m(x_i)$ corresponds to a weak learner at step m in the form of a greedy CART prediction.

The main steps for fitting the model are shown as pseudo code in Algorithm 5. The method exploits the fact that the residual corresponds to the negative gradient of the loss function. Thus, we calculate at each step m the so-called "pseudo residuals":

$$r_{im} = -\left[\frac{\partial \mathcal{L}(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad \text{for } i = 1, \ldots, n \tag{5.3}$$

In order to update the model, we fit a new weak learner $h_m(x)$ to those pseudo residuals and add it to the current model. This step is repeated until the algorithm converges.

Algorithm 5 Classical Gradient Boosting

Input: Training set $(x_i, s_i, y_i)_{i=1}^n$, a number of iterations M, a differentiable loss function $\mathcal{L}(y, F(x))$

Initialize: Calculate the constant value:

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} \mathcal{L}(y_i, \gamma)$$

for $m = 1$ **to** $M - 1$ **do**

(a) Calculate the pseudo residuals:

$$r_{im} = -\left[\frac{\partial \mathcal{L}(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad \text{for } i = 1, \ldots, n$$

(b) Fit a classifier $h_m(x)$ to pseudo residuals using the training set $(x_i, r_{im})\}_{i=1}^n$

(c) Compute multiplier γ_m by solving the following one-dimensional optimization problem:

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} \mathcal{L}\left(y_i, F_{m-1}(x_i) + \gamma * h_m(x_i)\right)$$

(d) Update the model:

$$F_m(x_i) = F_{m-1}(x_i) + \gamma_m * h_m(x_i)$$

end for

5.2 | Fair Adversarial Gradient Tree Boosting (FAGTB)

Our aim is to learn a classifier that is both effective for predicting true labels and fair, in the sense that it cares about metrics defined in section 3.1 for demographic parity or equalized odds. The idea is to leverage the great performance of GTB for classification, while adapting it for fair machine learning via adversarial learning.

5.2.1 | Min-Max formulation

The GTB processes sequentially by gradient iteration (see Section 5.1). This architecture allows us to apply for fair classification with decision tree algorithms the concept of adversarial learning, which corresponds to a two-player game with two contradictory components, such as in generative adversarial networks (GAN) (Goodfellow et al., 2014). In the vein of (Zhang et al., 2018; Louppe et al., 2017; Wadsworth et al., 2018) for fair classification, we consider a predictor function F, that outputs the probability of an input vector X for being labelled $Y = 1$, and an adversarial model A which tries to predict the sensitive attribute S from the output of F. Depending on the accuracy rate of the adversarial algorithm, we penalize the gradient of the GTB at each iteration. The goal is to obtain a classifier F whose outputs do not allow the adversarial function to reconstruct the value of the sensitive attribute. If this objective is achieved, the data bias in favor of some demographics disappeared from the output prediction.

The predictor and the adversarial classifiers are optimized simultaneously in a min-max game defined as:

$$\arg\min_{F} \max_{\theta_A} \sum_{i=1}^{n} \mathcal{L}_{F_i}(F(x_i)) - \lambda \sum_{i=1}^{n} \mathcal{L}_{A_i}(F(x_i); \theta_A) \tag{5.4}$$

where $\mathcal{L}_{F_i}$ and $\mathcal{L}_{A_i}$ are respectively the predictor and the adversary loss for the training sample i given $F(x_i) \in \mathbb{R}$, which refers to the output of the GTB predictor for input x_i. The hyperparameter λ controls the impact of the adversarial loss.

The targeted classifier outputs the label $\widehat{Y}$ which maximizes the posterior $P(\widehat{Y}|X)$. Thus, for a given sample x_i, we get:

$$\hat{y}_i = \arg\max_{y \in \{0;1\}} p_F(Y = y | X = x_i) \tag{5.5}$$

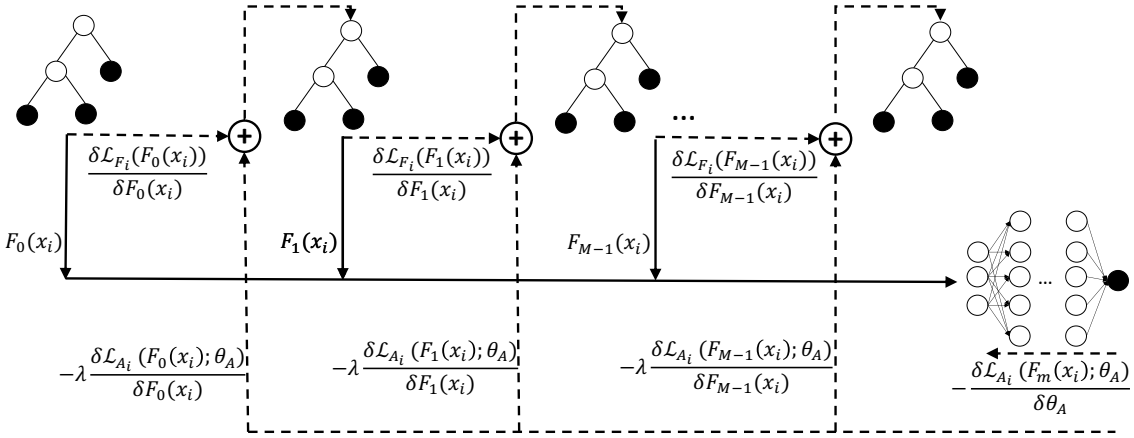


Figure 5.1: The architecture of the Fair Adversarial Gradient Tree Boosting (FAGTB). 4 steps are depicted, each one corresponding to a tree h that is added to the global classifier F. The neural network on the right is the adversary that tries to predict the sensitive attributes from the outputs of the classifier. Solid lines represent forward operations, while dashed ones represent gradient propagation. At each step m, gradients from the prediction loss and the adversary loss are summed to form the target for the next decision tree h_{m+1}.

where $p_F(Y = 1|X = x_i) = \sigma(F(x_i))$, with σ denoting the sigmoid function. Therefore, $\mathcal{L}_{F_i}$ is defined as a logistic regression loss:

$$\begin{aligned} \mathcal{L}_{F_i}(F(x_i)) &= -\log p_F(Y = y_i|X = x_i) \\ &= -\mathbf{1}_{y_i=1} \log(\sigma(F(x_i))) \\ &\quad -\mathbf{1}_{y_i=0} \log(1 - \sigma(F(x_i))) \end{aligned} \tag{5.6}$$

where $\mathbf{1}_{cond}$ equals 1 if *cond* is true, 0 otherwise.

The adversary A corresponds to a neural network with parameters θ_A, which takes as input the sigmoid of the predictor's output for any sample i (i.e., $P_F(Y = 1|X = x_i)$), and outputs the probability P_{F,θ_A} for the sensitive to equal 1:

- For the demographic parity task, $P_F(Y = 1|X = x_i)$ is the only input given to the adversary for the prediction of the sensitive attribute s_i. In that case, the network A outputs the conditional probability $P_{F,\theta_A}(S = 1|V = v_i) = A(v_i)$, with $V = (\sigma(F(X)))$.

- For the equalized odds task, the label y_i is concatenated to $P_F(Y = 1|X = x_i)$ to form the input vector of the adversary $v_i = (\sigma(F(x_i)), y_i)$, so that the function A could be able to output different conditional probabilities $P_{F,\theta_A}(S = 1|V = v_i)$ depending on the label y_i of i.

The adversary loss is then defined for any training sample i as:

$$\begin{aligned}\mathcal{L}_{A_i}(F(x_i);\theta_A) = &- \mathbf{1}_{s_i=1}\log(\sigma(A(v_i))) \\ &- \mathbf{1}_{s_i=0}\log(1-\sigma(A(v_i)))\end{aligned} \tag{5.7}$$

with v_i defined according to the task as detailed above.

Note that, for the case of demographic parity, if there exists (F^*, θ_A^*) such that $\theta_A^* = \arg\max_{\theta_A} P_{F^*,\theta_A}(S|V)$ on the training set, $P_{F^*,\theta_A^*}(S|V) = \widehat{P}(S)$ and $P_{F^*}(Y|X) = \widehat{P}(Y|X)$, with $\widehat{P}(S)$ and $\widehat{P}(Y|X)$ the corresponding distributions on the training set, (F^*, θ_A^*) is a global optimum of our min-max problem eq. (5.4). In that case, we have both a perfect classifier in training, and a completely fair model since the best possible adversary is not able to predict S more accurately than the estimated prior distribution. Similar observations can easily be made for the equalized odds task (by replacing $\widehat{P}(S)$ by $\widehat{P}(S|Y)$ and using the corresponding definition of V in the previous assertion). While such a perfect setting does not always exists in the data, it shows that the model is able to identify a solution when it reaches one. If a perfect solution does not exists in the data, the optimum of our min-max problem is a trade-off between prediction accuracy and fairness, controlled by the hyperparameter λ.

5.2.2 | Learning

The learning process is outlined as pseudo code in Algorithm 6. The algorithm first initializes the classifier F_0 with constant values for all inputs, as done for the classical GBT. Additionally, it initializes the parameters θ_A of the adversarial neural network A (a Xavier initialization is used in our experiments). Then, at each iteration m, beyond calculating the pseudo residuals r_{im} for any training sample i w.r.t. the targeted prediction loss $\mathcal{L}_{F_i}$, it computes pseudo residuals t_{im} for the adversarial loss $\mathcal{L}_{A_i}$ too. Both residuals are combined in $u_{im} = r_{im} - \lambda * t_{im}$, where λ controls the impact of the adversarial network. The algorithm then fits a new weak regressor h_m (a decision tree in our work) to residuals using the training set $\{(x_i, u_{im})\}_{i=1}^n$. This pseudo-residuals regressor is supposed to correct both prediction and adversarial biases of the previous classifier F_{m-1}. It is added to it after a line search step, which determines the best γ_m weight to assign to h_m in the new classifier F_m. Finally, the adversarial has to adapt its weights according to new outputs (i.e., using the training set $\{(F_m(x_i), s_i)\}_{i=1}^n$). This is done by gradient backpropagation. A schematic representation of our approach can be found in Figure 5.1.

Algorithm 6 Fair Adversarial Gradient Tree Boosting

Input: training set $(x_i, s_i, y_i)_{i=1}^{n}$, a number of iterations M, an adversarial learning rate α, a differentiable loss function $\mathcal{L}_F$ for the output classifier and $\mathcal{L}_A$ for the adversarial classifier.

Initialize: Calculate the constant value:

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} \mathcal{L}_{F_i}(\gamma)$$

Initialize parameters θ_A of the neural network $A(x)$

for $m = 1$ **to** $M - 1$ **do**

(a) Calculate the pseudo residuals:

$$r_{im} = -\left[\frac{\partial \mathcal{L}_{F_i}(F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad \text{for } i = 1, \ldots, n$$

(b) Calculate the pseudo residuals of the adversarial from the input $F_{m-1}(x_i)$:

$$t_{im} = -\left[\frac{\partial \mathcal{L}_{A_i}(F(x_i; \theta_A))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad \text{for } i = 1, \ldots, n$$

(c) Calculate the training loss derivative: $u_{im} = r_{im} - \lambda * t_{im}$

(d) Fit a classifier $h_m(x)$ to pseudo residuals using the training set $\{(x_i, u_{im})\}_{i=1}^{n}$

(e) Compute multiplier γ_m by solving the one-dimensional optimization problem:

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} \mathcal{L}_{F_i}\left(F_{m-1}(x_i) + \gamma * h_m(x_i)\right) - \lambda * \mathcal{L}_{A_i}(F_{m-1}(x_i) + \gamma * h_m(x_i); \theta_A).$$

(f) Update the learning model:

(g) Fit the adversarial A to the using the new outputs (i.e., using the training set $\{(F_m(x_i), s_i)\}_{i=1}^{n}$)

$$\theta_A := \theta_A - \alpha * \frac{\partial \mathcal{L}_{A_i}(F_m(x_i); \theta_A)}{\partial \theta_A}$$

end do

5.3 | Empirical Results

We evaluate the performance of our algorithm empirically with respect to regression accuracy and fairness. We conduct the experiments on a synthetic scenario, but also on real-world data sets. Finally, we compare the results with state-of-the-art algorithms.

Synthetic Scenario

We illustrate the fundamental functionality of our proposal with a simple toy scenario which was inspired by the Red Car example (Kusner et al., 2017). The subject is a pricing algorithm for a fictional car insurance policy. The purpose of this exercise is to train a fair classifier which estimates the claim likelihood without incorporating any gender bias. We want to demonstrate the effects of an unfair model versus a fair model.

We focus on the general claim likelihood and ignore the severity or cost of the claim. Further, we only consider the binary case of claim or not (as opposed to a frequency). We assume that the claim likelihood only depends on the aggressiveness and the inattention of the policyholder. To make the training more complex, these two properties are not directly represented in the input data but only indirectly available through correlations with other input features. We create a binary label Y with no dependence with the sensitive attribute S. Concretely, we use as features the protected attribute *gender* of the policyholder, and the unprotected attributes *color* of the car, and *age* of the policyholder. In our data distribution, the *color* of the car is strongly correlated with both *gender* and aggressiveness. The *age* is not correlated with *gender*. However, the *age* is correlated with the inattention of the policyholder. Thus, the latter input feature is actually linked to the claim likelihood. First, we generate the training samples $(x_i, s_i, y_i)_{i=1}^n$. The unprotected attributes $x_i = (c_i, a_i)$ represent the *color* of the car and the *age* of the policyholder, respectively. s is the protected variable *gender*. y is the binary class label, where $y = 1$ indicates a registered claim. As stated above, we do not use the two features aggressiveness (A) and inattention (I) as input features but only to construct the data distribution which reflects the claim likelihood. In order to make it more complex, we add a little noise ϵ_i. These training samples are generated as follows:

$$
\begin{aligned}
s_i &\sim \mathcal{B}(0,1) \\
\begin{pmatrix} I_i \\ a_i \end{pmatrix} &\sim \mathcal{N}\left[\begin{pmatrix} 0 \\ 40 \end{pmatrix}, \quad \begin{pmatrix} 1 & 4 \\ 4 & 20 \end{pmatrix}\right] \\
A_i &\sim \mathcal{N}(0,1) \\
c_i &= (1.5 * s_i + A_i) > 1 \\
y_i &= \sigma(A_i + I_i + \epsilon_i) > 0.5 \\
\epsilon_i &\sim \mathcal{N}(0, 0.1)
\end{aligned}
$$

where $\mathcal{B}(p)$ denotes the bernoulli distribution with probability p and $\mathcal{N}(\mu, \sigma)$ the normal distribution with μ and σ as mean and standard deviation parameters.

A correlation matrix of the distribution is shown in Table 5.1.

We execute first a classical GTB algorithm. In Figure 5.3, on the top-left graph, we can see the curves of accuracy and the fairness metric p-rule during the training phase. Even though there is no obvious link with the sensitive attribute, we notice that this model is unfair (p-rule of 67%). In fact, the outcome observations Y depend exclusively on A and I which should have no dependence with the sensitive feature S. To reconstruct the aggressiveness, the classifier has to consider the color of the car. Unfortunately, it incorporates the sensitive information too, resulting in a claim likelihood prediction one and a half times more for men than for women (1/0.67).

To solve this problem, we also plot in Figure 5.3 curves for our FAGTB model with 5 different values of λ, optimized for demographic parity. We observe that λ efficiently controls fairness against accuracy, with a p-rule that increases to 1 (perfectly fair model) for $\lambda \geq 0.016$. Of course this is at the cost of a slight loss of accuracy. Of course this is at the cost of a slight loss of accuracy. We note that gaining 29 points of p-rule leads to a decrease of accuracy of 10 points. To have a better understanding of what is happening when we consider the model as fair in this specific scenario, we plot the permutation feature importance (Breiman, 2001; Fisher et al., 2019) (calculated on the model performance reduction after randomly shuffling a feature) for the fair and the unfair model in Figure 5.2. With higher lambda values, the weight of color, which is indirectly correlated with the sensitive attribute, tends to be cancelled.

Comparison Against the State-of-the-Art

Table 5.1: Correlation matrix of the synthetic scenario

a	1.0				
A	0.01	1.0			
c	-0.01	0.68	1.0		
s	0.0	-0.01	0.36	1.0	
I	0.90	0.01	0.0	0.0	1.0
	a	A	c	s	I

The features are: age (a), agressivity (A), color (c), gender (s), inattention (I)

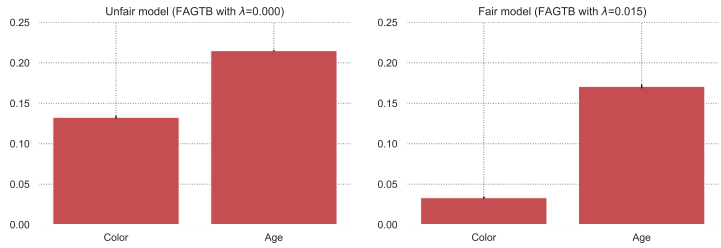


Figure 5.2: Synthetic Scenario: Feature importance for a biased model ($\lambda = 0$) and a fair model ($\lambda = 0.015$) optimized for demographic parity.

Data Sets

For our experiments we use 4 different popular data sets often used in fair classification:

- Adult: The Adult UCI income data set (Dua and Graff, 2017) in this specific experiment the class labels state if the income is higher than $50,000 or not. As sensitive attribute we use gender encoded as a binary attribute, male or female.

- Compas: The COMPAS data set (Angwin et al., 2016) with the same setting as Subsection 4.5

- Default: The Default data set (Yeh and Lien, 2009) with the same setting as Subsection 4.5

- Bank: The bank marketing data set (Moro et al., 2014) contains 16 features about 45,211 clients of a Portuguese banking institution. The goal is to predict if the client has subscribed or not to a term deposit. We consider the age as sensitive attribute, encoded as a binary attribute, indicating whether the client's age is between 33 and 60 years, or not.

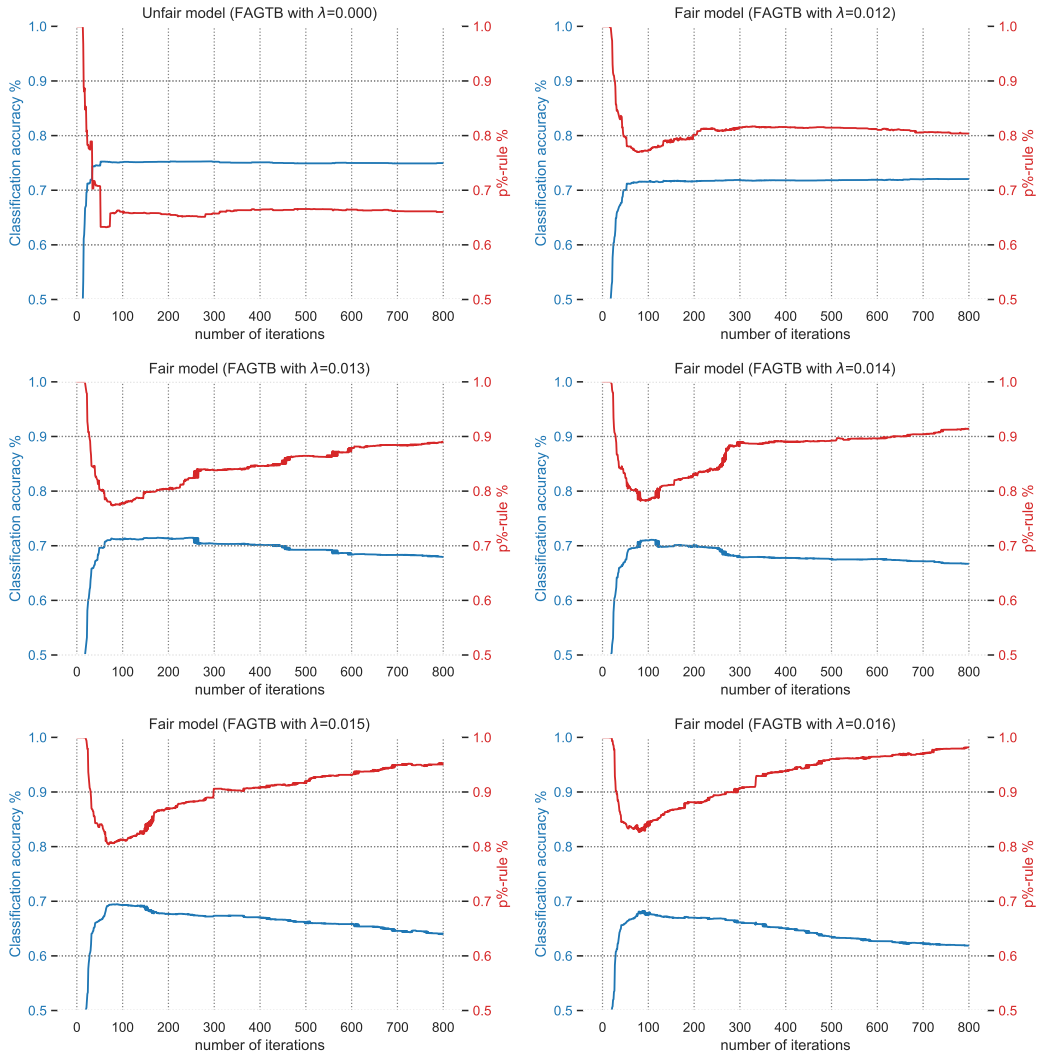


Figure 5.3: Synthetic scenario: Accuracy and p-rule metric for a biased model ($\lambda = 0$) and for several fair models with varying values of λ optimized for demographic parity.

For all data sets, we repeat 10 experiments by randomly sampling two subsets, 80% for the training set and 20% for the test set. Finally, we report the average of the accuracy and the fairness metrics from the test set.

Fairness Algorithms

Because different optimization objectives result in different algorithms, we run separate experiments for the two fairness metrics of our interest, demographic parity (Table 5.2) and equalized odds (Table 5.3). More specifically, for demographic parity we aim at a p-rule of 90% for all algorithms and then compare the accuracy. Optimiz-

ing for equalized odds, results are more difficult to compare. In order to be able to compare the accuracy, we have done our best to obtain, each time, a disparate level below 0.03.

As a baseline, we use a classical, "unfair" gradient tree boosting algorithm, Standard GTB, and a deep neural network, Standard NN.

Further, to evaluate if the complexity of the adversarial network has an impact on the quality of the results, we compare a simple logistic regression adversarial, FAGTB-1-Unit, with a complex deep neural network, FAGTB-NN. In this latter, the adversarial architecture consists of 3 hidden layers with 64, 32, and 16 units, respectively, and ReLU activations.

In addition to the algorithms mentioned above, we evaluate the following fair state-of-the-art in-processing algorithms: Wadsworth2018 (Wadsworth et al., 2018)[2], Zhang2018 (Zhang et al., 2018)[3], Kamishima (Kamishima et al., 2012)[1] Feldman (Feldman et al., 2015)[1], Zafar-DI (Zafar et al., 2017b)[1] and Zafar-DM (Zafar et al., 2017a)[1].

For each algorithm and for each data set, we obtain the best hyperparameters by grid search in 5-fold cross validation (specific to each of them).

For Standard GTB, we parameterize the number of trees and the maximum tree depth. For example, for the Bank data set, a tree depth of 3 with 800 trees is sufficient. For the Standard NN, we parameterize the number of hidden layers and units with a ReLU function and we apply a specific dropout regularization to avoid overfitting. Further, we use an Adam optimisation with a binary cross entropy loss. For the Adult UCI data set for example, the architecture consists of 2 hidden layers with 16 and 8 units, respectively, and ReLU activations. The output layer comprises one single output node with sigmoid activation.

For FAGTB, to accelerate the learning phase, we decided to sacrifice some performance by replacing the one-dimensional optimization γ_m by a specific fixed learning rate for the classifier predictor. All hyperparameters mentioned above, for trees and neural networks, are selected jointly. For FAGTB-NN, in order to achieve better results, we execute for each gradient boosting iteration several training iterations of the adversarial NN. This produces a more persistent adversarial algorithm. Otherwise, the predictor classifier GTB could dominate the adversary. At the first iteration, we begin with modeling a biased GTB and we then model the adversarial NN based on those biased predictions. This approach allows to have a better weight initialization of the adversarial NN. It is more suitable for the specific bias on the data set.

[1] https://github.com/algofairness/fairness-comparison
[2] https://github.com/equialgo/fairness-in-ml
[3] https://github.com/IBM/AIF360

Without this specific initialization we encountered some cases where the predictor classifier surpasses the adversarial too quickly and tends to dominate from the beginning. Compared to the FAGTB-NN, the adversary of the FAGTB-1-Unit is more simple. In this case, the 2 parameters of the adversarial are chosen randomly and for each gradient boosting iteration only one is computed for the adversarial unit.

Results

For demographic parity (Table 5.2), as expected Standard GTB and Standard NN achieve the highest accuracy. However, they are also the most biased models. For example, the classical gradient tree boosting algorithm achieves a 32.6% p-rule for the Adult UCI data set. In this particular case, the prediction for earning a salary above $50,000 is in average more than three times higher for men than for women. Using such a predictor for setting the salary for any new employee would thus perpetuate this bias for training data.

Comparing the mitigation algorithms, FAGTB-NN achieves the best result with the highest accuracy for a p-rule equality of about 90%. The choice of a neural network architecture for the adversary proved to be in any case better than a simple logistic regression. This is particularly true for the COMPAS data set where, for a similar p-rule, the difference in accuracy is considerable (2.7 points). Recall that for demographic parity the adversarial classifier only has one single input feature which is the output of the prediction classifier. It seems necessary to be able to segment this input in several ways to better capture information relevant to predict the sensitive attribute. The sacrifice of accuracy is less important for the Bank and the Default data set. The dependence between the sensitive attribute and the target label is thus less important than for the COMPAS data set. To achieve a p-rule of 90%, we sacrifice 4.6 points of accuracy (Comparing GTB and FAGTB-NN) for COMPAS, 0.7 points for Default and 0.6 points for Bank.

In Figure 5.5 we plot the distribution of the predicted probabilities for each sensitive attribute S for 3 different models: An unfair model with $\lambda = 0$, and 2 fair FAGTB models with $\lambda = 0.06$ and $\lambda = 0.15$, respectively. For the unfair model, the distribution differs most in lower probabilities. The second graph shows an improvement but there remain some differences. For the final one, the distributions are practically aligned.

For equalized odds, the min-max optimization is more difficult than for demographic parity. The fairness metrics DispFPR and DispFNR (e.q. 3.3 and 3.4 respectively) are not exactly comparable thus we did not succeed to obtain the same level

Table 5.2: Results for Demographic Parity

	Adult		COMPAS		Default		Bank	
	Accuracy	P-rule	Accuracy	P-rule	Accuracy	P-rule	Accuracy	P-rule
Standard GTB	86.8%	32.6%	69.1%	61.2%	82.9%	77.2%	90.8%	48.1%
Standard NN	85.3%	31.4%	67.5%	71.1%	82.1%	63.3%	90.3%	58.6%
FAGTB-1-Unit	84.4%	90.4%	61.8%	90.1%	81.5%	90.1%	90.1%	90.0%
FAGTB-NN	**84.9%**	90.3%	**64.5%**	90.0%	**82.2%**	90.2%	**90.2%**	90.0%
(Wadsworth et al., 2018)	83.1%	89.7%	63.9%	90.1%	81.8%	90.0%	**90.2%**	90.1%
(Zhang et al., 2018)	83.3%	90.0%	64.1%	89.8%	81.4%	90.0%	90.0%	90.0%
(Zafar et al., 2017c)	82.2%	89.8%	63.9%	89.7%	80.7%	89.8%	89.2%	90.1%
(Kamishima et al., 2012)	82.3%	89.9%	63.8%	90.0%	81.1%	90.0%	89.6%	89.9%
(Feldman et al., 2015)	-	-	61.4%	90.1%	72.2%	90.2%	-	-

Comparing our approach with different common fair algorithms by accuracy and fairness (p-rule metric) for the Adult UCI, the COMPAS, the Default and the Bank data set.

Table 5.3: Results for Equalized Odds

	Adult			COMPAS		
	Accuracy	DispFPR	DispFNR	Accuracy	DispFPR	DispFNR
Standard GTB	86.8%	0.06	0.07	69.1%	0.12	0.20
Standard NN	85.3%	0.07	0.10	67.5%	0.09	0.15
FAGTB-1-Unit	86.3%	0.02	0.02	65.1%	0.03	0.12
FAGTB-NN	**86.4%**	0.02	0.02	**66.2%**	0.01	0.02
(Wadsworth et al., 2018)	84.9%	0.02	0.03	65.4%	0.02	0.03
(Zhang et al., 2018)	84.8%	0.03	0.03	64.9%	0.03	0.02
(Zafar et al., 2017a)	83.9%	0.03	0.09	64.3%	0.09	0.17
(Kamishima et al., 2012)	82.6%	0.06	0.24	63.6%	0.08	0.11
(Feldman et al., 2015)	80.6%	0.07	0.05	61.1%	0.03	0.03

	Default			Bank		
	Accuracy	DispFPR	DispFNR	Accuracy	DispFPR	DispFNR
Standard GTB	82.9%	0.02	0.04	90.8%	0.04	0.06
Standard NN	82.1%	0.02	0.05	90.3%	0.05	0.08
FAGTB-1-Unit	82.1%	0.00	0.01	89.7%	0.02	0.07
FAGTB-NN	**82.5%**	0.00	0.01	**90.3%**	0.01	0.07
(Wadsworth et al., 2018)	81.2%	0.01	0.02	89.4%	0.01	0.07
(Zhang et al., 2018)	81.9%	0.00	0.01	89.8%	0.00	0.07
(Zafar et al., 2017a)	81.0%	0.00	0.03	89.5%	0.01	0.08
(Kamishima et al., 2012)	80.5%	0.00	0.04	89.3%	0.00	0.08
(Feldman et al., 2015)	71.8%	0.02	0.02	87.1%	0.05	0.06

Comparing our approach with different common fair algorithms by accuracy and fairness (DispFPR, DispFNR) for the Adult UCI, the COMPAS, the Default and the Bank data set.

of fairness. However, we notice that the FAGTB-NN achieves better accuracy with a reasonable level of fairness. Concretely, we achieve for the 4 data sets and for both metrics values below 0.02 or less, except for the Bank data set where DispFNR is equal to 0.07. For this data set, most of the state-of-the-art algorithms result in a DispFNR

between 0.06 and 0.08. It proves hard to achieve a low False Negative Rate (FNR) because the target feature deals with an imbalanced problem (11.7% of positive class). A possible way to handle this problem of imbalanced target class could be to add a specific weight directly in the loss function. We also notice that the difference in the results between FAGTB-1-Unit and FAGTB-NN is much more significant, a linear adversarial being not sufficient to predict the sensitive attribute accurately in that case.

In order to better understand the impact of hyperparameter λ, we illustrate its impact on the accuracy and the p-rule metric in Figure 5.4 for the Adult UCI data set. For that, we model the FAGTB-NN algorithm with 10 different values of λ and we run each experiment 10 times. In the graph, we report the accuracy and the p-rule fairness metric, and finally plot a polynomial regression of second order to demonstrate the general effect.

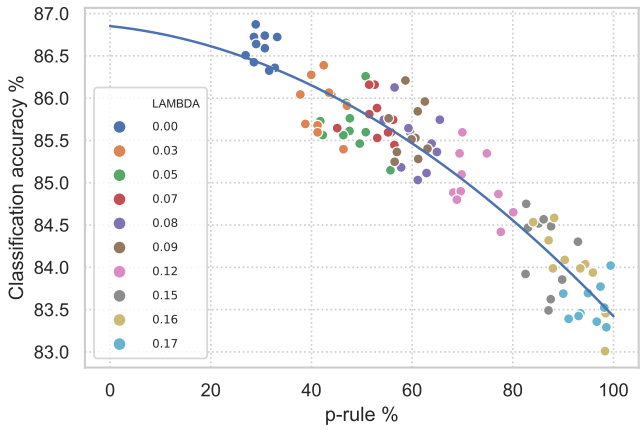


Figure 5.4: Impact of hyperparameter λ (Adult UCI data set): Higher values of λ produce fairer predictions, while λ near 0 allows to only focus on optimizing the classifier predictor.

5.4 | Conclusion

In this chapter, we developed a new approach to produce fair gradient boosting algorithms. Compared with other state-of-the-art algorithms, our method proved to be more efficient in terms of accuracy while obtaining a similar level of fairness.

Currently, we use a neural network architecture for the adversary. We chose this approach in order to recover the gradient of the input. Another possible strategy is to replace the adversarial neural network with deep neural decision forests (Kontschieder et al., 2015) which allow making an architecture of trees only.

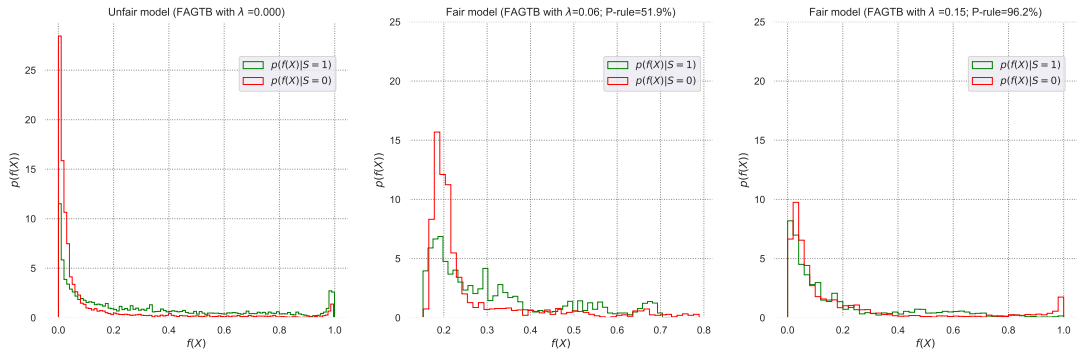


Figure 5.5: Distributions of the predicted probabilities given the sensitive attribute S (Adult UCI data set)

Another field left for further investigations is the mathematical identification of the optimal hyperparameter λ. The objectives are to automatically find the best trade-off between accuracy and fairness and improve optimization convergence. In addition, we are also interested in implementing this architecture by replacing the simple adversarial with our HGR adversarial network, which will allow us to deal with continuous and/or multidimensional features. Finally, it might be interesting to investigate a measure that does not only consider the general case of bias but can also spot and quantify bias that persists in specific sub-segments of the population.

6

Group Fairness Without the Sensitive Attribute

In this chapter, we investigate a new way to still satisfy some fairness criteria without having access to the sensitive attribute during the training process. In recent years, most fairness strategies in machine learning have focused on mitigating unwanted biases by assuming that sensitive information is available. However, this is not always the case in practice: due to privacy purposes and regulations such as RGPD in the EU, many personal sensitive attributes are frequently not collected. Yet, only a few prior works address the issue of mitigating bias in this difficult setting, in particular to meet classical fairness objectives such as Demographic Parity and Equalized Odds. By leveraging recent developments for approximate inference, we propose in this chapter an approach to fill this gap. To infer a sensitive information proxy, we introduce a new variational auto-encoding-based framework named SRCVAE that relies on knowledge of the underlying causal graph. The bias mitigation is then done after this inference step via an adversarial fairness approach. Our proposed method empirically achieves significant improvement over existing works in the field. We observe that the generated proxy's latent space correctly recovers sensitive information and that our approach achieves a higher accuracy while obtaining the same level of fairness on two real datasets.

Most of the work presented in this chapter is the subject of the paper *Fairness without the sensitive attribute via Causal Variational Autoencoder*, published at the IJCAI 2022 conference (Grari et al., 2021c).

The remainder of this chapter proceeds as follows. First, section 6.1 reviews papers related to our work and the motivation. Section 6.2 describes the methodology of our SRCVAE algorithm. Finally, section 6.3 presents experimental results of our

approach.

6.1 | Motivation and Related Work

Currently, the vast majority of these state-of-the-art approaches rely on having access to the sensitive information to be mitigated during training (though sometimes encrypted as in (Veale and Binns, 2017; Kilbertus et al., 2018)). However, in practice, it is often unrealistic to assume that this sensitive information is available or even collected. In Europe, for example, a car insurance company cannot ask a potential client about his/her origin or religion, as this is strictly regulated. Note that, as discussed in chapter 2, ignoring sensitive attributes as input of predictive models in order to achieve fairness, which is known as "fairness through unawareness" (Pedreshi et al., 2008), is not enough. Some complex correlations in the data may provide unexpected links to sensitive information (Dwork et al., 2012).

For this reason, some approaches have attempted to obtain a fair predictor model without the sensitive information. From the state-of-the-art literature, one possible way to achieve fairness despite the unavailability of sensitive attributes during training is to use transfer learning methods from external sources of data where the sensitive group labels are known. For example, (Madras et al., 2018) proposed to learn fair representations via adversarial learning on a specific downstream task and transfer it to the targeted one. (Schumann et al., 2019) and (Coston et al., 2019) focus on domain adaptation. (Mohri et al., 2019) considers an agnostic federated learning context by equalizing the performance of all participants through the lens of minimax optimization and fair resource allocation. However, this makes the actual desired bias mitigation highly dependent on the distribution of the external data. Other methods require prior knowledge on sensitive correlations. With prior assumptions, (Gupta et al., 2018) and (Zhao et al., 2021) mitigate the dependence of the predictions on the available features that are known to be likely correlated with the sensitive attribute. However, such strongly correlated features do not always exist in the data.

Finally, a few approaches address this objective without any prior knowledge on the sensitive information. Some of these works aim at improving the accuracy for the worst-case protected group (Rawlsian Max-Min objective) by leveraging techniques from distributionally robust optimization (Hashimoto et al., 2018) or adversarial learning (Lahoti et al., 2020). Other works act on the input data using a cluster-based balancing strategy in order to minimize the biases locally (Yan et al., 2020b). However, such methods are usually ineffective for traditional group fairness defi-

nitions such as *demographic parity* and *equalized odds*. Their blind way of mitigation affects non-sensitive information, likely implying a degradation of the predictor accuracy.

To overcome the limitations of these approaches (i.e., possessing relevant external data or correlated features and blind way of mitigation), we propose a novel approach that leverages a causal graph to reconstruct sensitive information using Bayesian variational autoencoders (VaEs). The inferred information is then used as a proxy for mitigating biases in an adversarial fairness training setting. We empirically show that this approach, based on sensitive reconstruction, is significantly more effective for achieving usual fairness objectives than its competitors, with a more direct control on mitigated biases. Our approach is inherently different from the aforementioned approaches. Based on minimal prior knowledge of causal relationships in the data, we perform Bayesian inference of latent sensitive proxies, whose dependencies with prediction outputs are mitigated in a second training step.

6.2 | Methodology

In this chapter, we consider training data where the binary sensitive attribute s_i for all i is unobserved. The only available training data is therefore $(x_i, y_i)_{i=1}^n$, where $x_i \in \mathbb{R}^p$ is the feature vector of the i-th example and y_i its binary outcome. In our context the training sample x_i is decomposed into two feature vectors $x_{c_i} \in \mathbb{R}^{p_c}$ and $x_{d_i} \in \mathbb{R}^{p_d}$. In addition, we consider an - unobserved - binary sensitive attribute s_i for all i.

In our approach, we first assume the existence and availability of a specific causal graph which underlies the training data, as discussed in subsection 6.2.1. The causal graph allows us to infer, through Bayesian inference, a latent representation containing as much information as possible about the sensitive feature. This process is described in subsection 6.2.2. Finally, we present in subsection 6.2.3 our methodology to mitigate fairness biases while preserving as much as possible prediction accuracy using this latent representation.

6.2.1 | Causal Structure of SRCVAE

Our work relies on the assumption of having an underlying causal graph describing the data, where causal interactions are indicated as directed edges between subsets of features (nodes). In particular, we suppose that the graph can be represented by the illustration shown in Figure 6.1. This structure is aimed to be generic enough to fit with most real world settings (slightly different graphs are studied in appendix C.2).

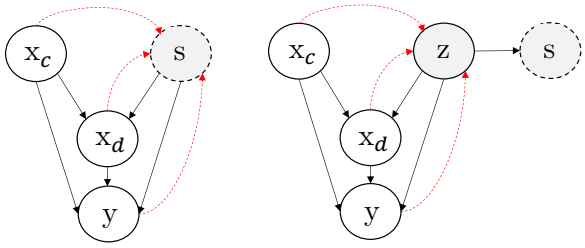


Figure 6.1: Causal graphs of SRCVAE: Left graph represents prior expert knowledge, where x is mapped into two components x_c and x_d. Right graph denotes the graph considered in our approach, with a multivariate confounder z inferred to be used as a proxy of the sensitive attribute s. Solid arrows denote causal links, red dashed arrows denote inference, grey circles denote missing attributes.

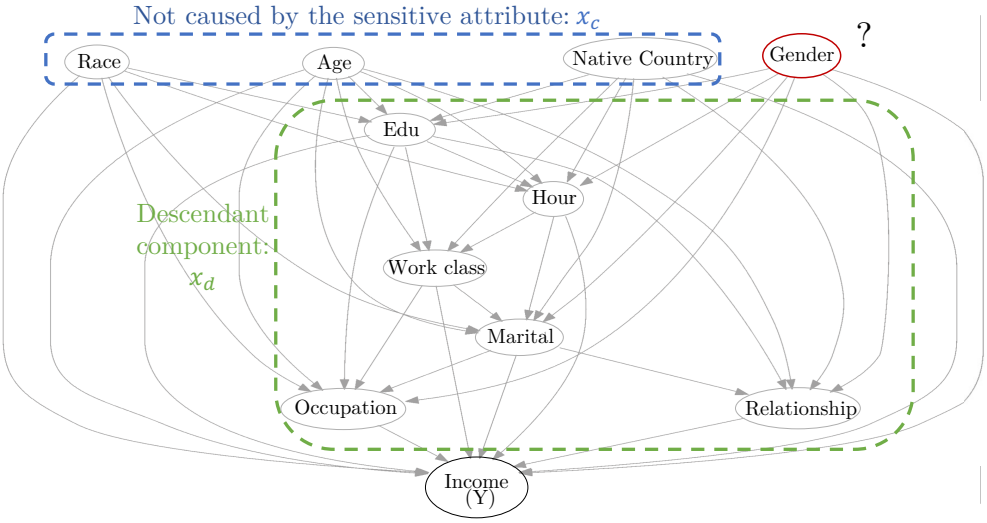


Figure 6.2: Causal Graph - Adult UCI

In the leftmost graph, parents of the output y are split into three components x_c, x_d and s. The subsets x_c and x_d regroup together all of the features that are given as input x to the model. The distinction between the two relies on the existence or absence of a causal relationship with the missing sensitive information s: no interaction is assumed with x_c, while some is with x_d. In addition, some causal relationship may exist between x_c and x_d.

To illustrate the generic aspect of this framework, we apply it to the Adult UCI dataset. The assumed causal graph of this dataset, with *Gender* as the sensitive attribute s and *Income* as the expected output y, is shown in Figure 6.2. In this context, x_c is the set of variables *Race*, *Age* and *Native_Country* which do not depend on the sensitive attribute, while x_d corresponds to all remaining variables that are generated from x_c and s (i.e., $x_d = \{Education, Work\_Class, ...\}$).

Assuming that all of the variables except s are available, our purpose is to recover all the hidden information not caused by the set x_c but responsible of x_d and y. In a real world scenario, it is noteworthy that the accuracy with which one can recover the real sensitive s depends on the right representation of the complementary set x_c. Yet, it is possible that the set x_c is under-represented. In such a case, there is a risk that the reconstruction of s may contain some of this missing additional information. For instance, assuming that the graph from Figure 6.2 is the exact causal graph that underlies the Adult UCI, let us consider a setting where the variable *Race* is hidden. Hence, this variable would be likely to leak in the sensitive variable reconstruction. In such a leakage setting, we argue that working with a binary sensitive proxy would strongly degrade the inferred sensitive information, by introducing noise in the reconstruction. This is what motivated us to rather consider the rightmost graph from Figure 6.1. It considers a multivariate continuous intermediate confounder z that both causes the sensitive s and the observed variables in x_d and y. As long as the confounder z contains the real sensitive information, removing the corresponding dependence with the output prediction is guaranteed to ensure fairness for the model (we prove this in 6.2.1). As we observe in the experiments section, such a multivariate proxy also allows for better generalization abilities for mitigated prediction.

6.2.2 | Reconstructing the Sensitive Attributes

We describe in this section the first step of our SRCVAE (Sensitive Retrieval Causal Variational Autoencoder) framework, which aims at generating a latent representation z that contains as much information as possible about the real sensitive feature s. As discussed above, our strategy is to use Bayesian inference approximation, using the pre-defined causal graph represented in Figure 6.1.

VAE Leveraging recent developments for approximate inference with deep learning, many different works proposed to use Variational Autoencoding methods (VAE) (Kingma and Welling, 2013) to model exogenous variables in causal graphs. It has been shown to achieve successful empirical results, in particular in the subfield of counterfactual fairness (Louizos et al., 2017; Grari et al., 2020b). We propose to apply VAE for our setting of fairness with hidden sensitive attribute.

Following the rightmost causal graph from Figure 6.1, the distribution $p_\theta(x_c, x_d, y|z)$ can be factorized as:

$$p_\theta(x_c, x_d, y|z) = p(x_c)p_\theta(x_d|x_c, z)p_\theta(y|x_c, x_d, z)$$

Given an approximate posterior $q_\phi(z|x_c, x_d, y)$, we obtain the following variational lower bound:

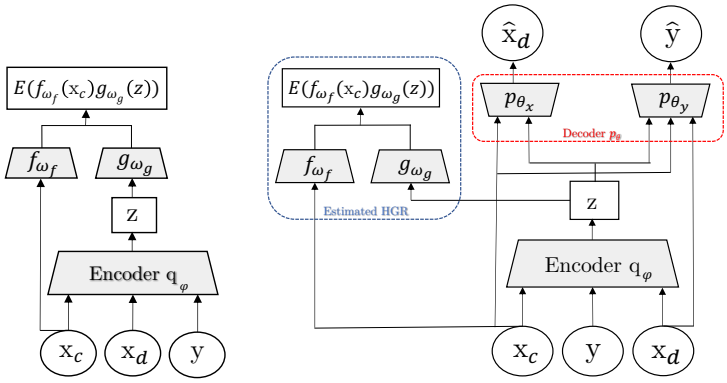
$$log(p_\theta(x_c, x_d, y)) \geq \mathbb{E}_{\substack{(x_c,x_d,y)\sim\mathcal{D}, \\ z\sim q_\phi(z|x_c,x_d,y)}} [\log p_\theta(x_d, y|x_c, z) + log(p(x_c)) - D_{KL}(q_\phi(z|x_c, x_d, y)||p(z))] \tag{6.1}$$

where D_{KL} denotes the Kullback-Leibler divergence of the posterior $q_\phi(z|x_c, x_d, y)$ from a prior $p(z)$, typically a standard Gaussian distribution $\mathcal{N}(0, I)$. The posterior $q_\phi(z|x_c, x_d, y)$ is estimated using a deep neural network with parameters ϕ, which typically outputs the mean μ_ϕ and the variance σ_ϕ of a diagonal Gaussian distribution $\mathcal{N}(\mu_\phi, \sigma_\phi I)$.

The likelihood term, which factorizes as $p_\theta(x_d, y|x_c, z) = p_\theta(x_d|x_c, z)p_\theta(y|x_c, x_d, z)$, is defined as the output of a neural network with parameters θ. Since attracted by a standard prior, the posterior is supposed to remove the probability mass for any information of z that is not involved in the reconstruction of x_d and y. Since x_c is given together with z as input of the likelihoods, all the information from x_c should be removed from the posterior distribution of z. In this chapter, we employ a variant of the ELBO optimization as done in (Pfohl et al., 2019), where the term $D_{KL}(q_\phi(z|x_c, x_d, y)||p(z))$ is replaced by a Maximum Mean Discrepancy (MMD) term $\mathcal{L}_{MMD}(q_\phi(z)||p(z))$ between the aggregated posterior $q_\phi(z)$ and the prior. This has been shown to be more powerful than the classical D_{KL} for ELBO optimization in (Zhao et al., 2017), as the latter may be too restrictive (Chen et al., 2016; Sønderby et al., 2016), and also tends to overfit the data.

HGR Minimization To be accurate, inference must ensure that no dependence is created between x_c and z (no arrow is linking x_c to z in the rightmost graph in Figure 6.1). This ensures the generation of a proper sensitive proxy that is not linked to the complementary x_c. However, by optimizing the ELBO Equation 6.1, some dependence may still be observed empirically between x_c and z, as we show in Section 6.3. This is due to some information from x_c leaking to the inferred z. In order to ensure some minimum independence level, we add a penalisation term in the proposed loss function. Leveraging our HGR estimation seen in chapter 3 by neural network (HGR_NN) for mitigating the dependence between continuous variables, we extend this main idea by adapting this penalization to the case of variational autoencoders.

In the following, we denote as $\widehat{HGR}^{w_f,w_g}_{U\sim\mathcal{D}_\mathcal{U},V\sim\mathcal{D}_\mathcal{V}}(U, V)$ the neural estimation of HGR between two variables U and V, computed via two inter-connected neural net-



(a) Max Phase: HGR (b) Min Phase: Reconstruction

Figure 6.3: Neural architecture of SRCVAE in max phase for the HGR estimation between x_c and z via gradient ascent (a) and Variational autoencoder structure of SRCVAE in min phase (b).

works f and g with parameters w_f and w_g:

$$\underset{U\sim\mathcal{D}_U,V\sim\mathcal{D}_V}{\widehat{HGR}}^{w_f,w_g}(U,V) = \max_{w_f,w_g} \mathbb{E}_{U\sim\mathcal{D}_U,V\sim\mathcal{D}_V}(\widehat{f}_{w_f}(U)\widehat{g}_{w_g}(V))$$

where $\mathcal{D}_U$ (resp. $\mathcal{D}_V$) is the distribution of U (resp. V), and $\widehat{f}$ (resp. $\widehat{g}$) refer to standardized outputs of network f (resp. g).

Reconstruction Objective Altogether, the final objective of our SRCVAE approach is given as:

$$\begin{aligned}\arg\min_{\theta,\phi}\max_{w_f,w_g} - \mathbb{E}_{\substack{(x_c,x_d,y)\sim\mathcal{D},\\ z\sim q_\phi(z|x_c,x_d,y)}}[\log p_\theta(x_d,y|x_c,z) \\ + \lambda_{mmd}\mathcal{L}_{MMD}(q_\phi(z)||p(z))] \\ + \lambda_{inf}\underset{\substack{(x_c,x_d,y)\sim\mathcal{D},\\ z\sim q_\phi(z|x_c,x_d,y)}}{\widehat{HGR}}^{w_f,w_g}(x_c,z)\end{aligned}$$

where λ_{mmd}, λ_{inf} are scalar hyperparameters. The additional MMD objective can be interpreted as minimizing the distance between moments of each aggregated latent code distribution and the prior distribution. Note that giving y as input of the inference scheme $q(z|x_c,x_d,y)$ is allowed since z is only used during training (see next section).

In Figure 6.3, we represent the min-max structure of SRCVAE. The left structure represents the max phase where the HGR between z and x_c is estimated by gradient

ascent with multiple iterations. The right graph represents the min phase where the reconstruction of x_d and y is performed by the decoder p_θ (red frame) via the generated latent space z from the decoder q_ϕ. The adversarial HGR component (blue frame) ensures independence between the generated latent space z and x_c. The network f takes the set x_c as input, while g takes the continuous representation space z. This way, for each gradient iteration of SRCVAE we capture the estimated HGR between the set x_c and the generated proxy latent space z. At the end of each iteration, the algorithm updates the parameters of the decoder parameters θ as well as the encoder parameters ϕ by one step of gradient descent. λ_{inf} controls the importance of the dependence loss in the optimization.

6.2.3 | Mitigating the Unwanted Biases

The sensitive reconstruction model can now be used for training a fair predictive function h_{w_h}. We propose to mitigate the unwanted bias via an adversarial penalization during the training phase that depends on the targeted fairness objective.

Demographic Parity We propose to find a mapping $h_{w_h}(x)$ that both minimizes the deviation with the expected target y and does not imply much dependency with the representation z, inferred from $q_\phi(z|x_c, x_d, y)$ as described in the previous section. We propose the following optimization, which considers a neural estimation of HGR as well, but this time applied to variables $h_{w_h}(x)$ (the output of the classifier) and z (the inferred latent representation):

$$\arg\min_{\theta} \max_{\psi_f, \psi_g} \mathcal{L}(h_{w_h}(x), y) + \lambda_{DP} \underset{\substack{(x_c, x_d, y)\sim\mathcal{D}, \\ z\sim q_\phi(z|x_c, x_d, y)}}{\widehat{HGR}^{\psi_f, \psi_g}} (h_{w_h}(x), z)$$

where $\mathcal{L}$ is the predictor loss function (the log-loss function in our experiments) of the output $h_{w_h}(x) \in \mathbb{R}$ w.r.t. the target label y. The hyperparameter λ_{DP} controls the impact of dependence between the output prediction $h_{w_h}(x) \approx p(y = 1|x_d, x_c)$ and the sensitive proxy z. To assess this correlation, K different representations are sampled for each observation (x_{c_i}, x_{d_i}, y_i) from the causal model (200 in our experiments). As in the inference phase, the backpropagation of the HGR adversary with parameters ψ_f and ψ_g is performed by multiple steps of gradient ascent.

Practice in real-world As mentioned in the first subsection, the assumed causal graph 6.1 requires the right representation of the complementary set x_c. If the set x_c is under-represented, some specific hidden attributes can be integrated with the

sensitive information in the inferred sensitive latent space z. The following Theorem 6.2.1 allows us to ensure that mitigating the HGR between z and $\hat{y}$ implies some upperbound for the targeted objective (proof in appendix C.1).

Theorem 6.2.1. *For two nonempty index sets S and Z such that $S \subset Z$ and $\hat{Y}$ the output prediction of the model, we have:*

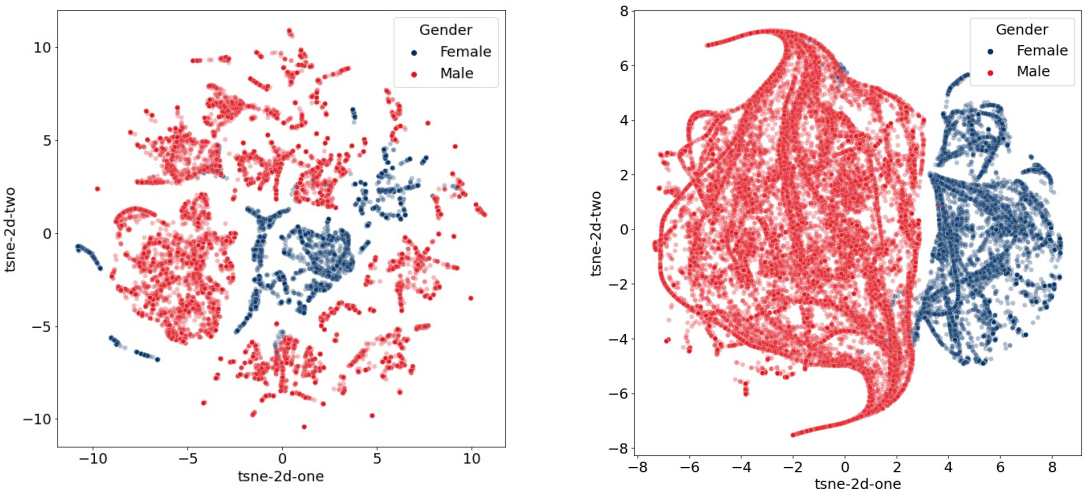
$$HGR(\hat{Y}, Z) \geq HGR(\hat{Y}, S) \tag{6.2}$$

Therefore, minimizing $HGR(\hat{Y}, Z)$ tends to reduce the real bias objective $HGR(\hat{Y}, S)$. Results on benchmark and real-world datasets demonstrate below in part 6.3 that such an assumed graph demonstrates good robustness properties. This property is also held for equalized-odds we consider below, with $HGR(\hat{Y}, Z|Y) \geq HGR(\hat{Y}, S|Y)$.

Equalized odds We extend the demographic parity optimization to the equalized odds task. The objective is to find a mapping $h_{w_h}(x)$ which both minimizes the deviation with the expected target y and does not imply too much dependency with the representation z conditioned on the actual outcome y. For the decomposition of disparate mistreatment, we propose to divide the mitigation based on the two different values of y. Identification and mitigation of the specific non linear dependence for these two subgroups leads to the same false positive and the same false negative rates for each demographic. We propose the following optimization:

$$\arg\min_{\theta} \max_{\psi_{f_0}, \psi_{g_0}, \psi_{f_1}, \psi_{g_1}} \mathcal{L}(h_{w_h}(x), y) + \lambda_0 \underset{\substack{(x,y)\sim\mathcal{D}_0,\\ z\sim q_\phi(z|x,y)}}{\widehat{HGR}^{\psi_{f_0}, \psi_{g_0}}}(h_{w_h}(x), z) + \lambda_1 \underset{\substack{(x,y)\sim\mathcal{D}_1,\\ z\sim q_\phi(z|x,y)}}{\widehat{HGR}^{\psi_{f_1}, \psi_{g_1}}}(h_{w_h}(x), z)$$

with $\mathcal{D}_0$ (resp. $\mathcal{D}_1$) corresponding to the observations set (x, y) verifying $y = 0$ (resp. $y = 1$). The hyperparameters λ_0 and λ_1 control the impact of the dependence loss for the false positive and the false negative objective respectively. The first penalisation (controlled by λ_0) enforces the independence between the output prediction $h_{w_h}(x) \approx p_\theta(y = 1|x)$ and the sensitive proxy z only for the cases where $y = 0$. It enforces the mitigation of the difference of false positive rates between demographics, since at optimum for w_h^* with no trade-off (i.e., with infinite λ_0) and $(x, y) \sim \mathcal{D}_0$, $HGR(h_{w_h^*}(x), z) = 0$ and implies theoretically: $h_{w_h^*}(x) \perp z|y = 0$. The second one enforces the mitigation of the difference between the true positive rates, since the dependence loss is performed between the output prediction $h_{w_h}(x)$ and the sensitive proxy only for cases where $y = 1$ (i.e., mitigation of Δ_{FNR}).



(a) λ_{inf}=0.0; $HGR(x_c, z)$=81.7%

(b) λ_{inf}=0.2; $HGR(x_c, z)$=22.6%

Figure 6.4: Inference phase for Adult UCI: t-SNE of the sensitive latent reconstruction Z. Blue points are males ($S = 1$), red ones are females ($S = 0$). Increasing λ_{inf} improves the independence of z from x_c. This leads to a better separation between male and female data points, which indicates a proper sensitive proxy.

6.3 | Experimental Results

For our experiments, we empirically evaluate the performance of our contribution on real-world data sets where the sensitive s is available. This allows to assess the fairness of the output prediction, obtained without the use of the sensitive attribute, w.r.t. this ground truth. For this purpose, we use the Adult UCI and Default datasets, presented in Subsection 4.5.

Sensitive Reconstruction In order to understand the interest of mitigating the dependence between the latent space z and the complementary set x_c during the inference phase, we plot the t-SNE of z with two different inference models for the Adult UCI dataset in Figure 6.4. We consider a version of our model trained without the penalization term ($\lambda_{inf} = 0.00$) as a baseline. It is then compared to a version trained with a penalization term equal to 0.20. As expected, training the inference model without the penalization term results in a poor reconstruction of the z proxy, where the dependence on x_c is observed. We can observe that the separation between the men (blue points) and women (red points) data is not significant. We also observe that increasing this hyper-parameter λ_{inf} allows to decrease the HGR estimation from

81.7% to 22.6% and to greatly increase the separation between male and female data points.

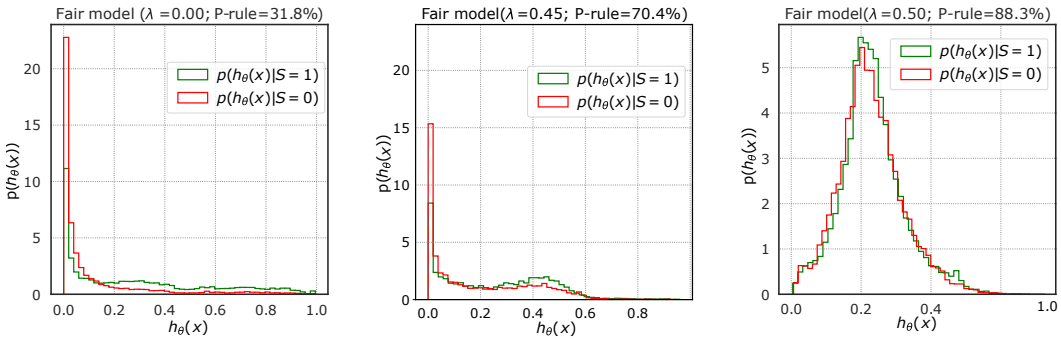
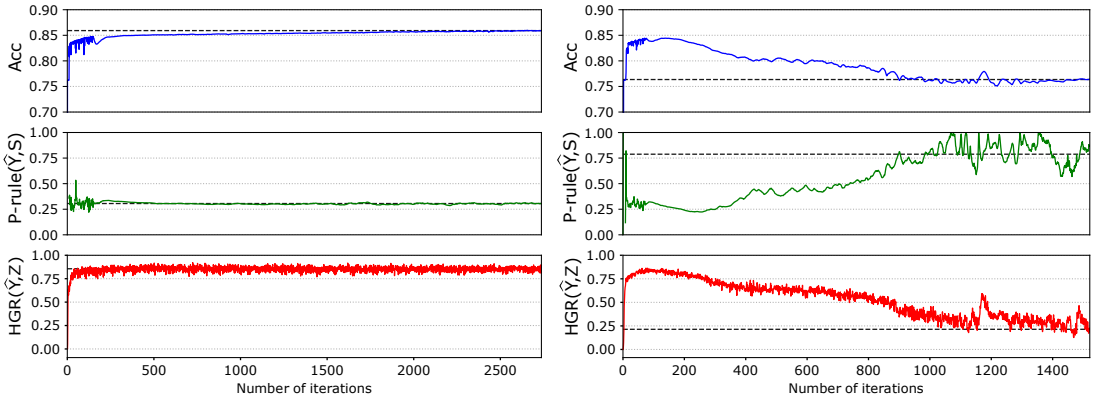


Figure 6.5: Distributions of the predicted probabilities given the real sensitive s (Adult UCI data set) for the Demographic Parity task.



(a) No mitigation ($\lambda_{DP} = 0$)

(b) Mitigation with $\lambda_{DP} = 0.5$

Figure 6.6: Dynamics of adversarial training

Bias Mitigation The dynamics of adversarial training for demographic parity is performed for Adult UCI with unfair ($\lambda_{DP} = 0$) and fair ($\lambda_{DP} = 0.5$) models as illustrated in Figure 6.6. Other values are presented in appendix C.3. We represent the accuracy of the model (top), the P-rule metric between the prediction and the real sensitive s (middle), and the HGR between the prediction and the latent space z (bottom). For the unfair model (leftmost graph) we observe that the convergence is stable and achieves a P-rule of 29.5%. As expected, the penalization loss decreases (mea-

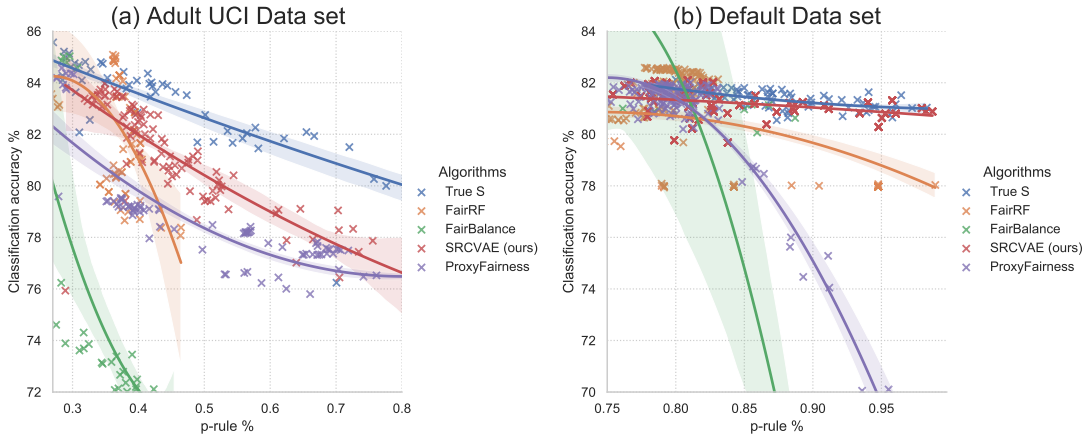


Figure 6.7: Demographic Parity task

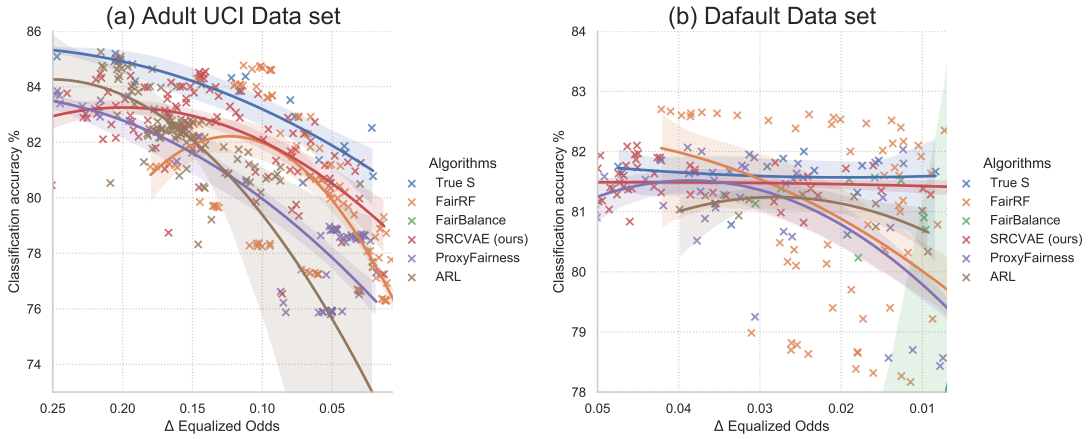


Figure 6.8: Equalized odds task

sured with the HGR estimator) when the hyperparameter λ_{DP} is increased. It allows to increase the fairness metric P-rule to 83.1% with a slight drop of accuracy.

In Figure 6.5 we plot the distribution of the predicted probabilities for each sensitive attribute s for three different models: an unfair model with $\lambda_{DP} = 0$, and two fair models with $\lambda_{DP} = 0.45$ and 0.50, respectively. For the leftmost graph (i.e. $\lambda_{DP} = 0$) the model appears to be very unfair, since the distribution between the sensitive groups differs importantly. As expected, we observe that the distributions are more aligned as λ_{DP} values increase.

For the two datasets, we test different models where, for each, we repeat five runs by randomly sampling two subsets, 80% for the training set and 20% for the test set. As different optimization objectives result in different algorithms, we run separate

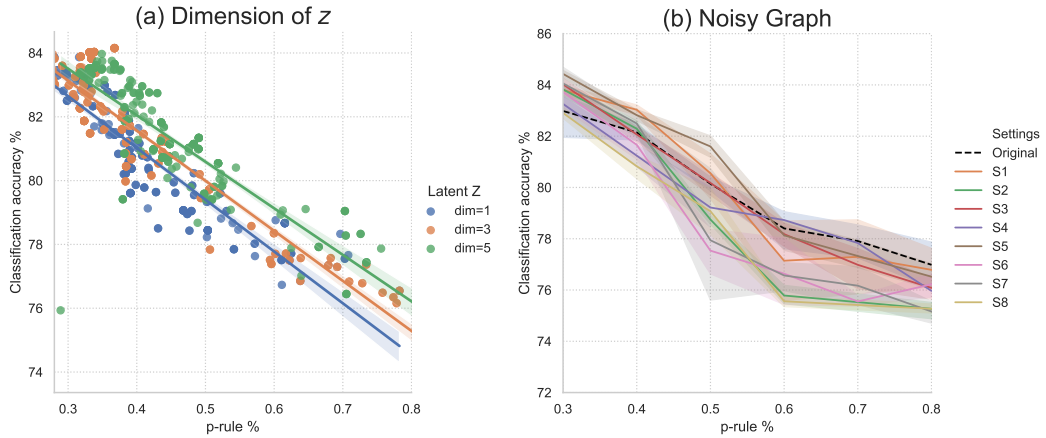


Figure 6.9: Additional Experiments

experiments for the two fairness objectives of our interest. As an optimal baseline to be reached, we consider the approach from (Adel et al., 2019) using observations of the sensitive s during training, which we denote as *True S*. We also compare various approaches specifically designed to be trained in the absence of the sensitive information during training: *FairRF* (Zhao et al., 2021), *FairBalance* (Yan et al., 2020b), *ProxyFairness* (Gupta et al., 2018) and *ARL* (Lahoti et al., 2020). The latter is only compared for the equalized odds task (i.e. discussion in (Zhao et al., 2021)). We plot the performance of these different approaches by displaying the Accuracy against the P-rule for Demographic Parity (Figure 6.7) and the Disparate Mistreatment (DM) for Equalized Odds (Figure 6.8). For all algorithms, we clearly observe that the Accuracy, or predictive performance, decreases when fairness increases. As expected, the baseline *True S* achieves the best performance for all the scenarios with the highest accuracy and fairness. We note that, for all levels of fairness (controlled by the mitigation weight in every approach), our method outperforms state-of-the-art algorithms for both fairness tasks (except some points for very low levels of fairness, on the left of the curves). We attribute this to the ability of SRCVAE to extract a useful sensitive proxy, while the approaches *FairRF* and *ProxyFairness* seem to greatly suffer from merely considering correlations present in the data for mitigating fairness. The approach *FairBalance*, which pre-processed the data with clustering, seems inefficient and degrades the predictive performance too significantly. The advantages of our approach are even more pronounced on the Default dataset, where a less obvious correlation exists between observed variables and the sensitive attribute. In that setting, leveraging the knowledge of a causal graph appears to be crucial.

Proxy dimensions In figure 6.9(a), we perform an additional experiment on the sensitive proxy. For the two datasets we observe that increasing z dimensions results in increased accuracy. Increasing the dimensions to 5 for Adult UCI (same experiment for Default in appendix C.3) allows to obtain better results in terms of accuracy and this for all levels of P-rule. We claim that mitigating biases in larger spaces allows better generalisation abilities at test time, as already observed in section 4.5. It supports the choice of considering a multivariate sensitive proxy z, rather than directly acting on a reconstruction of s as a univariate variable.

Noisy graph In figure 6.9(b), we analyse the impact of noise in the causal graph. To do this, we focus on cases where the decomposition of x in sets x_c and x_d is noisy, or sets of variables are under-represented. For this purpose, we experimented 8 scenarios on the Adult UCI data set. First, we removed features from x_c: the *race* (S1), the *age* (S2). Then, we removed features from x_d: the *education* (S3) and the *hour* (S4). Finally, we moved features from x_c to x_d and reversely: membership inversion between *race* and *education* (S5), membership inversion between *age* and *hour* (S6), inclusion of *age* in x_d (S7) and inclusion of *hour* in x_c (S8). From the results, our approach appears greatly robust to noise, with results in every scenario at least comparable to the best considered competitors (which all present settings where performances catastrophically drop as observed in Fig. 6.7 and 6.8). This robustness is partly achieved thanks to the use of a multivariate continuous proxy z, which limits the possible lack of sensitive information that would occur with a scalar proxy of s, if non-sensitive information leaks in the reconstruction. While the inclusion of variables from x_d to x_c may induce the removal of some useful sensitive information from the proxy, the inclusion of variables from x_c to x_d may lead to optimize the independence of some non sensitive information with model outputs. If fairness needs to be guaranteed, the expert must thus tend to favor false x_d variables rather than false x_c, the former only inducing a slight accuracy loss in most cases (as demonstrated in Theorem 6.2.1).

6.4 | Conclusion and Future Work

In this chapter we proposed a new way to mitigate undesired bias without the availability of the sensitive demographic information in training. To generate a latent representation which is expected to contain as much sensitive information as possible, the approach relies on a new variational auto-encoding based framework named SRCVAE. In a second phase, inferred proxies serve to mitigate biases in an adver-

sarial fairness training of a prediction model. Compared with other state-of-the-art algorithms, our method proves to be more efficient in terms of accuracy for similar levels of fairness. For further investigation, we are interested in extending this work to settings where the actual sensitive can be continuous (e.g. age or weight attribute) and/or multivariate.

7

Group Fairness for Insurance Pricing

This chapter discusses the importance of adapting the traditional fairness algorithms to specific real-life applications and, in particular, to insurance pricing. We claim that mitigating undesired biases with a generic fair algorithm can be counterproductive for specific applications. Fairness in insurance pricing is a relatively new and much-requested topic, especially in light of new laws and regulations and past issues encountered in practice (Dolman et al., 2021; Frees and Huang, 2021; Embrechts and Wüthrich, 2022; Block et al., 2008). Consequently, companies/regulators are looking for new methodologies to ensure a sufficient level of fairness while maintaining an adequate accuracy of predictive models. However, traditional Fair-ML as adversarial methods are not currently adequate for insurance pricing. Therefore, for these purposes, we have developed a more suitable and effective framework to satisfy a fairness objective while maintaining a sufficient level of predictor accuracy. Please note that, in insurance, the term pricing is underpinned by the "pure premium", which is the basic and essential element in assigning prices to policyholders. We are looking at this from the insurer's perspective, for whom the pricing target is an exercise in predicting future costs.

At the core of insurance business lies classification between risky and non-risky insureds, actuarial fairness meaning that risky insureds should contribute more and pay a higher premium than non-risky or less-risky ones. Actuaries, therefore, use econometric or machine learning techniques to classify, but the distinction between a fair actuarial classification and 'discrimination' is subtle. For this reason, there is a growing interest about fairness and discrimination in the actuarial community (Lindholm et al., 2022). Surprisingly, we will show that debiasing the predictor alone may be insufficient to maintain adequate accuracy (1). Indeed, the traditional pricing model is currently built in a two-stage structure that considers many potentially

biased components such as car or geographic risks. We will show that this traditional structure induce significant limitations in achieving fairness. For this reason, we have developed a novel pricing approach. Recently some approaches have shown the value of autoencoders in pricing (Blier-Wong et al., 2021b; Wuthrich and Merz, 2021). In this chapter, we will show that (2) these techniques can be generalized to multiple pricing factors (e.g., geographic, car type), and that (3) perfectly be adapted for the fair pricing context since it allows to debias the set of pricing components. We extend this main idea to a general framework in which a single whole pricing model is trained by generating the geographic and car pricing components needed to predict the pure premium while mitigating the unwanted bias according to the desired metric. In this context of big data, where insurance companies collect more and more data, aggregating features by representation learning techniques seems to be a judicious choice to ensure explainability, computational traceability and fairness. We present our approach on private car insurance but it can be generalized for commercial, health, and household insurance products.

In section 7.1, we briefly describe the actuarial pricing literature with the application of fair adversarial algorithms. In section 7.2, we present our general extension that is more adapted for applying fairness to actuarial pricing. Finally, section 7.3 presents the experimental results of the approaches.

Parts of the work presented in this chapter are the subject of the paper *A fair pricing model via adversarial learning* (Grari et al., 2022).

7.1 | Actuarial Pricing

Insurance is usually described as the contribution of many to the misfortune of the few by pooling risks together. A fair contribution that should be asked to policyholders, who purchased an insurance policy, is its expected loss over the coverage period of the contract (usually one year), the so-called pure premium. Insurance pricing relies essentially on the law of large numbers, but since risks have to be homogeneous, it is important to classify the risks properly, as explained for instance in (Thomas, 2012). For this risk classification, insurers try to split policyholders into different groups, and risks are pooled within each group. Those groups are supposed to be as homogeneous as possible. They are usually based on observable factors, such as in motor insurance, the age of the (main) driver, the power of the car, some information about the spatial location, possibly the value of the car, and maybe the gender of the driver. This classification is never perfect, but it should be accepted by policyholders

and as valid as possible to be a competitive premiums. Heterogeneity within groups means that policyholders cross-subsidy, which could yield adverse selection, where lower risks might be attracted by a competitor. Insurers can capture this residual heterogeneity by offering some more personalized premiums (possibly cheaper for some of them), while the more risky will remain in the portfolio, and cross-subsidy will not work anymore. Therefore, the goal of risk classification is to charge for a risk-based fair actuarial premium and to avoid unnecessary cross-subsidy. As explained in (Paefgen et al., 2013), "*In order to differentiate the risk of insurance policies, actuaries use a set of rate factors to separate policies into groups (i.e., tariff classes).*" *For each tariff class, actuaries analyze historical claims data to arrive at a reliable estimate of the corresponding pure premium, that is, the minimum required payment per policy to cover the expected losses from its class.*"

7.1.1 | Description of the Method

A classical technique in actuarial science is based on Generalized Linear Models, GLM, since they satisfy a "*balance property*", as called in (Wuthrich and Merz, 2021) for instance, stating that the sum of y_i's (on the training dataset) should equal the sum of $\widehat{y}_i$'s. The economic interpretation is that the first sum of the sum of losses, while the second one is the sum of premiums. Hence, using GLMs, we ensure that, on average, the insurer will be able to repay policyholders claiming a loss. Unfortunately, GLMs experience difficulties when categorical rating factors have a large number of levels, not only because of the computational cost of dealing with high-dimensional design matrices, but also because of the implied statistical uncertainty, both in parameter estimation and prediction (even if regularization techniques can be used, as in (Frees and Lee, 2015)).

In this insurance pricing context, the GLM preditor model is not fitted only once as in the traditional machine learning standard. Instead, as mentioned in (Taylor, 1989), (Boskov and Verrall, 1994), or more recently (Tufvesson et al., 2019) a standard approach in insurance is to consider a two-step procedure, where the first step initially considers the generation of geographic and automobile risk components. Then the second step is based on the predictive task. The insurer often has recourse to a large number of external variables, which can usually exceed a hundred (Boucher, 2016; Beraud-Sudreau, 2017). These can be geographical (e.g., the total number of thefts or crime rate in the area of residence) and car-specific (e.g., airbag, emergency braking, etc.). (Shi and Shi, 2021) recalls that it can be difficult to use, in classical actuarial pricing models, categorical variables with a large number of categories, such as a

ZIP code (spatial information) or type/variant/version/model/make of cars (vehicle information). In order to keep models under control (i.e., computational traceability and explainability), actuaries traditionally prefer to aggregate all this information into single variables. The generation of these components is performed by a first model based on the prediction of the target Y (i.e., frequency and severity), excluding all external information (spatial and unstructured effects). The idea is that the observed residuals of this model correspond to the missing information that it was unable to capture (i.e., geographical, car information). In this way, a second model is trained on these observed residues from these external data. This method allows assigning external risk prediction to new policyholders based on their external information. For example, suppose a policyholder comes from a residential area that the insurer has never had in its historical data. In this case, the insurer will still be able to generate a geographic risk level using only external information from the area. Note that actuaries, notably by K-Nearest Neighbors (KNN) methods, reprocess this residual to smooth the predictions (e.g, KNN on the spatial neighborhood for the geographic component). As discussed in (Blier-Wong et al., 2021b), early geographic models in actuarial science were models that smoothed the residuals of a regression model (also called "*correction models*") where geographic effects are captured after the main regression model, in a smoothing model, as in (Taylor, 1989). For example (Fahrmeir et al., 2003), and more recently (Wang et al., 2017), suggested using spatial interpolation, inspired by kriging techniques, to capture spatial heterogeneity. In addition, a quantile approach is applied to these components in order to divide these components into several risk levels. Finally, a final model is fitted to the objective task based on these generated risk components and policyholder information.

In this particular context, the training data $x_i \in \mathbb{R}^d$ is decomposed into to subsample $x_{p_i} \in \mathbb{R}^{d_p}$ corresponding to the policyholder's information, $x_{g_i} \in \mathbb{R}^{d_g}$ corresponding to the geographical information with p_g predictors and $x_{c_i} \in \mathbb{R}^{d_c}$ corresponding to the car information with d_c predictors.

In Figure 7.1, we have represented the general process of the two-step traditional pricing. For the first step, a model h_{w_0} with parameters w_0 is fitted to predict the target task Y from the policyholder's information X_p (i.e., without including the external information X_g and X_c). The residual of this model R is calculated (can be calculated on the relative difference in practice (Said, 2016)). In second, some models h_{w_g} and h_{w_c} with parameters w_g and w_c are trained to predict the observed residuals R from the corresponding set of variables X_g and X_c respectively. Some prereatreatments are realized on the output predictions of these models as KNN methods and/or quantiles. Finally, a last model h_{w_h} with parameter w_h is trained to predict the target task

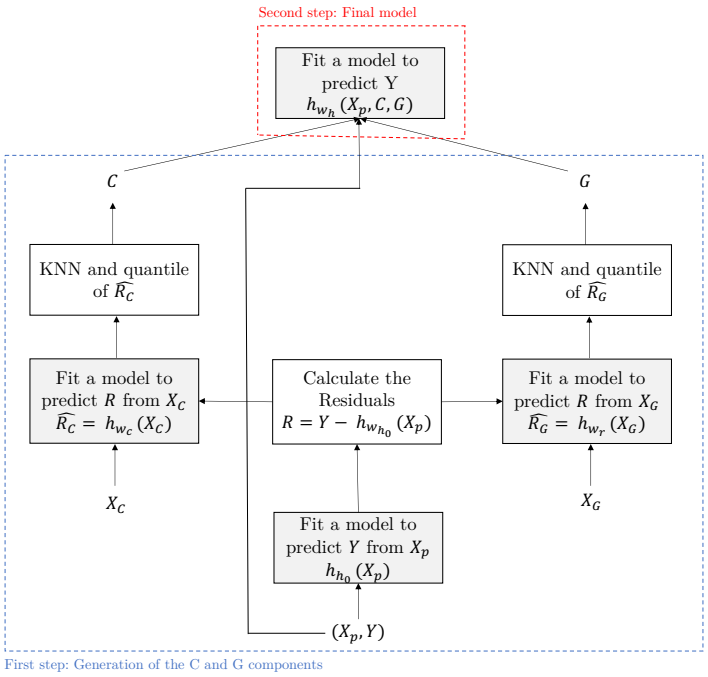


Figure 7.1: Two-Step Traditional Pricing model

Y from the policyholder's information X_p and the two generated features C and G.

7.1.2 | Fairness on the Traditional Two-Step Pricing Model

A classic methodology for enforcing fairness should be to act directly on the final predictor model h_{w_h}. An additional adversarial (e.g, simple or HGR seen in chapter 3) could mitigate the sensitive bias during the training process. However, we argue that this is not an appropriate process. Note that the car risk C and the geographic risk G have already been trained in the first step and are not re-trained in this optimization. Therefore, the unwanted bias is only mitigated on the predictor h_{w_h}. Consequently, if the predictor h_{w_h} is a GLM and if G and C are strongly dependent on the sensitive attribute S, the parameters w_h corresponding to G and C may tend towards 0 and thus nullify the effect of these factors. For improving fairness, the risk is to miss the relevant information about G and C for predicting Y.

This motivates us for a more robust model that breaks away from this drawback. So the goal here will be to define a new actuarial framework that is able to generate fair geographical and car variables, that are strongly correlated with the risk Y, but fair with respect to S.

7.2 | Pricing with an Autoencoder Structure

We propose in this section to generalize an actuarial model that would be better fitted for fairness. The main idea would be to train the geographical and car risk at the same time of the model. As discussed in (Wuthrich and Merz, 2021), a classical starting point is some initial feature engineering step, where some embedding of spatial components, and information relative to the vehicle, are considered, using principal component analysis (PCA). However, in this approach, the generated components are not trained with the predictor and cannot be specifically targeted to the objective task. Instead of considering several separate models stacked together as described above, it is possible to create a unique actuarial pricing model that can be trained as a whole. This has different advantages, which we will mention below, especially for fairness. Different approaches have recently focused on spatial embedding (Blier-Wong et al., 2021a,b) and have shown superior performance than traditional pricing strategies. These models propose to aggregate by deep neural network the geographic information into a unique/multidimensional representation by providing this information into the predictor model during the training. As said in (Blier-Wong et al., 2021b), those "*geographic embeddings are a fundamentally different approach to geographic models studied in actuarial science*", since the different models are trained at the same time with the objective to predict the pure premium. It allows having aggregate geographic risks adapted for the targeted risks. We extend this main idea to a general framework where a unique model is trained by generating the car (C) and geographic pricing (G) components required to predict the pure premium.

7.2.1 | Description of the Method

First in a pure predictive task, we propose to find a car and a geographical aggregation via a latent representation C and G respectively, which both minimizes the deviation between the target Y and the output prediction $\widehat{Y}$. In Figure 7.2, we represent our model extension. The output prediction is provided by a function $h_{w_h}(X_p, C, G)$ where h_{w_h} is a predictor with parameters w_h, which takes as input X_p the policies information, C and G. Let c_{w_c} and γ_{w_γ} be two neural networks with respective parameters w_c and w_γ, the latent representation C is generated as $c_{w_c}(X_c)$ (resp. G as $\gamma_{w_\gamma}(X_g)$) with X_c as the information about the car (resp. X_g as the geographical information). Depending on the task or objective, we can consider the latent representation C and G as multi-dimensional. This can therefore provide a rich representation for the geographical and car ratemaking. The mitigation procedure follows the optimization

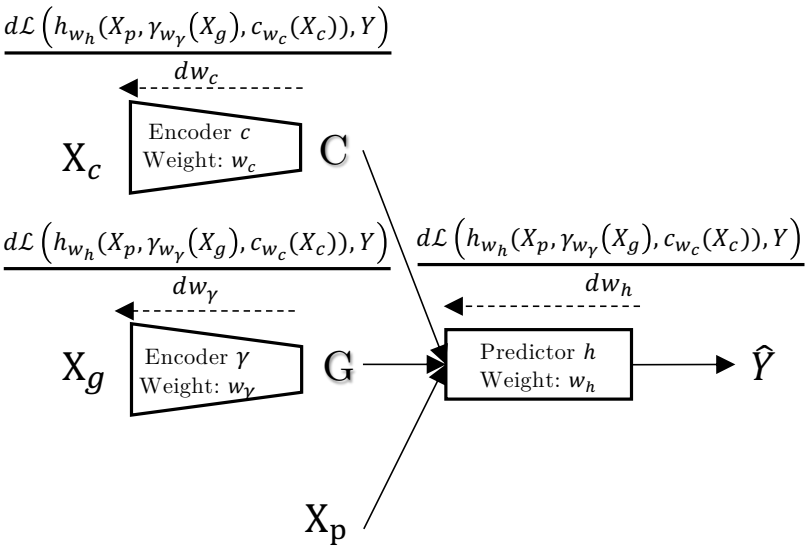


Figure 7.2: Pricing model via an autoencoder structure

problem:

$$\underset{h_{w_h}, g_{w_g}, c_{w_c}}{\arg\min} \left\{ \mathcal{L}(h_{w_h}(X_p, \gamma_{w_\gamma}(X_g), c_{w_c}(X_c)), Y) \right\} \tag{7.1}$$

where $\mathcal{L}$ is the predictor loss function between the prediction $h_{w_h}(X_p, \gamma_{w_\gamma}(X_g), c_{w_c}(X_c)) \in \mathbb{R}$ and the corresponding target Y. Note that smoothing can be performed on G such as (Blier-Wong et al., 2021b) to avoid that some nearby regions have too high premium volatility, especially when the risk exposure is very low.

7.2.2 | A Fair Pricing Model via an Autoencoder Structure

In this section, we adapt our autoencoder pricing model by adding an adversarial structure for fairness purposes. The objective is to find a mapping of a prediction $h_{w_h(X)}$ that both minimizes the deviation from the expected target Y and does not imply too much dependence on the sensitive attribute S, according to its definition for the desired fairness objective seen in section 2.2. This strategy is radically different from the previous two-stage traditional strategy since the training of the car risk and the geographic risk is done simultaneously with the learning of the predictor. In this case, the objective is, unlike the previous model, to recover the essential information from G and C to predict Y and solely neutralize the undesirable effects during the learning process. To achieve this, the back-propagation of the learning of the c_{w_c} and γ_{w_γ} encoders is performed at the same time as the penalization of the adversarial fairness component. It allows minimizing the deviation from the expected target Y and

does not imply too much dependence on the sensitivity S, as defined for the desired fairness objective 2.2. The predictor h_{w_h} is also back-propagated in the same way but takes as input the G and C attributes in addition to the policy contract information X_p.

Demographic Parity The predictor function is defined as $h_{w_h}(X_p, \gamma_{w_\gamma}(X_g), c_{w_c}(X_c))$ where h is a predictor model that takes as input X_p, the geographical risk $\gamma_{w_\gamma}(X_g)$ and the car risk $c_{w_c}(X_c)$).

The mitigation procedure follows the optimization problem:

$$\begin{aligned}\underset{c_{w_c}, \gamma_{w_\gamma}, h_{w_h}}{\arg\min} \Big\{ \max_{w_f, w_g} \Big\{ & \mathcal{L}(h_{w_h}(X_p, \gamma_{w_\gamma}(X_g), c_{w_c}(X_c)), Y) \\ & + \lambda \mathbb{E}(\widehat{f}_{w_f}(h_{w_h}(X_p, \gamma_{w_\gamma}(X_g), c_{w_c}(X_c))) * \widehat{g}_{w_g}(X_p)) \Big\} \Big\}\end{aligned} \tag{7.2}$$

where $\mathcal{L}$ is the predictor loss function between the output prediction and the corresponding target Y. We add in this objective optimization a second term representing our HGR estimation between the output prediction and the sensitive attribute S. It corresponds to the expectation of the products of standardized outputs of both networks ($\hat{f}_{w_f}$ and $\hat{g}_{w_g}$). The hyperparameter λ controls the impact of the correlation loss in the optimization.

Figure 7.3 gives the full architecture of our adversarial learning algorithm using the neural HGR estimator between the output predictions and the sensitive attribute S. It depicts the encoders functions γ_{w_γ} and c_{w_c}, which respectively outputs a latent variable G from X_g and C from X_c. The two neural networks f_{w_f} and g_{w_g}, which seek at defining the most strongly correlated transformations of the output predictions $h_{\omega_h}(X_p, \gamma_{w_\gamma}(X_g), c_{w_c}(X_c))$ and S. The model h_{ω_h} outputs the prediction $\widehat{Y}$ from the information X_p, G and C. The encoders aggregate the information of G and C from the information of X_g and X_c in order to maximize the performance accuracy for h_{w_h} and simultaneously minimize the HGR estimation finding with the adversary f_{w_f} and g_{w_g}. Left arrows represent gradient back-propagation. The learning is done via stochastic gradient, alternating steps of adversarial maximization, and global loss minimization.

The algorithm takes as input a training set from which it samples batches of size b at each iteration. At each iteration, it first standardizes the output scores of networks f_{w_f} and g_{w_g} to ensure 0 mean and a variance of 1 on the batch. Then it computes the HGR neural estimate and the prediction loss for the batch. At the end of each iteration, the algorithm updates the prediction parameters ω_h as well as encoder parameters ω_γ and ω_c by one step of gradient descent.

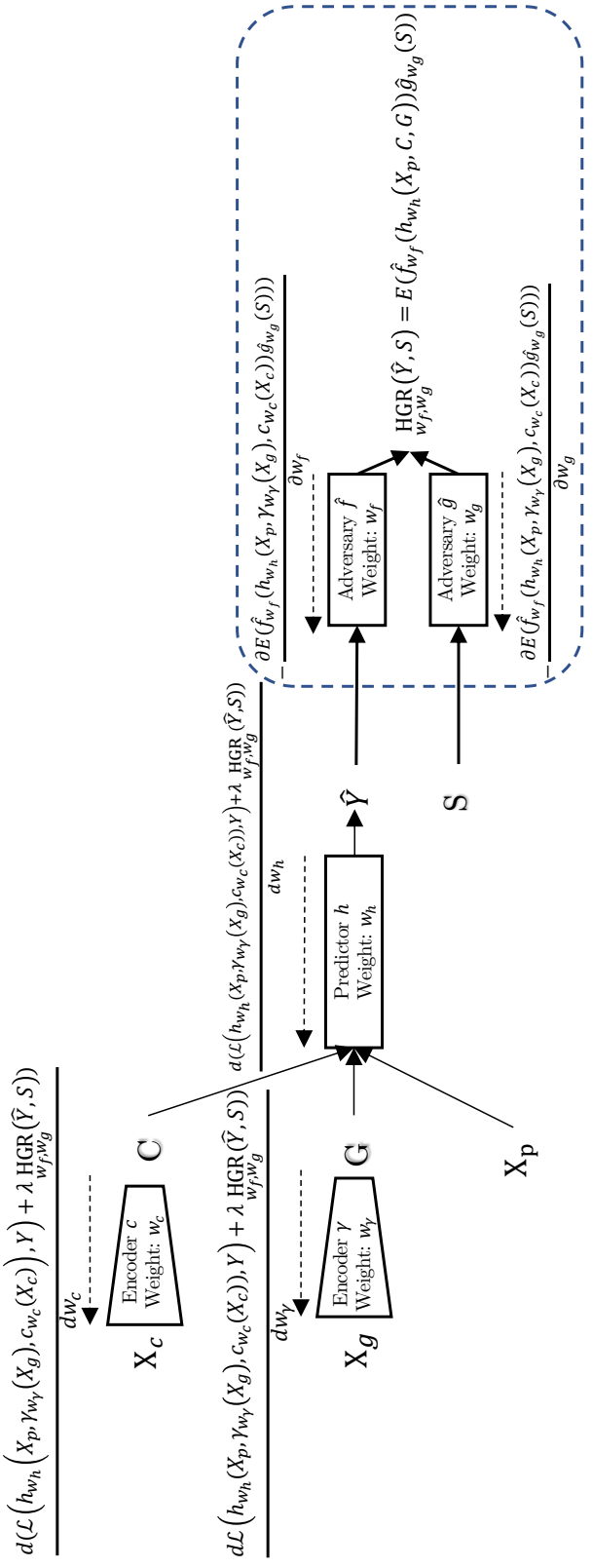


Figure 7.3: The Fair Demographic Parity Pricing Model via an autoencoder structure

Equalized odds For equalized odds, we also extend the main idea of mitigation seen in section 4.2.2.2. The mitigation procedure follows the optimization problem:

$$\underset{c_{w_c},\gamma_{w_\gamma},h_{w_h}}{\arg\min}\left\{\max_{w_{f_0},w_{g_0},\ldots,w_{f_K},w_{g_K}}\left\{\mathcal{L}(h_{w_h}(X_p,\gamma_{w_\gamma}(X_g),c_{w_c}(X_c)),Y)\right.\right. \\ \left.\left.+\frac{1}{K+1}\sum_{y\in\Omega_Y}\lambda_y\mathbb{E}_{(X,S)\sim\mathcal{D}_y}(\widehat{f}_{w_{f_y}}(h_{w_h}(X_p,\gamma_{w_\gamma}(X_g),c_{w_c}(X_c))*\widehat{g}_{w_{g_y}}(S))\right\}\right\} \tag{7.3}$$

where $\mathcal{D}_y$ corresponds to the distribution of pair (X,S) conditional on $Y=y$ and $K=\#\Omega_Y-1$. The hyperparameters λ_y control the impact of the dependence loss for the different number event objective. The penalization enforces the independence between the output prediction and the sensitive S only for the cases where $Y=y$. It enforces naturally the mitigation of equalized odds since it enforces the mitigation of biases for demographic parity for each number of events.

7.3 | Results and Discussion

We evaluate the performance of these fair algorithms empirically with respect to performance accuracy and fairness. We conduct the experiments on a synthetic scenario and real-world datasets.

7.3.1 | Synthetic Scenario

We illustrate the fundamental functionality of our proposal with a simple synthetic scenario that was inspired by the Red Car example illustrated in chapter 5. The purpose of this exercise is also to estimate the claim likelihood without incorporating any gender bias in terms of *Demographic Parity*. However, here, external information about geography and car is added to the model, making the fairness task more complex. We compare for this objective, the fair traditional pricing structure denoted as *BASE* and the fair pricing autoencoder structure as shown in the subsection 7.2.2 denoted as *OURS*. We focus on the general claim likelihood and ignore the severity or cost of the claim. Further, we only consider the binary case of claim or not (as opposed to a frequency). We assume that the claim likelihood only depends on the aggressiveness and the inattention of the policyholder. To make the training more complex, these two properties are not directly represented in the input data but are only indirectly available through correlations with other input features. We create a binary label Y with no dependence on the sensitive attribute S. Concretely, we use as protected attribute the *gender* of the policyholder. The unprotected attributes *color*, the maximum *speed* of the cars, and the average *salary* of the policyholder's area are

all caused by the sensitive attribute. In our data distribution, the *color* and the maximum *speed* of the car are strongly correlated with both *gender* and aggressiveness. *Age* is not correlated with *gender*. However, *age* is correlated with the inattention of the policyholder. Thus, the latter input feature is actually linked to the claim likelihood. First, we generate the training samples $(x_{p_i}, x_{g_i}, x_{c_i}, s_i, y_i)_{i=1}^n$. The unprotected attribute $x_{p_i} = age_i$ represents the policy information with the *age* of the policyholder, $x_{g_i} = Sal_i$ represents the geographical information with the average *salary* of the area and $x_{c_i} = (Col_i, Sp_i)$ represent the *colors* and the maximum *speed* of the car.

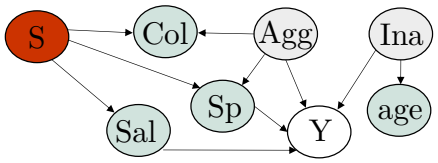


Figure 7.4: Causal Graph - Synthetic Scenario

We plot the performance of these different approaches in Figure 7.5. In the first graph, we can see the curves of accuracy against the fairness metric p-rule during the training phase. Note that on the top of the left of these curves, the λ hyper-parameters are fixed to 0 for the two models, therefore, only the performance accuracy is optimized during the training for this case. We observe that, for all levels of fairness (controlled by the mitigation weight λ in the two approaches), the model via autoencoder outperforms the traditional two-stage model. We attribute this to the ability of the autoencoder to extract a useful car and geographic risk. The traditional pricing structure has significant limitations in achieving fairness. In the middle left graph, as expected for the autoencoder model, we observe that the dependence between car risk and the sensitive S is lower for higher λ. However, this is not the case for the traditional model, where we observe stagnation since the minimax fairness optimization is performed only on the final predictor model. Furthermore, we observe on the two most right graphs that the bias reduction is at the expense of the essential unbiased information for the traditional one. This information is too strongly reduced for the latter, in contrast to the autoencoder model where the dependence between the car risk and the prediction is higher while being less biased.

7.3.2 | Real-World Datasets

We consider two real-world datasets: the pricingame 2015 (The Institute of Actuaries of France, 2015; Charpentier, 2014) and 2017 (The French Institute of Actuaries, 2017). Each of them contains 100,000 TPL policies for private motor insurance. In these data sets, we perform the two objectives of fairness *Demographic Parity* and *Equalized Odds* for three different tasks: Binary, frequency, and continuous objectives. For the binary

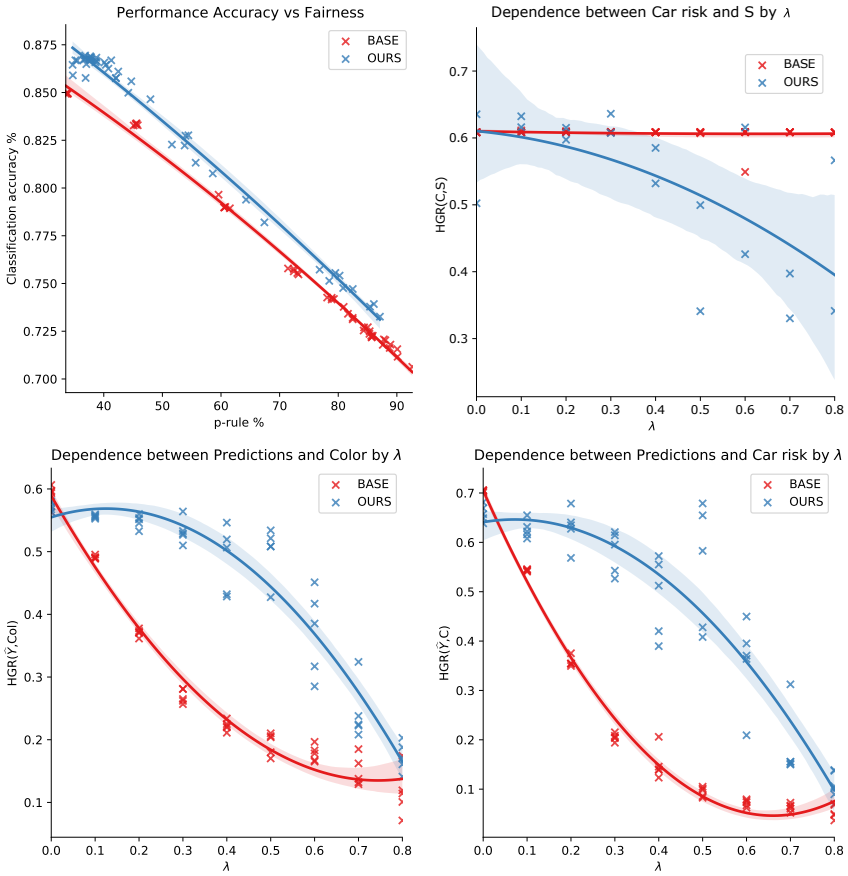


Figure 7.5: Impact of hyperparameter λ (Synthetic Scenario): (Left) For all levels of predictive performance accuracy, our proposed method outperforms the traditional fair pricing model. (Middle left) Higher λ values produce a fairer car risk to our approach than traditional car risk where biases mitigation is not performed. (Middle right and Right) Higher λ values significantly reduce the dependence between output predictions and the car risk, as well as for the geographical risk in comparison to our proposed model.

objective, we propose to predict the TPL claim of policyholders. For the continuous objective, we propose to predict the cost of third-party material claims (euro). For the frequency objective, we predict the number of third-party claims. For the *Pricingame 2017* dataset, we consider x_c as the set of variables *vh_age*, *vh_cyl*, *vh_din*, *vh_speed*, *vh_value*, *vh_weight*, *vh_make*, *vh_model* , *vh_sale_begin*,*sale_end*, *vh_type* , x_g as the set of variables representing more than 100 hundreds external features collected from French INSEE organism via the *INSEE_code* (more details in the code[1]) and all the remaining variables for x_p. For the *Pricingame 2015* dataset, we considers x_c as the set of variables *Type*, *Category*, *Group1*, *Value* of the car, x_g as the *Density* feature and all

[1]https://github.com/LeoPetrini/XGBoost-in-Insurance-2017

the remaining variables for x_p (more information in (Dutang and Charpentier, 2019)).

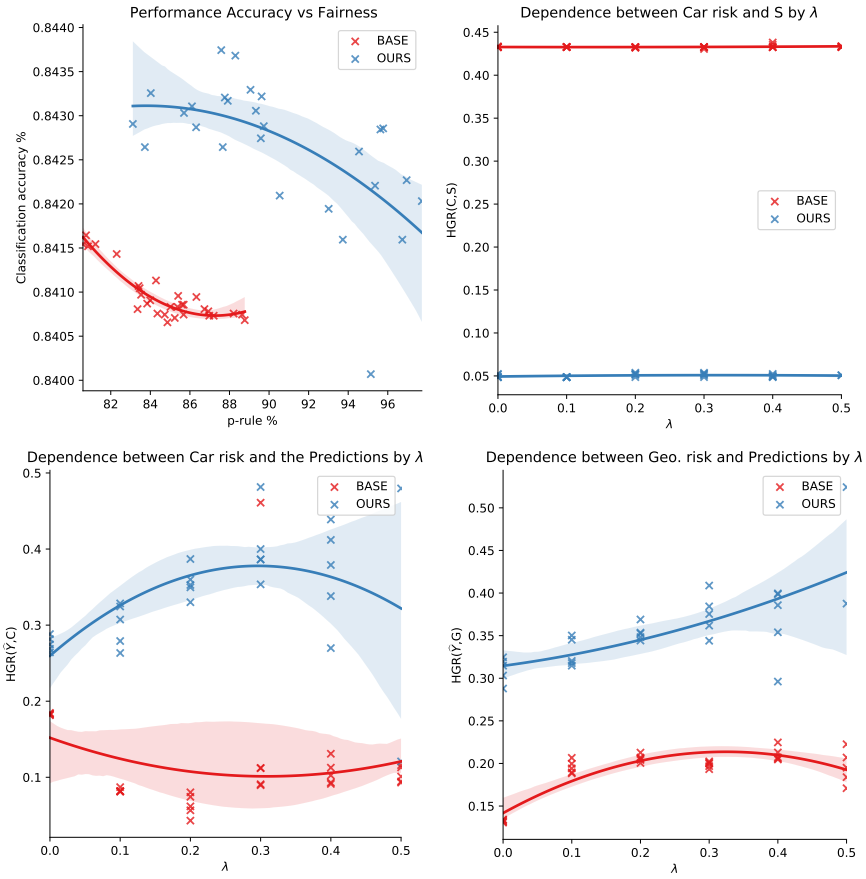


Figure 7.6: Scenario 1 - Binary objective - Pricingame 2015 dataset (similar results as synthetic scenario)

As the synthetic scenario, we compare for this objective, the fair traditional pricing structure denoted as *BASE* and the fair pricing autoencoder structure as shown in the subsection 7.2.2 denoted as *OURS*. For all data sets, we repeat five experiments by randomly sampling two subsets, 80% for the training set and 20% for the test set.

Depending on the objective, we report the average of the accuracy (*ACC*), the Expected Deviance Ratio (EDR), the average of the mean squared error (*MSE*), the *GINI*, the mean of the *HGR* metric HGR_{NN} and the *FairQuant* metric on the test set. We plot the performance of these different approaches by displaying the predictive performance (*ACC*, *EDR*, *MSE* and *GINI*) against the *FairQuant* for Demographic Parity in the two left most graphs in Figures 7.6,7.8,7.9 and with separate mistreatment for each cases of the target Y for Equalized Odds (Figure 7.10). For all algorithms, we clearly observe that the predictive performance decreases when fairness

increases. We note that, for all levels of fairness (controlled by the mitigation weight in every approach), our method outperforms the traditional algorithm for both fairness tasks. We attribute this result, as in the synthetic scenario, to the ability of our approach to extract useful components from the mitigated car and geographic risks. In contrast, the traditional approach suffers significantly from merely mitigating biases on the predictor model only. We observe in the middle right graphs of the Figures 7.6, 7.9 that the dependencies between the car risks and the predictions are all more important than the traditional one, and this is valid for all levels of fairness. In Figure 7.7 we plot the distribution of the predicted probabilities for each demographic group of the sensitive attribute S for 3 different models: An unfair model ($\lambda = 0$), and a mildly fair model ($\lambda = 1.1$) and a strongly fair one ($\lambda = 2.0$). For the unfair model, the distribution differs most for the lower probabilities. As expected, we observe that the distributions are more aligned with a higher λ. Figure 7.10 compares the results of (a) the demographic parity task and (b) the equalized odds task. We plot for this purpose the predictive performance (measured by *MSE*) against the fairness metric (measured by *FairQuant*) with separate mistreatment for each case of the target $Y = \{0, 1, 2^+\}$. For all cases, our method outperforms the traditional one. We also observe that for the demographic parity task and in particular when $Y = 0$, the bias mitigation is more pronounced (*FairQuant* closer to 0) than in other cases ($Y = \{1, 2^+\}$). Most observations have no claims (87.7%), therefore, the weight for the 0 case should be higher. In contrast, our proposed equalized odds strategy by separating the cases for different λ shows a significant improvement for the cases with claims ($Y = \{1, 2^+\}$). The *FairQuant* is closer to zero in these cases, except for the traditional one where the *FairQuant* stabilizes at 0.003 for $Y = 1$. Note that this is not at the sacrifice of the non-claimed observations where the results are close to the demographic parity results.

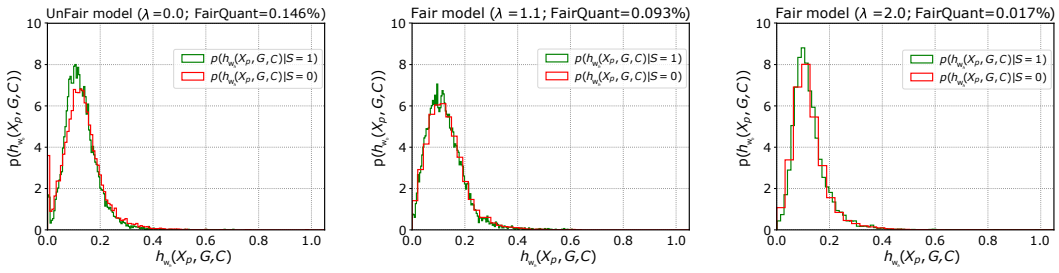


Figure 7.7: Distributions of the predicted probabilities given the sensitive attribute S (Pricingame 2015 dataset). Higher λ values produce more aligned distributions between men and women.

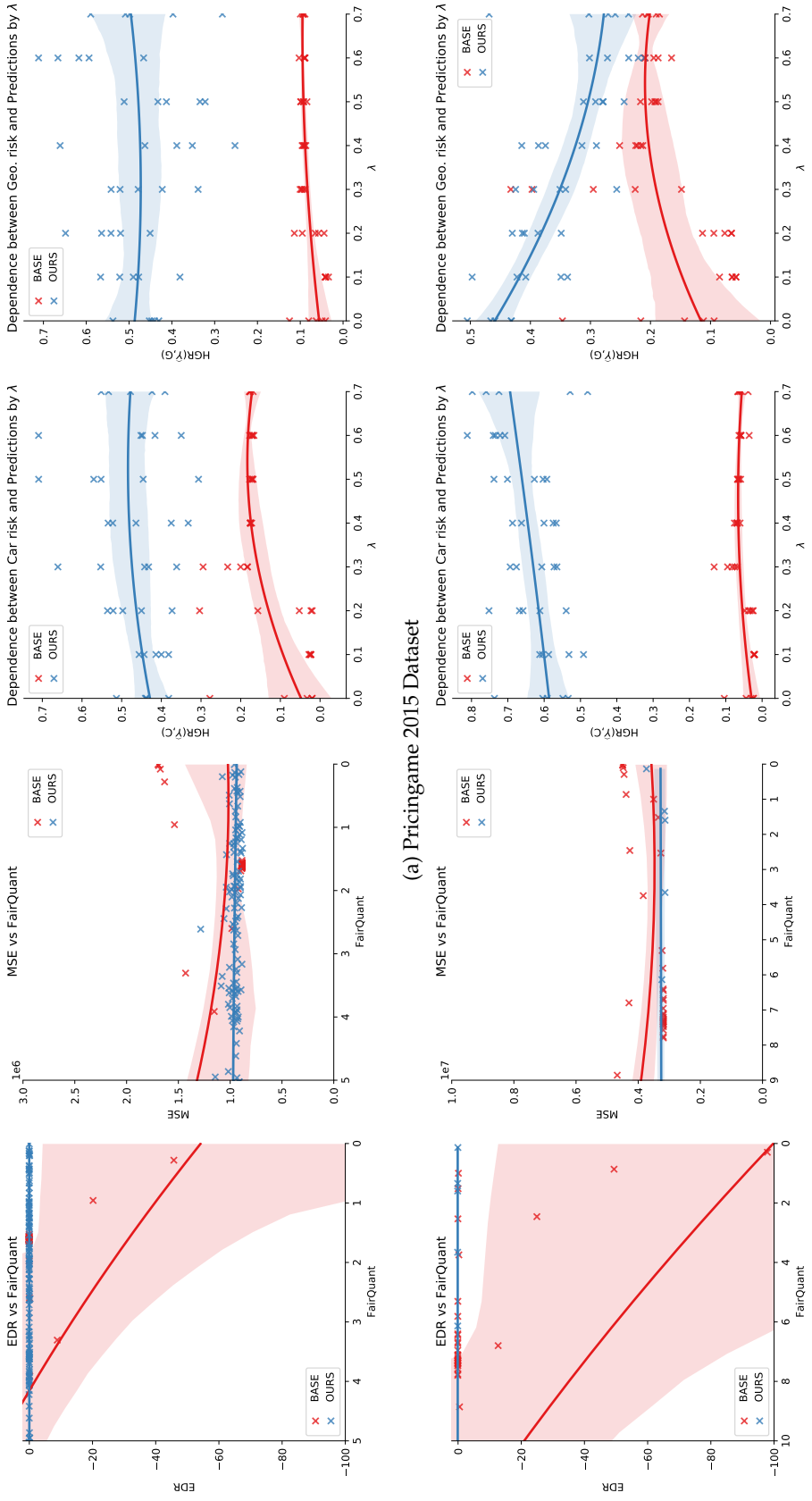


Figure 7.8: Scenario 3 - Continuous Objective (Average Cost) (Left and Middle left) For all levels of predictive performance, *EDR* or *MSE*, our proposed method outperforms the traditional fair pricing model. (Middle right and Right) Higher λ values significantly reduce the dependence between output predictions and the car risk, as well as for the geographical risk in comparison to our proposed model.

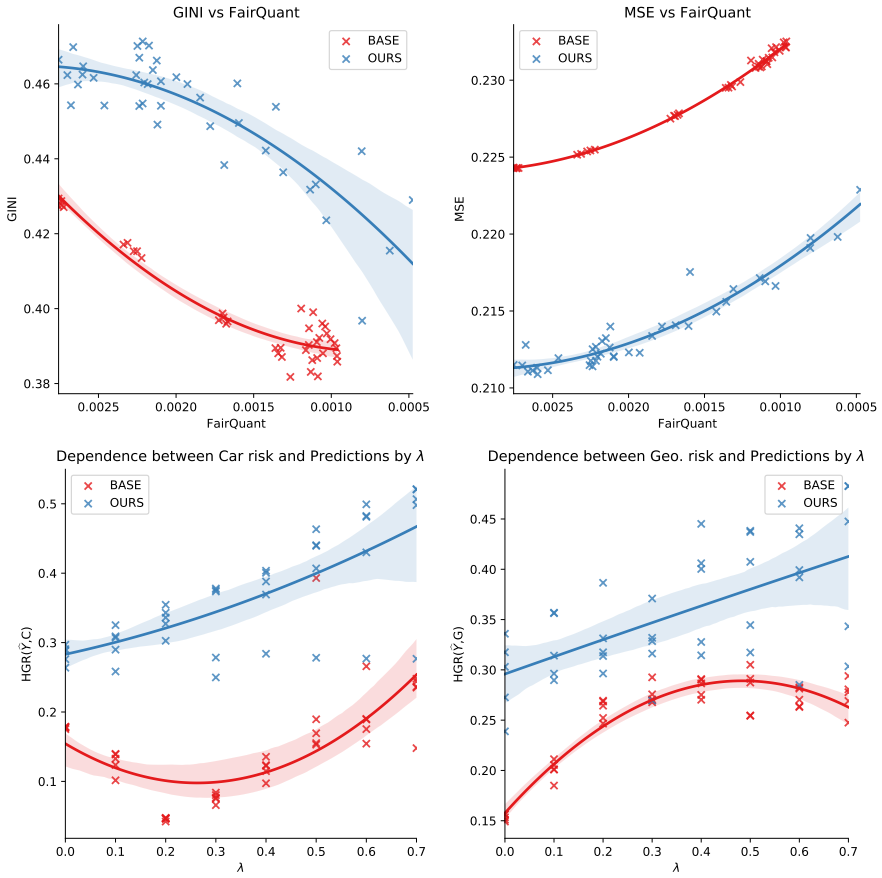
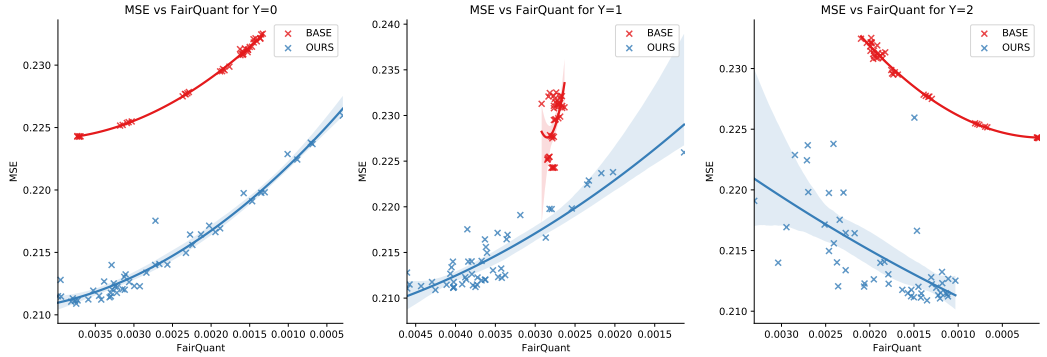
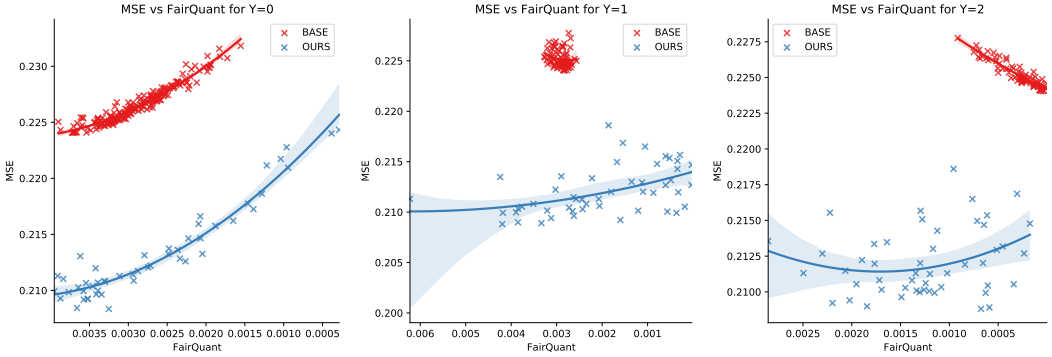


Figure 7.9: Scenario 2 - Frequency objective (Left and Middle left) For all levels of predictive performance, *GINI* or *MSE*, our proposed method outperforms the traditional fair pricing model. (Middle right and right) Higher λ values significantly reduce the dependence between output predictions and the car risk, as well as for the geographical risk in comparison to our proposed model.



(a) Demographic Parity task



(b) Equalized Odds task with $\lambda_0 = \lambda_1 = \lambda_{2+}$

Figure 7.10: Frequency task (Pricingame 2015 dataset) - The figures show the results between the prediction performance (*MSE*) against the fairness metric (*FairQuant)* for the demographic parity task (a) and for the equalized odds task (b). (Left) represents the results only for cases where $Y = 0$ (87.7% of the observations), (Middle) only for one-claim cases where $Y = 1$ (10.4% of the observations) and (Right) for high-claim cases where $Y \geq 2$ (1.9% of the observations). For all cases, our method outperforms the traditional one. We also observe that the demographic parity objective gives more weight to mitigation where there are more observations, especially for unclaimed ($Y = 0$). While for observations with claims, it seems to be more discarded. In contrast, our proposed method for equalized odds objective seems more consistent even for cases with fewer observations. We note that for the 3 cases, the fairness is very close to the optimum (*FairQuant* very close to 0).

7.4 | Conclusion

We developed a novel general framework designed specifically for insurance pricing. First, we extend autoencoders to generate the geographic and car rating components needed for pure premium prediction. To the best of our knowledge, this is the first such method to be applied in a single whole insurance pricing model, as it traditionally requires a two-step structure. Compared with the traditional pricing method, our method proves to be more efficient in terms of performance accuracy on various artificial and real data sets. We attribute this to the ability of the autoencoder to extract useful car and geographic risks, while the traditional pricing structure has significant limitations in correctly modeling the risks associated with these deep and complex components. Furthermore, this general framework has proven to be highly interesting for applying an adversarial learning approach specifically designed to improve the fairness of insurance pricing. We show empirically that the different components predicted by the model are debiased in contrast to traditional approaches that might remove the important information for predicting the pure premium. This approach shows to be more efficient in terms of accuracy for similar levels of fairness on various data sets. As future work, it might be interesting to consider a generalization of our proposal for telematics insurance where some biases can be mitigated on different aggregation scores.

8

Individual Fairness

This chapter considers two ways of looking at individual fairness.

First, we are interested to look at the link between the traditional adversarial group fairness models (discussed in section 4) and models from individual fairness through awareness. Most of the current approaches involves comparing individuals from a distance $d_{\mathcal{X}}$. This distance can come from expert knowledge, or can be learnt from linear projection on data as discussed in Section 2.2.2. We will see that fair representations algorithms like (Adel et al., 2019; Grari et al., 2021b) mitigate the underlying bias not only for group fairness but also on the individual sense for some implicit similarity distance $d_{\mathcal{X}}$. In this comparison, we also present a new framework method, that leverages the Variational Autoencoder (VAE) algorithm and the Hirschfeld-Gebelein-Renyi (HGR) maximal correlation coefficient for enforcing individual fairness without access to the fair distance $d_{\mathcal{X}}$. We demonstrate the effectiveness of our approach in enforcing individual fairness on several machine learning tasks prone to algorithmic bias, even compared with the same distance $d_{\mathcal{X}}$ as the one used by competitors approaches (e.g., (Yurochkin et al., 2019))

In the second section we are interested to apply fairness in a causal perspective, counterfactual fairness, which is close to the general individual fairness since it aims at building prediction models which ensure fairness at the most individual level. Rather than globally considering equity over the entire population, the idea is to imagine what any individual would look like with a variation of a given attribute of interest, such as a different gender or race for instance. Therefore, the difference with the general individual fairness is that it requires an *intervention* from the sentive feature. Existing approaches rely on Variational Auto-encoding of individuals, using Maximum Mean Discrepancy (MMD) penalization to limit the statistical dependence of inferred representations with their corresponding sensitive attributes. This enables

the simulation of counterfactual samples used for training the target fair model, the goal being to produce similar outcomes for every alternate version of any individual. In this work, we propose to rely on an adversarial neural learning approach, that is particularly well fitted for the continuous setting, where values of sensitive attributes cannot be exhaustively enumerated. Experiments show significant improvements in term of counterfactual fairness for both the discrete and the continuous settings.

8.1 | Fairness Through Awareness

The method that we propose in order to enforce individual fairness is based on Variational Autoencoding methods (VAE), which is the basis of the first step of our algorithm (Figure 8.1). While some recent work leverage VAE methods for counterfactual inference (Chiappa, 2019; Madras et al., 2019; Pfohl et al., 2019; Louizos et al., 2015), we make use of this technique to generate similar individuals via the encoding-decoding process. We learn an unobserved confounder U of which we mitigate the bias w.r.t the sensitive variable, and generate new individuals based on the latent variable U. Then, we make use of this learnt generator in the prediction step, by adding a regularization term representing differences of outputs between similar individuals.

8.1.1 | Step 1: Rényi Variational Inference

The original formulation of VAE consists in optimizing the classical lower bound (ELBO):

$$\mathcal{L}_{ELBO} = -E_{u\sim q_\phi(u|x)}(\log p_\theta(x|u)) + D_{KL}(q_\phi(u|x)||p(u)) \tag{8.1}$$

D_{KL} is defined as the Kullback-Leibler divergence, which computes distances between distributions, and the prior $p(u)$ is typically a standard Gaussian distribution. $q_\phi(u|x)$ is represented as a neural network, referred to as the encoder, (see Figure 8.1), that outputs the mean μ_ϕ and variance σ_ϕ of a Gaussian distribution, which allow to infer, stochastically, the variable U. The decoding process, corresponding to $p_\theta(x|u)$ and materialized by a neural network, consists in predicting X based on U. The overall objective consists in both minimizing the reconstruction error (first term) and the divergence (distance) with a standard Gaussian.

We adapt the ELBO by making two changes to the global step 1 objective. If we directly learnt the VAE with Eq. 8.1, we would obtain a generator that does generate,

for a given x, a similar individual, but not similar in an acceptable sense for individual fairness. The underlying notion of similarity would be closer to an euclidean one, and thus would not be necessarily fair. As discussed in (Castelnovo et al., 2022), dealing with euclidean comparison implies that any smooth predictor function h_{w_h} shall satisfy individual fairness condition (Eq. 4) and is therefore not a very interesting notion of fairness.

However, by debiasing the latent variable U, generated individuals could be similar to the original one (depending on the level of debiasing) in a fairer sense. The assumption is that, if U both adequately represents the input data and eliminates information about the sensitive attribute, individuals generated from U will be close to the original one in terms of debiased attributes. As shown in previous sections, a good candidate for debiasing the latent variable is the HGR, which has the ability to capture non-linear dependencies. We add the term $HGR(U, S)$ to Eq.8.1, with a hyperparameter λ_{HGR} that controls the trade-off between debiasing and reconstruction, and therefore use the HGR_NN to estimate the HGR coefficient. Also, similarly to (Pfohl et al., 2019), we replace the KL-divergence with a Maximum Mean Discrepancy (MMD) term (Gretton et al., 2012) $\mathcal{L}_{MMD}(q_\phi(u)||p(u))$. As shown in (Zhao et al., 2017; Chen et al., 2016), the D_{KL} can be too restrictive (uninformative latent code problem) and tends to overfit the data.

Therefore, the final minimization objective for the VAE step is:

$$\begin{aligned}\mathcal{L}_{VAE} = &- E_{u\sim q_\phi(u|x)}(\log p_\theta(x|u)) \\ &+ \lambda_{MMD}\mathcal{L}_{MMD}(q_\phi(u)||p(u)) \\ &+ \lambda_{HGR}HGR(U, S)\end{aligned} \tag{8.2}$$

This heuristic approach does not make use of any given distance $d_{\mathcal{X}}$, but actually learns its own notion of similarity: individuals can be considered similar if they are close, in an euclidean sense, on the latent space. Additionally, we can also note that a significant difference with counterfactual inference is that the sensitive attribute is not used as input for the VAE. In counterfactual inference, having the sensitive attribute as input allows to generate counterfactual individuals with different values of S.

8.1.2 | Step 2: Individually Fair Prediction

Once the inference model is learnt, we can generate similar versions of each training individual, and use them to learn an individually fair predictive function h_{w_h}. The global objective function for step 2 is:

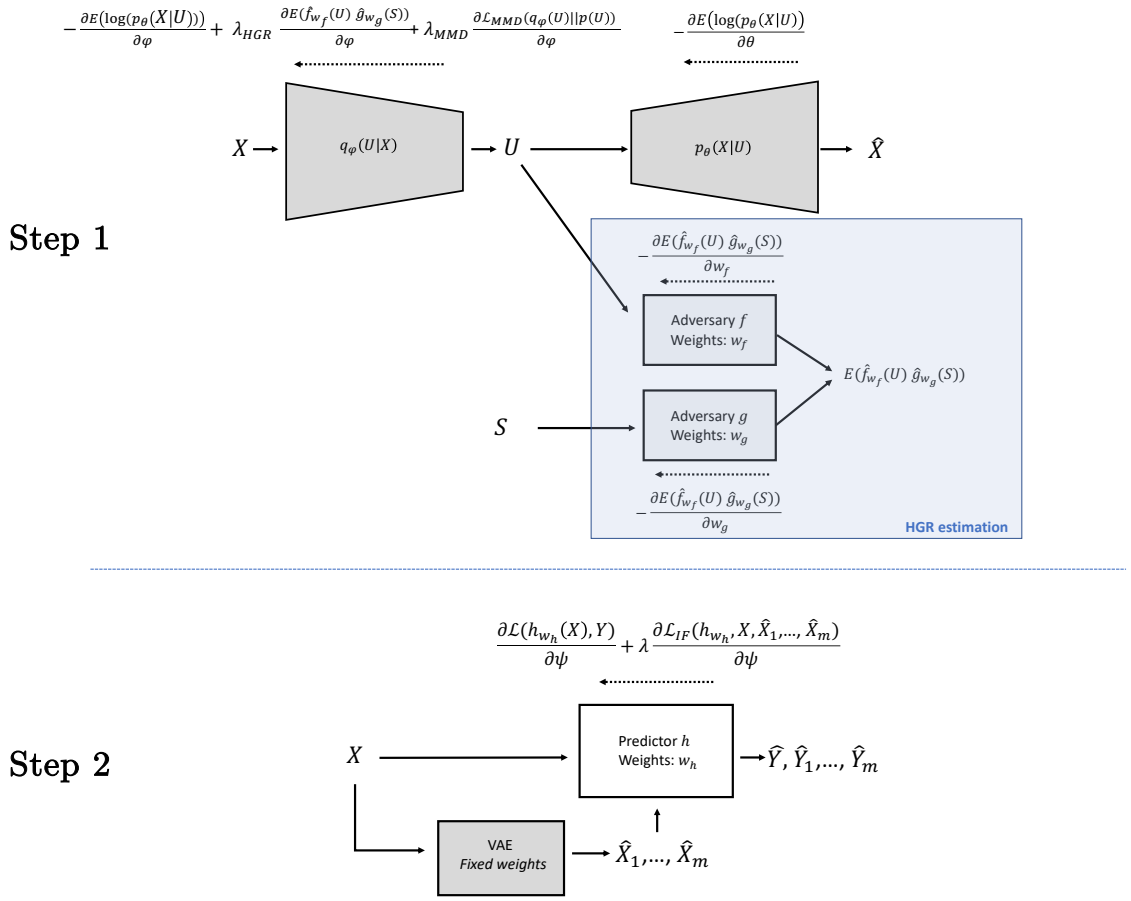


Figure 8.1: VAE-IF - Two steps algorithm for individual fairness

$$\mathcal{L}_{global} = \mathcal{L}(h_{w_h}(X), Y) + \lambda \mathcal{L}_{IF}(h_{w_h}, X, \widehat{X_1}, \ldots \widehat{X_m}) \tag{8.3}$$

where $\mathcal{L}$ is a suitable loss function for the problem at hand (regression or classification), $\widehat{X_1}, \ldots \widehat{X_m}$ are m individuals generated by the inference model learnt at step 1 and similar to X, and $\mathcal{L}_{IF}$, referred to as the Individual Fairness Loss, can be defined as:

$$\mathcal{L}_{IF}(h_{w_h}, X, \widehat{X_1}, \ldots \widehat{X_m}) = \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}(h_{w_h}(\widehat{X_i}), h_{w_h}(X)) \tag{8.4}$$

$\mathcal{L}_{IF}$ is a regularization term that, if minimized, ensures that similar individuals (in the sense of the step 1 VAE) have similar outputs. The hyperparameter λ controls the

trade-off individual fairness/accuracy. Therefore, for sufficiently high values of λ, the algorithm is individually fair w.r.t a distance $d_{\mathcal{X}}$ if the generated individuals are close to X in the sense of $d_{\mathcal{X}}$. Since we assume no access to $d_{\mathcal{X}}$, we do not directly enforce, in step 1, closeness of generated individuals in the sense of $d_{\mathcal{X}}$, and we therefore have to rely on experimental results to assess the individual fairness of our algorithm w.r.t to a reasonable distance $d_{\mathcal{X}}$ for the task at hand.

8.1.3 | Experiments

We empirically evaluate the performance of our contribution on 3 real world data sets. For the discrete scenario and specifically in the binary case ($Y \in \{0,1\}, S \in \{0,1\}$), we use 3 different popular data sets: For the discrete scenario and specifically in the binary case ($Y \in \{0,1\}, S \in \{0,1\}$), we use the popular Adult UCI data set. For the continuous setting (Y and S are continuous), we use the data sets Motor and Crime data set. Please not that all of these data set are with the same setting as Subsection 4.5.

Metrics and experimental conditions

We repeat five experiments with random 80%/20% train-test splits. We report the averages of mean squared error (MSE) or balanced accuracy (B-Acc%), the Individual Fairness Loss (IFL) defined in 8.4, the MRD and MDRD metrics defined in subsection 3.3.1 (choices of α and β are reported in the Appendix in section D.1), ΔDP for binary sensitives or the HGR for continuous sensitives. Therefore, we can assess both the individual fairness of algorithms w.r.t the intrinsic metric of the step 1 VAE (IFL) and the distance $d_{\mathcal{X}}$ (MRD, MDRD). The latter is learned with the subspace method of (Yurochkin et al., 2019) presented in chapter 3. We fit a logistic regression model to predict the sensitive S, and we calculate the projection matrix onto the span of A corresponding to the weight of the fitted model (more information in subsection 2.2.2). While this distance deals with a linear link, it is still reasonable for our comparison since it also compares individuals independently of the sensitive S.

We can also assess the group fairness with the HGR and ΔDP, as it is interesting to observe whether models trained for individual fairness perform well in terms of group fairness. Note that, for the Adult experiment, we report 2 group fairness metrics.

The baseline we use is a classic deep neural network (Standard NN). We compare our method with state of the art algorithms, among which the SenSR (Yurochkin

et al., 2019) algorithm which enforces Individual Fairness with knowledge of the fair distance $d_{\mathcal{X}}$, as well as group fairness algorithms. Among the group fairness algorithms, we compare our approach to two adversarial methods (Zhang et al., 2018; Adel et al., 2019) which both involve an adversary that aims at predicting the sensitive attribute (from either a representation of the input, or from the prediction outputs). We also compare our method to the bias mitigation method introduced in section 4.5 for group fairness.

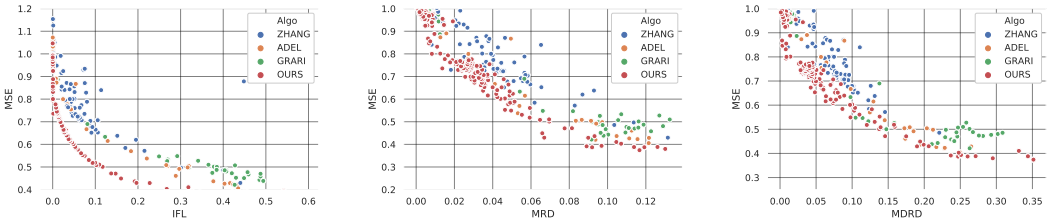
All hyper-parameters for every approach were tuned by 5-fold cross-validation. For the Adult UCI data set for our approach for instance, in step 1, the encoder q_ϕ architecture is an MLP of only one hidden layers with 256 units and a ReLU activation. The latent variable U is a 20 dimensional vector. On this dataset, the decoder p_θ is an MLP of also one hidden layer with 256 units and a ReLu activation function and the output consists in 95 units to reconstruct X (number of features). The two sub-networks f and g in the HGR adversarial neural network are both made of three hidden layers with 15 units. Notice that f takes only a one-dimensional input which corresponds to the sensitive feature S and g takes a 20 dimensional input which corresponds to the latent space U. In step 2, the predictor corresponds to a MLP with three hidden layers of 64, 32 and 8 units respectively with ReLu activation functions and one single output node with a sigmoid activation to reconstruct Y.

For the step 1, we used a Gaussian radial basis function kernel for the MMD constraint. For the Motor dataset, the prior distribution $p(U)$ considered for training the models is a 20-dimensional standard Gaussian and, for Crime, it is a 5-dimensional standard Gaussian.

Results

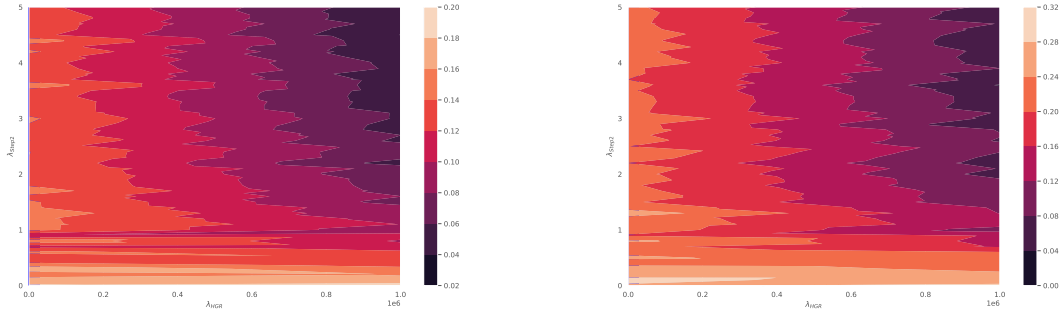
Results can be found in tables 8.1 and 8.2. For all of the fair algorithms, we attempted to obtain comparable predictive performances by giving similar balanced accuracies (B-Acc%) for classification or mean squared errors (MSE) for regression for all the algorithms, in the same settings. Each of the algorithms considered has a hyperparameter that allows to balance the relative importance of accuracy and fairness while learning. Best performances among fair algorithms are in bold.

For the three data sets, our algorithm enforces individual fairness at the best level among the other algorithms, not only with respect to its inner metric (IFL), but also and most importantly with respect to the $d_{\mathcal{X}}$ distance as shown by the MRD and MDRD metrics. In particular, our method VAE-IF outperforms SenSR in individual



(a) (b) (c)

Figure 8.2: Pareto fronts for IFL, MRD, MDRD. Higher values of λ produce fairer predictions, while λ near 0 only focuses on optimizing the predictor.



(a) Iso-MRD surfaces between λ_{HGR} and λ_{step2} (b) Iso-MDRD surafces between λ_{HGR} and λ_{step2}

Figure 8.3: Impact of the hyperparameters λ_{HGR} of the step 1 and the λ_{step2} of the step 2 on individual fairness metrics for the Crime data set.

		B-Acc%	IFL	MRD	MDRD	ΔDP_G	ΔDP_R
Adult	Standard NN	82.6% $\pm$ 0.2	0.239 $\pm$ 0.028	0.031 $\pm$ 0.012	0.052 $\pm$ 0.010	0.249 $\pm$ 0.002	0.129 $\pm$ 0.001
	SenSR (Yurochkin et al., 2019)	79.2% $\pm$ 0.7	0.327 $\pm$ 0.057	0.020 $\pm$ 0.008	0.015 $\pm$ 0.010	**0.143** $\pm$ 0.023	0.110 $\pm$ 0.016
	Simple PR (Zhang et al., 2018)	79.1% $\pm$ 2.3	0.362 $\pm$ 0.054	0.026 $\pm$ 0.019	0.081 $\pm$ 0.029	0.264 $\pm$ 0.042	0.126 $\pm$ 0.032
	Simple FR (Adel et al., 2019)	79.9% $\pm$ 0.6	0.348 $\pm$ 0.013	0.013 $\pm$ 0.008	0.023 $\pm$ 0.022	0.209 $\pm$ 0.016	0.110 $\pm$ 0.006
	Rényi FR (Grari et al., 2021b)	80.1% $\pm$ 0.9	0.321 $\pm$ 0.011	0.022 $\pm$ 0.004	0.018 $\pm$ 0.027	0.154 $\pm$ 0.015	**0.089** $\pm$ 0.003
	VAE-IF (Grari et al., 2021a)	**80.2**% $\pm$ 0.6	**0.001** $\pm$ 0.001	**0.010** $\pm$ 0.006	**0.010** $\pm$ 0.013	0.249 $\pm$ 0.040	0.113 $\pm$ 0.019

Table 8.1: Experimental results for the discrete dataset

fairness while having better accuracy, even knowing that the SenSR algorithm makes use of the fair distance $d_{\mathcal{X}}$. Unsurprisingly, in terms of group fairness metrics, our algorithm does not perform as well as the group fairness algorithms such as our Rényi fair representation (*Rényi FR*) as shown with the ΔDP_R and the HGR for the three data sets. Our algorithm still achieves group fairness debiasing (better than baseline) except for the Gender variable in the Adult experiment. Note that simple fair representation (*Simple FR*) (Adel et al., 2019) and our *Rényi FR* obtain good performances on individual fairness metrics even though these methods were initially intended for

		MSE	IFL	MRD	MDRD	HGR
Motor	Standard NN	0.946 ± 0.003	0.023 ± 0.005	0.043 ± 0.004	0.118 ± 0.026	0.208 ± 0.045
	SenSR (Yurochkin et al., 2019)	0.996 ± 0.029	0.039 ± 0.005	0.065 ± 0.003	0.209 ± 0.031	0.168 ± 0.024
	Simple PR (Zhang et al., 2018)	0.976 ± 0.016	0.042 ± 0.020	0.116 ± 0.026	0.139 ± 0.045	0.196 ± 0.015
	Simple FR (Adel et al., 2019)	0.981 ± 0.009	0.001 ± 0.001	0.009 ± 0.004	**0.028** ± 0.012	0.150 ± 0.076
	Rényi FR (Grari et al., 2021b)	**0.972** ± 0.004	0.008 ± 0.003	0.019 ± 0.005	0.042 ± 0.014	**0.079** ± 0.018
	VAE-IF (Grari et al., 2021a)	**0.972** ± 0.001	**0.001** ± 0.000	**0.008** ± 0.001	**0.028** ± 0.008	0.169 ± 0.022
Crime	Standard NN	0.387 ± 0.008	0.448 ± 0.080	0.185± 0.015	0.164 ± 0.095	0.772 ± 0.022
	SenSR (Yurochkin et al., 2019)	0.999 ± 0.084	0.379 ± 0.028	0.123 ± 0.041	0.274 ± 0.063	0.608 ± 0.061
	Simple PR (Zhang et al., 2018)	0.990 ± 0.069	0.051 ± 0.012	0.014 ± 0.004	0.028 ± 0.020	0.496 ± 0.031
	Simple FR (Adel et al., 2019)	0.996 ± 0.021	0.001 ± 0.000	**0.002** ± 0.000	**0.002** ± 0.001	0.549 ± 0.012
	Rényi FR (Grari et al., 2021b)	1.001 ± 0.017	**0.000** ± 0.000	**0.002** ± 0.001	**0.002** ± 0.002	**0.254** ± 0.096
	VAE-IF (Grari et al., 2021a)	**0.972** ± 0.004	**0.000** ± 0.000	**0.002** ± 0.001	**0.002** ± 0.001	0.531 ± 0.050

Table 8.2: Experimental results for the continuous dataset

group fairness. This is due to the fact that they both predict the outcome from a debiased representation. The prediction module being a lipschitz function, these two methods give similar predictions for individuals with similar representations.

In Figure 8.2, we plot 3 Pareto fronts displaying the IFL, MRD and the MDRD against the MSE with different values of the hyperparameter λ. These plots were obtained on the Crime data set with 4 algorithms: VAE-IF, *Simple FR* , *Simple FR* and *Rényi FR*. Varying the hyperparameter λ allows to control the fairness/accuracy trade-off. Here, we clearly observe for all algorithms that the MSE, or predictive performance, decreases when fairness increases. Higher values of λ produce fairer predictions w.r.t the three fairness metrics, while near 0 values of the hyperparameter λ result in the optimization of the predictor loss with no fairness consideration. We note that, for all levels of predictive performance, our method outperforms the state of the art algorithms in terms of the three fairness metrics. The gap is even higher (in our favour) on the IFL metric, since our method consists in mitigating this latter metric.

In Figure 8.3, we plot the iso-MRD and iso-MDRD surfaces in the $(\lambda_{HGR}, \lambda_{step2})$ plane, λ_{HGR} being the hyperparameter in step 1 and λ_{step2} the regularization hyperparameter in the prediction step. We can make several observations based on these two plots. First, for any value of λ_{HGR}, if the value of λ_{step2} is not high enough, then MRD and MDRD are not mitigated. That is to say: adding importance to the regularization term in step 2 improves fairness. Second, these plots allow us to assess the performance of the step 1 of the algorithm. Indeed, even for very high values of λ_{step2}, if λ_{HGR} is not high enough, then both the MRD and MDRD are not mitigated. This highlights the necessity of proper tuning of the λ_{HGR} in step 1. The more step 1 mitigates the bias in the confounder, the more the algorithm is individually fair. Note that the optimal values of MRD and MDRD are obtained for high values of λ_{HGR} and

λ_{step2}.

8.2 | Adversarial Counterfactual Fairness

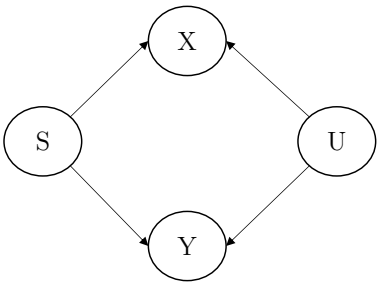


Figure 8.4: Graphical causal model. Unobserved confounder U has effect on both X and Y.

In this section, we focus on an other sub field of individual fairness: *Counterfactual Fairness*. In particular, counterfactual fairness aims at imagine what any individual would look like with a variation of a given attribute of interest, such as a different gender or race for instance. As discussed in section 2.2.2, it involves inferring causal effects from a causal model. Existing approaches rely on Variational Auto-encoding of individuals and use a Maximum Mean Discrepancy (MMD) penalization to limit the statistical dependence of inferred representations with their corresponding sensitive attributes. This enables the simulation of counterfactual samples used for training the target fair model, the goal being to produce similar outcomes for every alternate version of any individual. In this work, we propose to rely on an adversarial neural learning approach, that enables more powerful inference than with MMD penalties, and is particularly better fitted for the continuous setting, where values of sensitive attributes cannot be exhaustively enumerated. Experiments show significant improvements in term of counterfactual fairness for both the discrete and the continuous settings.

8.2.1 | Background

We focus on the classical causal graph depicted in Fig.8.4, often used in the counterfactual fairness literature (Kusner et al., 2017; Pfohl et al., 2019; Chiappa, 2019), which can apply for most applications. For more specific tasks, note further that our approach could be easily adapted for different graphs, such as those explored in (Kusner et al., 2017) for instances. In this causal graph, both input X and outcome Y only depend on the sensitive attribute S and a latent variable U, which represents

all the relevant knowledge non dependent on the sensitive feature S. In that setting, the knowledge of U can be used during training to simulate various versions of the same individual, corresponding to different values of S, in order to obtain a predictive function h_{w_h} which respects the fairness objective from definition 5. For any training individual, U has to be inferred since only X, S and Y are observed. This inference must however ensure that no dependence is created between U and S (no arrow from U to S in the graph from Fig.8.4), unless preventing the generation of proper alternative versions of X and Y for any values S.

Concerning causal effect identifiability (i.e., whether a joint distribution of latent and observed confounder variables can be uniquely inferred from observations), sufficient conditions as raised in (Louizos et al., 2017; Madras et al., 2019; Kilbertus et al., 2020) imply strong assumptions which require specific directed acyclic graphs different from ours. As in (Pfohl et al., 2019), that considers the same causal graph, we make no formal guarantee on identification even in the case where these assumptions hold (more information in their article). However, we argue that, given any distribution $P(U, S, X, Y)$ exactly inferred from a sufficiently large amount of observations (X, Y, S), with a constant prior on U, the counterfactual quantities $P(X_{S\leftarrow s'}|X, Y, S)$ and $P(Y_{S\leftarrow s'}|X, Y, S)$ are identifiable, whenever U is independent from S. From this, if the prior $P(U)$ is the true one, and the decoding is sufficiently powerful, a classifier can be trained to minimize counterfactual unfairness according to the inferred model (step 2 in the following).

Several current approaches (Louizos et al., 2017; Kim et al., 2021; Kocaoglu et al., 2018; Xu et al., 2019) enforce fairness on counterfactual data generated by their model. These works, which do not focus on the final predictor itself, assume that giving fair generated counterfactual observations as input to a traditional machine learning algorithm is sufficient to maintain the fairness objective. We argue that it is not always the case and the final predictions need to be evaluated to ensure a good fairness level. For this reason, we rather leverage a two-step method, as already considered in (Russell et al., 2017; Pfohl et al., 2019), that focus separately on Causal Inference (step 1) and Model Learning (step 2). We develop and discuss the general principles of this family of methods in the following.

Step 1: Counterfactual Inference

The goal is to define a way to generate counterfactual versions of original individuals. As discussed above, this is usually done via approximate Bayesian inference, according to a pre-defined causal graph.

The initial idea to perform inference was to suppose with strong hypothesis a non deterministic structural model with some specific distribution for all the causal links (Kusner et al., 2017). In this setting, the posterior distribution of U was estimated using the probabilistic programming language Stan (Team et al., 2016). Then, leveraging recent developments for approximate inference with deep learning, many works (Chiappa, 2019; Pfohl et al., 2019; Madras et al., 2019; Louizos et al., 2017) proposed to use Variational Autoencoding (Kingma and Welling, 2013) methods (VAE) to generalize this first model and capture more complex - non linear - dependencies in the causal graph.

Following the formulation of VAE, it would be possible to directly optimize the classical lower bound (ELBO) (Kingma and Welling, 2013) on the training set $\mathcal{D}$, by minimizing:

$$\mathcal{L}_{ELBO} = - \mathbb{E}_{\substack{(x,y,s)\sim\mathcal{D}, \\ u\sim q_\phi(u|x,y,s)}} [\log p_\theta(x,y|u,s)] + D_{KL}(q_\phi(u|x,y,s)||p(u)) \tag{8.5}$$

where D_{KL} denotes the Kullback-Leibler divergence of the posterior $q_\phi(u|x,y,s)$ from a prior $p(u)$, typically a standard Gaussian distribution $\mathcal{N}(0,I)$. The posterior $q_\phi(u|x,y,s)$ is represented by a deep neural network with parameters ϕ, which typically outputs the mean μ_ϕ and the variance σ_ϕ of a diagonal Gaussian distribution $\mathcal{N}(\mu_\phi,\sigma_\phi I)$. The likelihood term factorizes as $p_\theta(x,y|u,s) = p_\theta(x|u,s)p_\theta(y|u,s)$, which are defined as neural networks with parameters θ. Since attracted by a standard prior, the posterior is supposed to remove probability mass for any features of the latent representation U that are not involved in the reconstruction of X and Y. Since S is given together with U as input of the likelihoods, all the information from S should be removed from the posterior distribution of U.

However, many state of the art algorithms (Chiappa, 2019; Louizos et al., 2017; Madras et al., 2019; Pfohl et al., 2019) show that the independence level between the latent space U and the sensitive variable S is insufficient with this classical ELBO optimization. Some information from S leaks in the inferred U. In order to ensure a high level of independence, a specific TARNet (Shalit et al., 2017) architecture can be employed (Madras et al., 2019) or a penalisation term can be added in the loss function. For example, (Chiappa, 2019; Pfohl et al., 2019) add a Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) constraint. The MMD term can be used to enforce all the different aggregated posterior to match the prior distribution(Pfohl et al., 2019): $\mathcal{L}_{MMD}(q_\phi(u|S=s_k)||p(u))$ for all $s_k \in \Omega_S$ (referred to as MMD wrt $P(U)$ in the following). Alternatively, the constraint can directly enforce the matching between pairs of posteriors (Chiappa, 2019): $\mathcal{L}_{MMD}(q_\phi(u|S=s_k)||q_\phi(u|S=s))$ for all

$s_k \in \Omega_S$, with s standing for the original sensitive value of the considered individual (referred to as MMD wrt U_s in the following). Notice that while this additional term can improve independence, it can also encourage the model to ignore the latent confounders U, by being too restrictive. One possible approach to address this issue is to apply weights λ (hyperparameters) to control the relative importance of the different terms. In addition, as already done in chapter 6, we employ a variant of the ELBO optimization, where the $D_{KL}(q_\phi(u|x,y,s)||p(u))$ term is replaced by a MMD term $\mathcal{L}_{MMD}(q_\phi(u)||p(u))$ between the aggregated posterior $q_\phi(u)$ and the prior.

Finally, the inference for counterfactual fairness can be optimized by minimizing (Pfohl et al., 2019):

$$\begin{aligned}\mathcal{L}_{CE-VAE} = - \mathop{\mathbb{E}}_{\substack{(x,y,s)\sim\mathcal{D},\\ u\sim q_\phi(u|x,y,s)}} \left[\begin{array}{l} \lambda_x \log(p_\theta(x|u,s)) + \\ \lambda_y \log(p_\theta(y|u,s)) \end{array} \right] \\ + \lambda_{MMD}\, \mathcal{L}_{MMD}(q_\phi(u)||p(u)) \\ + \lambda_{ADV}\, \frac{1}{m_s} \sum_{s_k \in \Omega_S} \mathcal{L}_{MMD}(q_\phi(u|s=s_k)||p(u))\end{aligned} \tag{8.6}$$

where λ_x, λ_y, λ_{MMD}, λ_{ADV} are scalar hyperparameters and $m_s = |\Omega_S|$.

Note that we chose to present all models with a generic inference scheme $q(U|X,Y,S)$, while most approaches from the literature only consider $q(U|X,S)$. The use of Y as input is allowed since U is only used during training, for generating counterfactual samples used to learn the predictive model in step 2. Various schemes of inference are considered in our experiments (section 8.2.4).

Step 2: Counterfactual Predictive Model

Once the causal model is learned, the goal is to use it to learn a fair predictive function h_{w_h}, by leveraging the ability of the model to generate alternative versions of each training individual. The global loss function, $\mathcal{L}_{GL}$, is usually composed of the traditional predictor loss $l(h_{w_h}(x_i,s_i),y_i)$ (e.g. cross-entropy for instance i) and a counterfactual unfairness estimation term $\mathcal{L}_{\mathcal{CF}}(w_h)$:

$$\mathcal{L}_{GL}(w_h) = \frac{1}{m}\sum_{i}^{m} l(h_{w_h}(x_i,s_i),y_i) + \lambda \mathcal{L}_{\mathcal{CF}}(w_h) \tag{8.7}$$

where λ is an hyperparameter which controls the impact of the counterfactual loss in the optimization. The counterfactual loss $\mathcal{L}_{\mathcal{CF}}(w_h)$ considers differences of predictions for alternative versions of any individual. For example, (Russell et al., 2017)

considers the following Monte-Carlo estimate from B samples for each individual i and each value $s \in \Omega_S$:

$$\mathcal{L}_{\mathcal{CF}}(w_h) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{m_s}\sum_{s_k\in\Omega_S}\frac{1}{B}\sum_{b=1}^{B}\Delta_{s_k}^{i,b} \tag{8.8}$$

where $\Delta_{s_k}^{i,b} = \Delta(h_{w_h}(x_{i,S\leftarrow s_i}^b, s_i), h_{w_h}(x_{i,S\leftarrow s_k}^b, s_k))$ is a loss function that compares two predictions, $x_{i,S\leftarrow s}^b$ denotes the b-th sample from the causal model for the i-th individual of the training set and the sensitive attribute value s. Following the causal model learned at step 1, $x_{i,S\leftarrow s}^b$ is obtained by first inferring a sample u from $q_\phi(u|x_i, s_i, y_i)$ and then sampling $x_{i,S\leftarrow s}^b$ using $p_\theta(x|u, s)$ with the counterfactual (or factual) attribute value s. According to the task, Δ can take various forms. For binary classification, it can correspond to a logit paring loss as done in (Pfohl et al., 2019): $\Delta(z, z') = (\sigma^{-1}(z) - \sigma^{-1}(z'))^2$, where σ^{-1} is the logit function. For continuous outcomes, it can simply correspond to a mean squared difference.

Discussion

For now, state-of-the-art approaches have focused specifically on categorical variables S. Unfortunately, the classical methodology for CounterFactual Fairness as described above cannot be directly generalized for continuous sensitive attributes, because the two steps involve enumerations of the discrete counterfactual modalities s_k in the set Ω_S. Particularly in step 1, sampling S from a uniform distribution for approximating the expectation $E_{s\sim p(S)}\mathcal{L}_{MMD}(q_\phi(u|S = s)||p(u))$ is not an option since this requires to own a good estimation of $q_\phi(u|S = s)$ for any $s \in \Omega_S$, which is difficult in the continuous case. While such a posterior can be obtained for discrete sensitive attributes (at least when $|\Omega_S| << m$) by aggregating the posteriors $q_\phi(u|x_i, s_i, y_i)$ over training samples i such that $s_i = s$, such a simple aggregation over filtered samples is not possible for continuous attributes. On the other hand existing approaches based on MMD costs imply to infer codes U from a distribution that takes S as input, in order to be able to obtain the required aggregated distributions via: $q_\phi(u|s) = \mathbb{E}_{p_{data}(x,y|s)}[q_\phi(u|x, y, s)]$. Omitting S from the conditioning of the generator would correspond to assume the mutual independence of u and s given x and y, which is usually wrong. On the other hand, passing S to the generator of U can encourage their mutual dependency in some settings, as we observe in our experiments.

8.2.2 | Adversarial Learning for Counterfactual Fairness

In this section we revisit the 2 steps shown above by using adversarial learning rather than MMD costs for ensuring Counterfactual Fairness. Our contribution covers a broad range of scenarios, where the sensitive attribute S and the outcome value Y can be either discrete or continuous.

Step 1: Counterfactual Inference

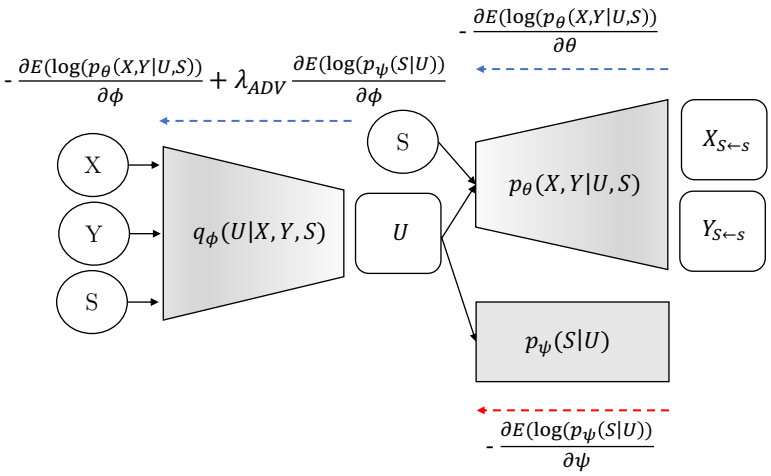


Figure 8.5: Architecture of our Counterfactual inference process. Blue arrows represent the retro-propagated gradients for the minimization of the global objective. The red one corresponds to the gradients for the adversarial optimization. Circles are observed variables, squares are samples from the neural distributions.

To avoid the comparison of distributions for each possible sensitive value, which reveals particularly problematic in the continuous setting, we propose to employ an adversarial learning framework, which allows one to avoid the enumeration of possible values in Ω_S. The idea is to avoid any adversarial function to be able to decode S from the code U inferred from the encoder q_ϕ, which allows one to ensure mutual independence of S and U. This defines a two-players adversarial game, where the goal is to find parameters ϕ which minimize the loss to reconstruct X and Y, while maximizing the reconstruction loss of S according to the best decoder $p_\psi(S|U)$:

$$\arg\min_{\theta,\phi}\max_{\psi}\mathcal{L}_{ADV}(\theta,\phi,\psi) \tag{8.9}$$

with, for the graphical causal model from figure 8.4:

$$\begin{aligned}\mathcal{L}_{ADV}(\theta,\phi,\psi) = -\underset{\substack{(x,y,s)\sim\mathcal{D},\\ u\sim q_\phi(u|x,y,s)}}{\mathbb{E}}&\left[\begin{array}{l}\lambda_x \log(p_\theta(x|u,s)) + \\ \lambda_y \log(p_\theta(y|u,s))\end{array}\right] \\ &+\lambda_{MMD}\,\mathcal{L}_{MMD}(q_\phi(u)||p(u)) \\ &+\lambda_{ADV}\underset{\substack{(x,a)\sim\mathcal{D},\\ u\sim q_\phi(u|x,y,s)}}{\mathbb{E}}[log(p_\psi(s|u))]\end{aligned} \tag{8.10}$$

where λ_x, λ_y, λ_{MMD}, λ_{ADV} are scalar hyperparameters. Compared to existing approaches presented in previous section, the difference is the last term which corresponds to the expectation of the log-likelihood of S given U according to the decoder with parameters ϕ. This decoder corresponds to a neural network which outputs the parameters of the distribution of S given U (i.e., the logits of a Categorical distribution for the discrete case, the mean and log-variance of an diagonal Gaussian in the continuous case).

All parameters are learned conjointly. Figure 8.5 gives the full architecture of our variational adversarial inference for the causal model from figure 8.4. It depicts the neural network encoder $q_\phi(U|X,Y,S)$ which generates a latent code U from the inputs X, Y and S. A neural network decoder $p_\theta(X,Y|U)$ reconstructs the original X and Y from both U and S. The adversarial network p_ψ tries to reconstruct the sensitive attribute S from the confounder U. As classically done in adversarial learning, we alternate steps for the adversarial maximization and steps of global loss minimization (one gradient descent iteration on the same batch of data at each step). Optimization is done via the re-parametrization trick (Kingma and Welling, 2013) to handle stochastic optimization.

8.2.3 | Step 2: Counterfactual Predictive Model

As described in section 2.3, the counterfactual fairness in the predictive model learned at step 2 is ensured by comparing, for each training individual, counterfactual predictions $Y_{S\leftarrow s'}$ for all $s' \in \Omega_S$. For the discrete case (i.e., S is a Categorical variable), we keep this process for our experiments. However, for the continuous setting (i.e., S is for instance generated from a Gaussian), such an approach must be somehow adapted, due to the infinite set Ω_S. In that case, we can consider a sampling distribution $P'(S)$ to formulate the following loss, which can be optimized via Monte-Carlo

sampling and stochastic gradient descent (SGD):

$$\mathcal{L}_{GL}(w_h) = \frac{1}{m}\sum_{i}^{m} l(h_{w_h}(x_i), y_i) + \lambda \underset{\substack{u\sim P(u|x_i,s_i,y_i),\\ \tilde{x}\sim P(x|u_i,s_i),\\ s'\sim P'(S),x'\sim P(x|u,s')}}{\mathbb{E}} [(h_{w_h}(\tilde{x}) - h_{w_h}(x'))^2] \quad (8.11)$$

This formulation is equivalent to the one from Eq. 8.8, for continuous outcomes $\hat{Y}$ (thus considering a least squared cost as Δ) and for continuous attributes S (thus using the sampling distribution $P'(S)$ rather than considering every possible $s \in \Omega_S$).

Note that using a non-uniform sampling distribution $P'(S)$ would enforce the attention of the penalisation near the mass of the distribution. This prevents using the prior of S estimated from the training set, since this would tend to reproduce inequity between individuals: counterfactual predictions for rare S values would be be little taken into account during training. We therefore consider a uniform $P'(S)$ in our experiments for the continuous setting when using the $\mathcal{L}_{CF}(w_h)$ objective at step 2.

However, for the specific case of high-dimensional sensitive attributes S, using a uniform sampling distribution $P'(S)$ could reveal as particularly inefficient. The risk is that a high number of counterfactual samples fall in easy areas for the learning process, while some difficult areas - where an important work for fairness has to be performed - remain insufficiently visited.

To tackle this problem, we propose to allow the learning process to dynamically focus on the most useful areas of Ω_S for each individual. During learning, we consider an adversarial process, which is in charge of moving the sampling distribution $P'(S)$, so that the counterfactual loss is the highest. This allows the learning process to select useful counterfactuals for ensuring fairness. Who can do more can do less: dynamically focusing on hardest areas allows one to expect fairness everywhere. Again, we face a two-players adversarial game, which formulates as follows:

$$\arg\min_{w_h} \arg\max_{\phi} \mathcal{L}_{DynCF}(w_h, \phi) \quad (8.12)$$

with:

$$\mathcal{L}_{DynCF}(w_h, \phi) = \frac{1}{m}\sum_{i}^{m} l(h_{w_h}(x_i), y_i) + \lambda \underset{\substack{u\sim P(u|x_i,s_i,y_i),\\ \tilde{x}\sim P(x|u,s_i),\\ s'\sim P_\phi(s|u),x'\sim P(x|u,s')}}{\mathbb{E}} [(h_{w_h}(\tilde{x}) - h_{w_h}(x'))^2]$$

Compared to Eq. 8.11, this formulation considers an adversarial sampling distribution $P_\phi(S|U)$ rather than a uniform static distribution $P'(S)$. It takes the form of a

neural network that outputs the parameters of the sampling distribution for a given individual representation U. In our experiments we use a diagonal logit-Normal distribution $sigmoid(\mathcal{N}(\mu_\phi(u), \sigma^2_\phi(u)I))$, where $\mu_\phi(u)$ and $\sigma^2_\phi(u)$ stand for the mean and variance parameters provided by the network for the latent code u. Samples from this distribution are then projected on the support Ω_S via a linear mapping depending on the shape of the set. Passing U as input for the network allows the process to define different distributions for different codes: according to the individual profiles, the unfair areas are not always the same. This also limits the risk that the adversarial process gets stuck in sub-optima of the sensitive manifold. As done for adversarial learning in step 1, all parameters are learned conjointly, by alternating steps for the adversarial maximization and steps of global loss minimization. The re-parametrization trick (Kingma and Welling, 2013) is also used, for the adversarial optimization of $P_\phi(S|U)$.

8.2.4 | Experiments

We empirically evaluate the performance of our contribution on 6 real world data sets. For the discrete scenario and specifically in the binary case ($Y \in \{0,1\}, S \in \{0,1\}$), we use 3 different popular data sets: the Adult UCI income data set (Dua and Graff, 2017) with a gender sensitive attribute (male or female), the COMPAS data set (Angwin et al., 2016) with the race sensitive attribute (Caucasian or not-Caucasian) and the Bank dataset (Moro et al., 2014) with the age as sensitive attribute (age is between 30 and 60 years, or not). For the continuous setting (Y and S are continuous), we use the 3 following data sets: the US Census dataset (US Census Bureau, 2019) with gender rate as sensitive attribute encoded as the percentage of women in the census tract, the Motor dataset (The Institute of Actuaries of France, 2015) with the driver's age as sensitive attribute and the Crime dataset (Dua and Graff, 2017) with the ratio of an ethnic group per population as sensitive attribute.

Additionally to the 6 real-world datasets, we consider a synthetic scenario, that allows us to perform a further analysis of the relative performances of the approaches. The synthetic scenario subject is a pricing algorithm for a fictional car insurance policy, which follows the causal graph from figure 8.4. We simulate both a binary and a continuous dataset from this scenario. The main advantage of these synthetic scenarios is that it is possible to get "ground truth" counterfactuals for each code U, obtained using the true relationships of the generation model while varying S uniformly in Ω_S. This will allow us to evaluate the counterfactual fairness of the models without depending on a given inference process for the evaluation metric, by relying on prediction differences between these true counterfactuals and the original individual. The

objective of this scenario is to achieve a counterfactual fair predictor which estimates the average cost history of insurance customers. We suppose 5 unobserved variables (Aggressiveness, Inattention, Restlessness, Reckless and Overreaction) which corresponds to a 5 dimensional confounder U. The input X is composed of four explicit variables $X_1, ..., X_4$ which stand for vehicle age, speed average, horsepower and average kilometers per year respectively. We consider the policyholder's age as sensitive attribute S. The input X and the average cost variable Y are sampled from U and S as depicted in figure 8.4. We propose both a binary (reported in the Appendix in section D.2) and a continuous version of this scenario. For both of them, 5000 individuals are sampled. Details of distributions used for the continuous setting of this synthetic scenario are given below:

$$U \sim \mathcal{N}\left[\begin{pmatrix}0\\0.5\\1\\1.5\\2\end{pmatrix}, \begin{pmatrix}1&0&0&0&0\\0&4&0&0&0\\0&0&2&0&0\\0&0&0&3&0\\0&0&0&0&2\end{pmatrix}\right]$$

$$\begin{aligned}
X1 &\sim \mathcal{N}(7 + 0.1 * S + U_1 + U_2 + U_3, 1);\\
X2 &\sim \mathcal{N}(80 + S + U_2^2, 10);\\
X3 &\sim \mathcal{N}(200 + 5 * S + 5 * U_3, 20);\\
X4 &\sim \mathcal{N}(10^4 + 5 * S + U_4 + U_5, 1000)\\
X &\sim [X1, X2, X3, X4];\\
S &\sim \mathcal{N}[45, 5];\\
Y &\sim \mathcal{N}(2 * (7 * S + 20 * \sum_j U_j), 0.1)
\end{aligned}$$

8.2.4.1 | Step 1: Counterfactual Inference

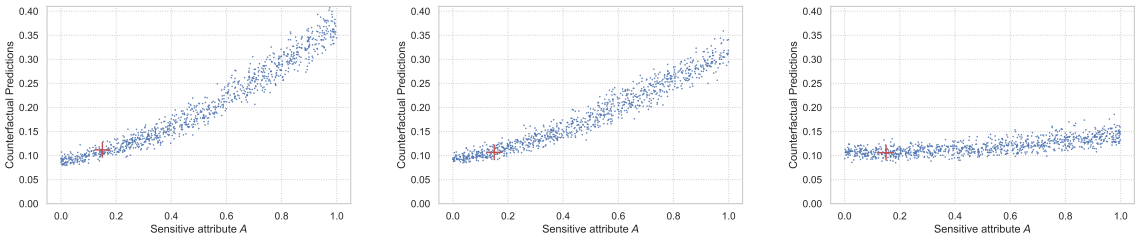
In this section, we report experiments performed for assessing our adversarial approach for Counterfactual Inference (step 1 of the previous section). We compare our adversarial approach with two version of the approach in Eq. 8.6, each using one of the two MMD constraints MMD wrt $P(S)$ or MMD wrt U_s as presented in section 8.2.1 (step 1). Note that these approaches are not applicable for continuous datasets as discussed at the end of section 8.2.1. For every approach, we compare three different inference schemes for U: $q_\phi(u|x, y, s)$, $q_\phi(u|x, y)$ and $q_\phi(u|x, s)$. As a baseline,

we also use a classical Variational Autoencoder inference without counterfactual independence constraint (i.e., Eq. 8.6 without the last term).

All hyper-parameters for every approach have been tuned by 5-fold cross-validation. For the US Census data set for our approach for instance, the encoder q_ϕ architecture is an MLP of 3 hidden layers with 128, 64 and 32 units respectively, with ReLU activations. On this dataset, the decoder p_θ is an MLP of only one hidden layer with 64 units with a ReLu activation function and the output consists in one single output node with linear activation to reconstruct Y and 37 units to reconstruct X (number of features). The adversarial neural network p_ψ is an MLP of two hidden layers with 32 and 16 units respectively. For the binary datasets, a sigmoid is applied on the outputs of decoders for S and Y. For both MMD constraints we used a Gaussian radial basis function kernel. For all datasets, the prior distribution $p(U)$ considered for training the models is a five-dimensional standard Gaussian.

In order to evaluate the level of dependence between the latent space U and the sensitive variable S, we compare the different approaches by using the neural estimation of the HGR correlation coefficient given in chapter 3. The estimator is trained for each dataset and each approach on the train set, comparing observed variables S with the corresponding inferred codes U.

For all data sets, we repeat five experiments by randomly sampling two subsets, 80% for the training set and 20% for the test set. Finally, we report the average reconstruction loss for X and Y on the test set, as long as the HGR between inferred test codes and the corresponding sensitive attributes. Results of our experiments can be found in table 8.3 for the discrete case and table 8.4 for the continuous case. For all of them, we attempted via the different hyperparameters (λ_x, λ_y, λ_{MMD}, λ_{ADV}) to obtain the lower dependence measure while keeping the minimum loss as possible to reconstruct X and Y.



(a) $\lambda = 0.00$; $CF = 72\%$ (b) $\lambda = 0.10$; $CF = 44\%$ (c) $\lambda = 0.30$; $CF = 04\%$

Figure 8.6: Impact of λ (Crime data set) on a specific instance i. Blue points are counterfactual predictions $h_{w_h}(x_{i,S\leftarrow s'})$ from 1.000 points $S \leftarrow s'$ generated randomly. The red cross represents the prediction $h_{w_h}(x_{i,S\leftarrow s})$ for the real $S = s$ of instance i.

Table 8.3: Inference results in the discrete case

		Adult UCI			Compas			Bank			Synthetic Scenario		
		Loss X	Loss Y	HGR	Loss X	Loss Y	HGR	Loss X	Loss Y	HGR	Loss X	Loss Y	HGR
$(\{x,y,s\})$	No Constraint, $q(u\|x,y,s)$	0.0781	0.0006	0.6984	0.0278	0.0041	0.6952	0.0963	0.0001	0.5988	0.2681	0.0085	0.9725
	Adv. Constraint, $q(u\|x,y,s)$	0.1091	0.0009	**0.5453**	0.0254	0.0020	**0.2693**	0.2038	0.0005	**0.3423**	0.2669	0.0721	**0.4167**
	MMD wrt $P(U)$, $q(u\|x,y,s)$	0.1286	0.0012	0.7017	0.0252	0.0029	0.6565	0.2002	0.0002	0.4521	0.2535	0.0839	0.6623
	MMD wrt U_s, $q(u\|x,y,s)$	0.0938	0.0009	0.7181	0.0259	0.0098	0.8892	0.1263	0.0003	0.5188	0.2762	0.0351	0.5697
$(\{x,y\})$	No Constraint, $q(u\|x,y)$	0.0786	0.0008	0.6077	0.0274	0.0133	0.3817	0.0957	0.0001	0.4989	0.2577	0.0022	0.6418
	Adv. Constraint, $q(u\|x,y)$	0.1272	0.0329	**0.1811**	0.0245	0.0013	**0.1728**	0.1858	0.0073	**0.2476**	0.2649	0.1015	0.4521
	MMD wrt $P(U)$, $q(u\|x,y)$	0.1287	0.0016	0.6092	0.0259	0.0055	0.4470	0.1898	0.0003	0.3716	0.2567	0.0885	0.6868
	MMD wrt U_s, $q(u\|x,y)$	0.0872	0.0013	0.6852	0.0266	0.0094	0.3109	0.1415	0.0003	0.3929	0.2674	0.0553	**0.4473**
$(\{x,s\})$	No Constraint, $q(u\|x,s)$	0.0982	0.3534	0.6689	0.0288	0.8246	0.3726	0.1391	0.2101	0.5572	0.2686	0.0128	0.7040
	Adv. Constraint, $q(u\|x,a)$	0.0995	0.3462	0.5259	0.0271	0.6889	0.4344	0.1880	0.2110	**0.3061**	0.2589	0.0980	**0.4264**
	MMD wrt $P(U)$, $q(u\|x,s)$	0.1308	0.3559	**0.3586**	0.0288	0.7611	0.4365	0.2141	0.2129	0.3386	0.2506	0.1176	0.6298
	MMD wrt U_s, $q(u\|x,s)$	0.0940	0.3603	0.5811	0.0278	0.7314	**0.3345**	0.1485	0.2135	0.5536	0.2584	0.1076	0.4692

Table 8.4: Inference results in the continuous case

	US Census			Motor			Crime			Synthetic Scenario		
	Loss X	Loss Y	HGR	Loss X	Loss Y	HGR	Loss X	Loss Y	HGR	Loss X	Loss Y	HGR
No Cons. $q(u\|x,y,s)$	0.1685	0.0019	0.5709	0.2526	0.0024	0.9023	0.4558	0.0016	0.9059	0.6788	0.0076	0.9523
No Cons. $q(u\|x,y)$	0.1690	0.0005	0.4163	0.3068	0.0034	0.9479	0.4523	0.0018	0.8998	0.6495	0.0003	0.6227
No Cons. $q(u\|x,s)$	0.1726	0.2886	0.8252	0.3377	0.9381	0.9728	0.4634	0.3999	0.9076	0.6751	0.4554	0.8650
Adv $q(u\|x,y,s)$	0.1617	0.0004	0.3079	0.4702	0.0035	0.2941	0.4865	0.0701	0.5268	0.6804	0.0088	0.2280
Adv $q(u\|x,y)$	0.1663	0.0009	0.2980	0.3694	0.0057	0.3314	0.4835	0.0571	0.6024	0.6633	0.1196	0.3175
Adv $q(u\|x,s)$	0.1828	0.2891	0.3285	0.4706	0.9878	0.2478	0.4904	0.3933	0.5810	0.6862	0.8819	0.5148

As expected, the baseline without the independence constraint achieves the best X and Y reconstruction loss, but this is also the most biased one with the worst dependence in term of HGR in most datasets. Comparing the different constraints in the discrete case, the adversarial achieves globally the best result with the lower HGR while maintaining a reasonable reconstruction for X and Y. It is unclear which MMD constraint performs better than the other. We observe that the best results in terms of independence are obtained without the sensitive variable given as input of the inference network (inference only with X and Y). Note however that for the MMD constraints, this setting implies to make the wrong assumption of independence of U w.r.t. S given X and Y for the estimation of the constraint (as discussed at the end of section 8.2.1). This is not the case for our adversarial approach, which obtains particularly good results on this setting for discrete datasets. On continuous datasets, our approach succeeds in maintaining reasonable reconstruction losses for important gains in term of HGR compared to the classical VAE approach (without constraint). Interestingly, on these datasets, it appears that our approach obtains slightly better results when using the full information (X, Y and S) as input of the inference network. We explain this by the fact that removing the influence of a binary input is harder than the one of a smoother continuous one, while this can reveal as a useful information for generating relevant codes.

8.2.4.2 | Step 2: Counterfactual Predictive Model

This section reports experiments involving the training procedure from step 2 as described in section 8.2.2. The goal of these experiments is threefold: 1. assess the impact of the adversarial inference on the target task of counterfactual fairness, 2. compare our two proposals for counterfactual bias mitigation (i.e., using a uniform distribution or an adversarial dynamic one for the sampling of counterfactual sensitive values) and 3. assess the impact of the control parameter from Eq.8.12.

The predictive model used in our experiments is a MLP with 3 hidden layers. The adversarial network P_ϕ from Eq.8.13 is a MLP with 2 hidden layers and RELU activation. For all our experiments, a single counterfactual for each individual is sampled at each iteration during the training of the models. Optimization is performed using ADAM.

Tables 8.5 and 8.6 report results for the discrete and the continuous case respectively. The inference column refers to the inference process that was used for sampling counterfactuals for learning the predictive model. For each setting, we use the best configuration from tables 8.3 and 8.4. The mitigation column refers to the type of counterfactual mitigation that is used for the results: No mitigation or L_{CF} (Eq.8.8) for the discrete case; No mitigation, L_{CF} (Eq.8.11) or L_{DynCF} (Eq.8.13) for the continuous setting. Results are reported in terms of accuracy (for the discrete case) or MSE (for the continuous case) and of Counterfactual Fairness (CF). The CF measure is defined, for the m_{test} individuals from the test set, as:

$$CF = \frac{1}{m_{test}} \sum_{i}^{m_{test}} \mathbb{E}_{(x',s')\sim C(i)}[\Delta(h_{w_h}(x_i, s_i), h_{w_h}(x', s'))] \tag{8.13}$$

where $C(i)$ is the set of counterfactual samples for the i-th individual of the test set. This corresponds to counterfactuals sampled with the Adversarial inference process defined at step 1 (with the best configuration reported in tables 8.3 and 8.4). As discussed above, the synthetic datasets allow one to rely on "true" counterfactuals for the computation of counterfactual fairness, rather than relying on an inference process which may include some bias. For these datasets, we thus also report an additional RealCF metric, which is defined as in Eq. 8.13, but using these counterfactuals sampled from the true codes used to generate the test data. For both CF and RealCF, for every i from the test set, $|C(i)|$ equals 1 for binary settings and $|C(i)|$ equals 1000 for the continuous one. Δ is a cost function between two predictions, the logit paring cost for the binary case (more details given in section 8.2.1 step 2) and a simple squared difference for the continuous setting.

Results from both tables first confirm the good behavior of our inference model from step 1, which allows one to obtain greatly better results than other inference processes for both the discrete and the continuous settings. Our adversarial counterfactual inference framework allows one to get codes that can be easily used to generate relevant counterfactual individuals. For this observation, the most important results are those given for the synthetic scenarios, for which the RealCF metric shows good results for our method, while strongly reliable since relying on counterfactuals sampled from true codes of individuals.

Secondly, results from table 8.4 show that, even in the continuous setting where the enumeration of all values from Ω_S is not possible, it is possible to define counterfactual mitigation methods such as our approaches L_{CF} and L_{DynCF}. These two methods, used in conjunction with our Adversarial Inference, give significantly better results than no mitigation on every dataset. Interestingly, we also observe that L_{DynCF} allows one to improve results over L_{CF}, which shows the relevance of the proposed dynamic sampling process.Furthermore, note that we can reasonably expect even better results compared to L_{CF} on data with higher-dimensional sensitive attributes.

To illustrate the impact of the hyperparameter λ on the predictions accuracy (MSE Error) and the counterfactual fairness estimation (CF), we plot results for 10 different values of λ (5 runs each) on figure 8.7 for the Crime data set. It clearly confirms that higher values of λ produce fairer predictions, while a value near 0 allows one to only focus on optimizing the predictor loss. This is also observable from Fig. 8.6 which plots counterfactual predictions for a specific instance i from the test set. Higher values of λ produce clearly more stable counterfactual predictions.

Table 8.5: Counterfactual Fairness Results for the Discrete Case

Inference	Mitigation	Adult UCI		Compas		Bank		Synthetic Scenario		
		Accuracy	CF	Accuracy	CF	Accuracy	CF	Accuracy	CF	Real CF
Without Constraint	None	84.22%	0.0096	67.12%	0.0102	90.64%	0.0369	99.49%	0.1087	0.1810
	L_{CF}	83.28%	0.0008	66.20%	0.0051	90.46%	0.0024	95.89%	0.0757	0.1327
MMD	None	84.22%	0.0116	67.12%	0.0076	90.64%	0.0469	99.49%	0.1074	0.1775
	L_{CF}	83.84%	0.0024	65.91%	0.0041	90.64%	0.0043	99.29%	0.0893	0.1557
Adversarial	None	84.22%	0.0114	67.12%	0.0118	90.64%	0.0376	99.49%	0.1426	0.1838
	L_{CF}	83.74%	0.0002	66.73%	0.0001	90.60%	0.000	93.19%	0.0001	0.0014

In figure 8.8, we consider the distribution of considered counterfactual samples w.r.t. to the sensitive variable S for the uniform sampling strategy from $P'(S)$ and the dynamic strategy as defined in Eq.8.12. This is done on the Motor dataset and for a specific randomly sampled instance i with sensitive attribute $a_i = 75$, at a given point of the optimization, far before convergence (the model is clearly

Table 8.6: Counterfactual Fairness Results for the Continuous Case

Inference	Mitigation	US Census		Motor		Crime		Synthetic Scenario		
		Accuracy	CF	MSE	CF	MSE	CF	MSE	CF	Real CF
Adversarial	None	0.274	0.0615	0.938	0.0285	0.412	0.7412	0.454	0.2490	1.1248
	L_{CF}	0.289	0.0009	0.941	0.0009	0.452	0.0154	0.572	0.0014	0.2013
	L_{DynCF}	0.290	0.0008	0.940	0.0005	0.445	0.0076	0.568	0.0013	0.2000
Without Constraint	None	0.274	0.0433	0.938	0.0271	0.381	0.7219	0.454	0.2919	1.1338
	L_{CF}	0.307	0.0010	0.939	0.0021	0.407	0.2938	0.531	0.1968	0.3303
	L_{DynCF}	0.310	0.0008	0.942	0.0016	0.418	0.2881	0.546	0.1743	0.3188

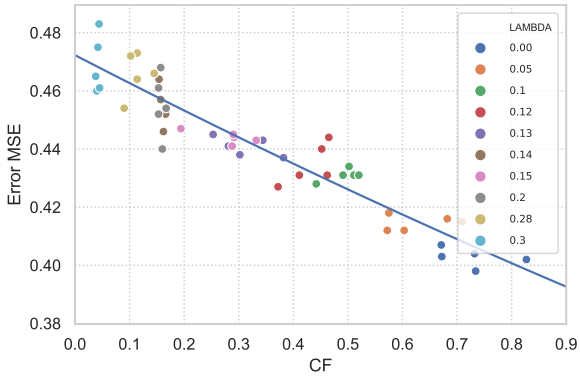


Figure 8.7: Impact of hyperparameter λ (Crime data set)

unfair at this point). The blue points are the counterfactual fairness estimation $(h_\theta(X_{i,S\leftarrow s}, s) - h_\theta(X_{i,S\leftarrow s'}, s'))$ for each counterfactual sampled s'b (1.000 points) from the uniform distribution $P'(S)$. The red points are the counterfactual fairness estimations for counterfactuals corresponding to s' values (30 points) sampled from our dynamic distribution $P_\phi(s'|u) = \mathcal{N}(\mu_\phi(u), \sigma^2_\phi(u)I)$, where ϕ are the parameters of the adversarial network which optimizes the best mean and variance for each latent code u ($\mu_\phi(u)$ and $\sigma^2_\phi(u)$). Being optimized to maximize the error at each gradient step, the red points are sampled on lower values of S where the error is the most important. More importantly, very few points are sampled in the easy area, near the true sensitive value of i which is 75. This demonstrates the good behavior of our dynamic sampling process.

8.2.5 | Total and Counterfactual Effect

In addition, we propose to compare performances of our approach with works based on fair data generation (Louizos et al., 2017; Kim et al., 2021; Kocaoglu et al., 2018; Xu et al., 2019), to emphasize the benefits of our two-steps process for learning counterfactually fair prediction models.

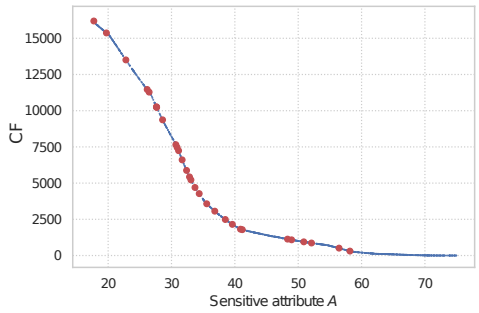


Figure 8.8: Dynamic Sampling Visualization for a randomly sampled individual whose age S is 75. Red points are sampled counterfactuals from the dynamic distribution $P_\phi(s'|u)$ with u the inferred confounding for this individual.

Traditionally, these methods are evaluated in terms of total and counterfactual causal effect of the sensitive on the data generated by the models. Total causal effect (TCE) aims at assessing the statistical parity on the outcomes generated from causal intervention. TCE for binary sensitives is defined as:

$$TCE = P(Y_{S\leftarrow s_1}) - P(Y_{S\leftarrow s_0}) \tag{8.14}$$

where $Y_{S\leftarrow s}$ corresponds to generated causal transformation of input Y, resulting from setting s as the sensitive attribute to the corresponding individual, according to the causal graph G (i.e., obtained via distribution $P(Y_{S\leftarrow s}|X, Y, S)$).

A limit of such a metric is that it only considers fairness in the data given as training set for learning the predictor model. We claim that this is not enough since any residual bias in these data may strongly impact the final prediction (on both training and testing data). To overcome this limitation, and assess total effect of sensitives on predictions rather than on training data only, we introduce the Total Predictions Effect (TPE), which refers to the statistical parity of the output prediction from intervention. The metric is defined in the binary case as:

$$TPE = P(h_\theta(X_{S\leftarrow s_1})) - P(h_\theta(X_{S\leftarrow s_0})) \tag{8.15}$$

which takes into account the fairness of the predictor from transformed data $X_{S\leftarrow s}$.

Following (Kim et al., 2021; Xu et al., 2019), we also consider counterfactual effects, which depend on the effect of the sensitive on the outcome for specific individuals (or groups of individuals). Similarly as for the total effect, for any observation o, we consider the Counterfactual Causal Effect defined as: $CCE = P(Y_{S\leftarrow s_1}|o) - P(Y_{S\leftarrow s_0}|o)$ and introduce the Counterfactual Prediction Effect as: $CPE = P(h_\theta(X_{S\leftarrow s_1})|o) - P(h_\theta(X_{S\leftarrow s_0})|o)$.

Causal Effect In Table 8.7, we represent the results from the different generated data observations on the Adult UCI dataset. We consider the condition observations o as the concatenation of the features *race* and *native_country* as in (Xu et al., 2019; Kim et al., 2021) ($O = \{race, native\_country\}$). We report the chi-square distance χ^2 that indicates the similarity between the generated and the real dataset. We consider three baselines that are unaware of the fairness constraint: CausalGan (Kocaoglu et al., 2018) that preserves the causal structures, the DCEVAE WR that represents the DCVAE architecture (Kim et al., 2021) without any fairness regulation term (i.e., $\beta_f = 0$ according to notations in (Kim et al., 2021)) and the original data. Our approach that contains no fairness penalty on the generated outcomes (in step 1) is also designed to reflect the causal structure. We also analyze the impact of CFGAN CE (Xu et al., 2019), which aims at decreasing the TCE in the generated data, CFGAN CE (Xu et al., 2019) which in turn aims at decreasing the 4 different (CCE), and finally DCEVAE, which corresponds to the DCVAE model with a fairness penalization set to $\beta_f = 0.3$.

As expected, only the three last methods, which act on the data rather than on the predictor itself, are able to mitigate TCE and CCE. Our method does not seek at mitigating biases in inferred outcomes, but seeks at leveraging inferred variables that allow it to learn a fair predictor. This is thus without any surprise that the reconstructed Y are not unbiased with regards to the sensitive; this is even a good indication of no information loss in the step 1 of our process, despite mitigating correlation between the latent confounder U and the sensitive A. In the following, we compare these observations to results in prediction effects.

Table 8.7: Total Causal Effect and Counterfactual Causal Effect on Adult UCI

	Total Causal Effect (TCE)	Counterfactual Causal Effect (CCE)				X^2
		o_{00}	o_{01}	o_{10}	o_{11}	
Real Data	0.1936	0.1785	0.1266	0.1293	0.2023	0
Causal GAN	0.1729	0.0717	0.1201	0.1326	0.1856	20388
DCEVAE WR	0.1819	0.1694	0.1472	0.1522	0.1899	20822
OURS	0.1834	0.1783	0.1803	0.1778	0.1845	21641
CFGAN CE	0.0135	0.0586	0.0087	0.003	0.0148	20591
CFGAN TE	0.0171	0.007	0.0168	0.0201	0.0169	20541
DCEVAE	0.0050	0.0051	0.0040	0.0043	0.0051	21142

Predictions Effect In this part, we focus on the level of fairness of the final predictor model. A Logistic Regression (LR), a Neural Network (NN) and a classification tree (CART) are considered in the following. These predictors are either trained on the datasets produced from generation-focused models (i.e., CausalGAN, CFGAN TE,

CFGAN CE, DCVAE RW and DCEVAE), or trained in the second step as described in subsection 8.2.3 for our two-steps model. Please note that our algorithm can only handle derivative gradient during the optimization, therefore we have discarded the tree CART.

We report the results for each prediction model in Table 8.8, in terms of TPE, CPE (measured on generated data for test samples) and prediction accuracy (measured on the original test dataset). From this table, we observe completely different results than those from previous table, with generation based models such as CFGAN greatly penalized compared to two-steps methods such as ours. This confirms our intuition that, even if produced data have biases well mitigated on the test set (as seen in table 8.7), some small residuals of these biases can stay in the data. Then, the learning process is free to assign important emphasis on these problematic features, if this helps to achieve good prediction accuracy. In two-steps approaches such as ours, this is not the case, since biases of the outcomes are mitigated while learning prediction models, which enables more fairness robustness on test data.

Table 8.8: Total Predictions Effect and Counterfactual Predictions Effect on Adult UCI

	Total Predictions Effect (TPE)	Counterfactual Predictions Effect (CPE)				Accuracy
		o_{00}	o_{01}	o_{10}	o_{11}	
Causal GAN - NN	0.1834	0.1148	0.134	0.1353	0.1965	0.8138
Causal GAN - LR	0.1368	0.0634	0.0985	0.0576	0.1535	0.7997
Causal GAN - CART	0.2204	0.0163	0.112	0.1252	0.2482	0.8082
DCEVAE RW - NN	0.1782	0.1758	0.1768	0.1771	0.1786	0.8133
DCEVAE RW - LR	0.1867	0.1237	0.1866	0.16474	0.1912	0.8040
DCEVAE RW - CART	0.2161	0.0662	0.1726	0.22638	0.22742	0.8119
CFGAN CE - NN	0.1394	0.1312	0.1339	0.0968	0.1463	0.8085
CFGAN CE - LR	0.1486	0.0603	0.1161	0.0597	0.1662	0.8153
CFGAN CE - CART	0.1501	0.101	0.0993	0.1119	0.1612	0.8143
CFGAN TE - NN	0.1415	0.0637	0.1266	0.1059	0.1498	0.8129
CFGAN TE - LR	0.1793	0.0295	0.1603	0.0528	0.2029	0.8116
CFGAN TE - CART	0.1794	0.0244	0.1463	0.0802	0.2004	0.8096
DCEVAE - NN	0.0047	0.027	0.0205	0.0205	0.0021	0.7997
DCEVAE - LR	0.0172	0.0525	0.0169	0.0169	0.0157	0.8019
DCEVAE - CART	0.0297	0.0265	0.0243	0.0243	0.0255	0.7999
Ours - NN	**0.0007**	**0.0044**	**0.0009**	**0.0017**	**0.0014**	**0.8441**
Ours - LR	0.0139	0.0179	0.0175	0.0166	0.013	0.8279

8.3 | Conclusion

We have seen in this chapter two different look of individual fairness. First, we have presented a new method based on variational inference for enforcing Fairness Through Awareness without accessing to a distance metric. In a first step, entitled Rényi Variational Inference, we infer an unbiased confounder by combining ELBO

optimization with HGR minimization. In a second step, we mitigate individual bias by adding a regularization term, representing output discrepancies between similar individuals, to global objective of a predictive neural network. This method proved to be very efficient on 3 real-world data sets for several individual fairness metrics that we proposed.

In addition, we have developed a new adversarial learning approach for counterfactual fairness. To the best of our knowledge, this is the first such method that can be applied for continuous sensitive attributes. The method proved to be very efficient for different dependence metrics on various artificial and real-world data sets, for both the discrete and the continuous settings. Finally, our proposal is applicable for any causal graph to achieve generic counterfactual fairness. As future works, it might be interesting to consider a generalization of our proposal for Path Specific (Chiappa, 2019) counterfactual fairness in the continuous case.

9

Conclusion and Perspectives

9.1 | Summary of the Contributions

Despite the fairness field being vast, we identified weaknesses in current traditional approaches that may produce an undesirable results. For example, we identify a lack of work for predictor models adapted explicitly to the tabular datasets or for continuous settings. We also raised the point that current fair algorithms are often restricted to simple contexts where sensitive variables are supposed to be present. Also, considering generic fair algorithm can be counterproductive for specific applications, as we have observed in insurance pricing.

Finally, the different contributions made in this thesis can be organized into two themes: those that focus on identifying and completing unstudied or partially studied sub-contexts in the Fair-ML community and those that provide rigorous methodologies to avoid counterproductive practices in a real context.

We use this point of view to summarize our contributions below. Additionally, we present how these contributions can be used to draw conclusions in a more general context of the field of fairness.

9.1.1 | Completing the Gaps

Fair Algorithms adapted to Tabular Datasets Most current fairness approaches focus more on tabular data than on text, images, or video. Collecting personal and individual information leads to direct or indirect discrimination in output prediction. Surprisingly we have noticed a lack of work for fair classifiers based on decision trees even though they have proven very efficient for tabular dataset. For this reason, we developed a new approach to produce fair gradient boosting algorithms. Our gra-

dient boosting framework has allowed us to consider any regression machine, by iteratively feeding it with both prediction and fairness residuals as target outputs. This enables the use of very effective machines such as CART decision trees for fair machine learning. To the best of our knowledge, this is the first adversarial learning method for generic classifiers, including decision trees. Compared with other state-of-the-art algorithms, our Fair Gradient Tree Boosting approach proves to be more efficient in terms of accuracy while obtaining a similar level of fairness. Since our publication, we note that efforts have been made on this point, for example, by combining trees with mixed-integer optimization (MIO) (Aghaei et al., 2021) or by enforcing group fairness on XGBoost model (Ravichandran et al., 2020).

Unwanted biases in continuous case In addition, we have identified a main issue for applying fairness for any continuous sensitive features: The traditional state-of-the-art adversarial algorithms are theoretically not able to optimize the most classical fairness objective as *demographic parity*. To address this issue we have contributed to a new estimation of the Hirschfeld-Gebelein-Rényi (HGR) maximal correlation. This estimation has allowed us to measure and mitigate non linear dependencies between features. Our estimation, the HGR_NN, is an approximation by deep neural network which has shown to be very useful for debiasing a predictor model with adversarial learning. We have theoretically shown the interest of using this fairness metric as a penalization term compared to the simple adversarial methods for the continuous case. It therefore allows us to use it for mitigating the underlying bias on the output prediction (Grari et al., 2020b), on an intermediary representation space (Grari et al., 2021b) and finally for a general framework in individual fairness (Grari et al., 2021a).

No access to the sensitive attribute Not having access to the sensitive attribute is a classic context in practice and is a challenge to overcome to ensure fairness. The need of new approaches for algorithmic fairness that break away from the prevailing assumption of observing sensitive characteristics has been many times highlighted, as in (Tomasev et al., 2021). By leveraging recent developments for approximate inference via variational auto-encoding, we have inferred a sensitive information proxy with a new framework named SRCVAE. The bias mitigation is done in a second step in an adversarial fairness approach. Our proposed method has empirically achieved significant improvement over existing works in the field. We have observed that the generated proxy's latent space correctly recovers sensitive information and that our approach achieves a higher accuracy while obtaining the same level of fairness on two real datasets.

9.1.2 | Understanding and Avoiding Counterproductive Practices

Beware of the generalisation of traditional fair models Furthermore, we claim that mitigating undesired biases with generic fair algorithms can be counterproductive for specific applications. We have shown that mitigating unwanted biases in insurance pricing via adversarial algorithms is not adaptable. Some components essential to the predictive actuarial pricing have been unfairly neutralized on the predictive task. Acting unwanted biases via adversarial learning on autoencoder seems a promising choice for specific cases. An autoencoder structure has allowed us to generate multiple aggregated pricing factors and, at the same time, debiasing them with adversarial learning. The results show the relevance of the method compared to the traditional one. We argue that this type of framework can be generalized to many other applications that require the creation of specific components, such as telematics, human behaviors, or credit scoring.

Beware of the Importance of the Architecture First, among this zoo of attenuation algorithms, we wanted to understand why some approaches such as fair representation perform significantly better than predictive reprocessing. To this end, we proposed the first work to compare attenuation at different levels of neural architectures. We argue that operating at intermediate levels of neural representations offers the best trade-off between expressiveness and generalization for bias mitigation.

Beware of what is being assessed We report that a large majority of counterfactual approaches do not evaluate the level of fairness on the final output prediction. Instead, they ensure and evaluate only an intermediary step called causal graph generation. The problem is that even if the counterfactual-generated observations are fair, it is not guaranteed that the predictor model is fair. We have shown an interest in penalizing the predictor model rather than the generated data. We have also contributed to a new framework that can suit either continuous or binary settings.

9.2 | Overview of Future Works and Perspectives

The contributions of this thesis open several promising directions for further works. Beyond some perspectives announced in the different chapters, these include prospective works on the proposed fairness approaches and criteria.

Six main research directions are identified, developed in the following sections. First, we discuss the questions opened by the transparence induces by the fair algorithm. Then, we propose prospective studies to extend the work conducted in thesis to the privacy domain. Next, we discuss the difficulty on the real application of fairness on a deployment phase. Finally, we identify perspectives opening the discussion on the consequences of the contributions of this thesis for the discovery and malicious unwanted biases.

More Transparent Algorithms

Currently, most fairness algorithms focus exclusively on the fairness criteria. However, as mentioned at different levels of this thesis, many tensions coexist when increasing fairness, such as decreasing model performance or privacy. Moreover we note that additionally there is an issue of transparency. Increasing fairness comes at the expense of something else, and not knowing the precise induced change between a biased and an unbiased model can be problematic. For example, adding an adversarial structure may lead to an increase in model complexity and reduce its confidence in domains like health care, finance, and security, which can be harmful. We argue that understanding the induced changes is one of the future challenges in the Fair-ML community. We identify some prospective studies for this purpose. A recent work (Wang et al., 2020) proposes a hierarchical rule-based model for classification tasks, Concept Rule Sets (CRS), with a strong transparent inner structure. To develop a model that achieves three objectives: a high classification performance, low complexity, and fair predictions. It would be interesting to implement this contribution with adversarial neural network architecture. By taking up the general idea of our different frameworks, the negative gradient from the adversarial (e.g., adversarial simple or our HGR) could be added to the predictor gradient of the discrete CRS via continuous Multilayer Logical Perceptron (MLLP) and Random Binarization (RB). Finally, it might be interesting to investigate a measure that does not only consider the general case of bias but can also spot and quantify bias that persists in specific sub-segments of the population.

More Private Algorithms

Moreover, in an ethical context, it also seems essential to address the anonymization of datasets, especially when we know, for example, that *"99.98% of Americans would be correctly re-identified in any dataset using 15 demographic attributes"* (Rocher et al.,

2019). A typically formalized notion used in the privacy literature is the *differential privacy* (Cynthia, 2006). It guarantees that a predictor model is trained on aggregate individuals and does not encode individuals' information. As raised by (Cummings et al., 2019; Alves et al., 2021; Pujol et al., 2020), there is tension between fairness and differential privacy. For example, (Alves et al., 2021) proves that it is impossible to simultaneously satisfy exact differential privacy and fairness (on equal opportunity) while maintaining a non-trivial accuracy of the predictor model. For future work, we are interested in working on a relaxed version of differential parity that would also include the fairness objective while keeping a correct level of the model's predictive performance. The privacy link with our HGR work is very close, especially with the Renyi-Divergence. For example, recent work uses Renyi-Divergence (Mironov, 2017) to relax this strict definition, so it seems interesting to study these two objectives at the same time.

Difficulty on Deployment

We believe that, although research on fair machine learning is progressing very rapidly, there is still a long way to go before its application becomes a reality. We see many obstacles that prevent achieving this aim.

First of all, as discussed in this thesis, fairness can be perceived differently by everyone. For this reason, currently there are 21 criteria that measure fairness. Choosing one of them may negatively impact others, and finding a generic measure that satisfies everyone is impossible.

Second, we expect the auditor and practitioner to possess the sensitive variable to certify fairness or even mitigate biases as adequately as possible. However, we are skeptical that this can be achieved in many areas. For example, requesting the origin or religion of policyholders for financial institutions or insurers in France is unacceptable, even in a good faith context. Some papers have recently focused on not giving the true sensitive variable to the practitioner while allowing bias mitigation (Kilbertus et al., 2018; Veale and Binns, 2017). Here, the sensitive variable is encrypted so that the practitioner can build a fair prediction model. The regulator that is controlling the model is the only one who has access to the sensitive variable in the deployment phase. Although this solution seems to be interesting for privacy reasons, it is not enough, since individuals still have to give their sensitive information to the regulator. Due to the increasing number of cyberattacks and possible mistrust in the regulator, we are dubious that it will work. For example, we are skeptical that policyholders would want to give information on their religion or race just to get a

car insurance, especially in EU. Applying fairness from a sensitive proxy such as ours (Grari et al., 2021c) in real-world applications does not seem to be possible yet as they are not mature enough. There is no guarantee of the resulting fairness level. A possible solution would be to give an option to the individuals wishing to benefit from fairness. But those who have not given the sensitive variable could be unjustifiably penalized.

In addition, some recent papers such as (Mishler and Dalmasso, 2022) have shown that there can be a significant instability between the level of fairness observed in the training and deployment phases. Naturally, one might expect this to be related exclusively to a change in distribution between the training and the test datasets. However, this is not the only reason. In particular, they show that the algorithms to which fairness has been applied are very unstable, even when the marginal distribution of features remains the same. The reason for that could be a conceptual shift where only the conditional distribution is different.

Furthermore, it is still challenging to understand what happens when a fairness metric is satisfied. For example, although the adversarial algorithms we have used in this work are very efficient, we know very little about the impacts on prediction. Therefore, in the future work, we would like to study the changes induced on local boundaries. In particular, by looking at the direction of the gradients of the adversarial network. This aim to detect segments that are most modified by the increase in fairness.

Finally, improving a fairness criterion requires using a testing framework that can be applied in an audit proposal. As mentioned in the paper (Wachter et al., 2021), the main reason for which fairness automation does not work in Europe is that, by design, the law does not provide *"a static or homogeneous framework suitable for discrimination testing in AI system"*. We suspect that this is related to the fact that each AI application requires specific fairness criteria which complicates the task. Choosing the right criteria for a specific situation has always been a central issue in human history. Therefore, we believe that it must be tailored to each task. For example, in medicine, a false negatives has a dramatic impact on patient's life. In this case choosing equalizing opportunity is more suitable. On the other hand, choosing demographic parity is more suitable for income assignments to obtain the same average for each demographic group. A future direction could be that judges, regulators, researchers and private sectors will increasingly need to work together and develop standards and certifying procedures to ensure that algorithmic disparity does not remain uncontrolled.

Discovering and/or Increasing Sensitive Biases : a Dangerous Shift of Paradigm

We caution that the type of approaches presented in this thesis can lead to opposite objectives or harmful practices. For example, practitioners intending to increase specific effects of variables on the prediction may achieve the opposite of the fairness objective. Note that increasing a feature effect on the prediction can be achieved by reversing the positive sign to the negative one of the second term corresponding to the HGR penalty (equation 4.6). In that case, as previously, the algorithm captures the estimated HGR dependence between the prediction and the sensitive attribute for each gradient iteration. However, instead of decreasing, this dependency will increase during training. The hyperparameter λ still allows controlling the trade-off level between the dependence loss and the accuracy. For example, a λ that tends to infinity will only consider the effect of the variable on the prediction. It should be noted that this objective goes beyond the Fair-ML field since some real applications are subject to the interest of increasing the effect of certain variables. For example, in non-life insurance and particularly in price elasticity, actuaries often recourse to increasing variables' effects for business purposes. They maximize the individual price effect of the policyholder, regardless of the set of other variables (Krolikowski, 2021).

It should also be noted that some practitioners may attempt to approximate a sensitive variable by inference, such as our SCRCVAE architecture seen in chapter 7 for not necessarily ethical reasons. Note that the regulations or laws are not clear on this subject, and for the moment, in many real-world applications, nothing prohibits a practitioner from increasing the effect of a proxy or even the sensitive variable itself on the prediction.

Excessive Focus on Popular Real-World Datasets.

Some real-world datasets, such as the Adult UCI (Dua and Graff, 2017) or Compass (Angwin et al., 2016) datasets, have gained a dramatic rise in popularity for testing the performance of algorithms in the Fair-ML community. However, we would like to point out that giving them our full attention can be problematic. We tend to overestimate them, and we are skeptical about the generalization of some algorithms. We need to be careful to not over-adjust the accuracy/fairness trade-off only for these data. These datasets follow particular distribution laws and causality graphs, so it seems problematic to consider them as a generality. Although we have used lesser-known data such as Pricingame and Default in this thesis, we believe that

other publicly available datasets would benefit from generalization in the future. We also note that the commonly used datasets in the Fair-ML community are easy to mitigate. They do not exhibit atypical scenarios. A less discussed topic in this thesis that would be interesting to investigate is extreme dependency cases or atypical graphs that would complexify mitigation. For example, if the dependence between the sensitive and the target is high (e.g., HGR$\geq$ 0.9), one could ask whether the trade-off between accuracy and fairness is still feasible. One of the two would probably take over the other. A first direction could be to conduct an extensive empirical comparative study of fair algorithms on many datasets, as has often been done on pure performance comparisons (e.g., a benchmark of classifiers on 71 datasets (Zhang et al., 2017)). A second direction could be to study atypical synthetic scenarios to understand how the different state-of-the-art algorithms behave.

Mitigating Biases Algorithms Beyond Fairness

Ensuring independence between variables has much broader applications than fairness alone. Many areas could benefit from this type of practice. For example, one widespread case is the unwanted noise interfering with machine learning prediction. This can be encountered, for example, in high-resolution microscopic, nuclear magnetic, medical, astronomical and satellite images or any type of sounds. We note that a sub-field of mitigation algorithms, named *Pileup models*, aims to reduce the noise of stacked particles and improve the performance of key physics observables. The link is often very close to fair machine learning algorithms. We note, for example, that the very first adversarial mitigation algorithm came from this field and was created in 2017 "Learning to Pivot with Adversarial Networks" (Louppe et al., 2017). As stated in many papers (Tamba et al., 2022; Wagner, 2021), the pileup separation is often a complicated nonlinear relationship. It would be interesting to investigate the interest of HGR in this field. Another application could be in Fluorescence microscopy, that is often noisy due to illuminating light (i.e., shot noise) (Laine et al., 2021). It might be interesting to explore how to extract and mitigate this biased source of information via classical fair adversarial learning models. Finally, as microeconomics and human behavior often require predictions from *all else being equal*, the use of an HGR adversarial algorithm from multidimensional individual information seems to be a promising choice.

Finally, retrieving information not present in the training set is a vast area. Our SCRCVAE algorithm can search for specific external effects in many other applications. For example, a company seeking to know the health status of a potential cus-

tomer could deduce this information from a knowledgeable causal graph. In this thesis, we also conduct some work on telematics insurance purposes for retrieving external information on policyholders' behavior (Corradin et al., 2022). We have used an EM algorithm. In the future, we are interested to see the potential added value of our VAE-architecture (Grari et al., 2021c).

Appendix

A

Supplementary Material of Chapter 3

This is the supplementary material of the chapter 3, Measuring Fairness. It provides proofs of our theoretical claims about the HGR neural estimator.

A.1 | Consistency of the HGR NN Estimator

The domains $\mathcal{U}$ and $\mathcal{V}$ of the random variables U and V are assumed to be compact.

We define the theoretical HGR as follows:

$$HGR(U,V) = \sup_{f:\mathcal{U}\to\mathbb{R}, g:\mathcal{V}\to\mathbb{R}} \rho(f(U), g(V)) \tag{A.1}$$

where ρ is the Pearson's correlation coefficient and f, g are (measurable) functions with finite and positive variance w.r.t the distributions of U and V.

We define the theoretical neural HGR measure associated to a family of neural networks F_Θ:

$$HGR_{F_\Theta}(U,V) = \sup_{(f_{\theta_f}, g_{\theta_g})\in F_\Theta} \rho(f_{\theta_f}(U), g_{\theta_g}(V)) \tag{A.2}$$

Θ is a compact domain of $\mathbb{R}^k$ for a given k.
$F_\Theta \subset \{(f_{\theta_f}, g_{\theta_g}), f_{\theta_f}$ and g_{θ_f} neural networks with parameters $(\theta_f,\theta_g) \in \Theta\}$.
We use the abuse of notation $HGR_\Theta(U,V)$ to refer to $HGR_{F_\Theta}(U,V)$. $HGR_\Theta(U,V)$ is well-defined when $f_{\theta_f}(U)$ and $g_{\theta_g}(V)$ are not constant for all $(\theta_f,\theta_g) \in \Theta$.

We define the empirical HGR neural measure, given n *i.i.d* samples of (U,V) and a family F_Θ, as:

$$\widehat{HGR(U,V)}_n = \sup_{(\theta_f,\theta_g)\in\Theta} \rho_n(f_{\theta_f}(U), g_{\theta_g}(V)) \tag{A.3}$$

where ρ_n is the sample correlation computed using the samples of (U, V). ρ_n is well-defined *iff* the sample variances are positive.

Lemma A.1.1. *(approximation) Let $\eta > 0$. There exists a family of continuous neural networks F_Θ parametrized by a compact domain $\Theta \subset R^k$, such that $HGR_\Theta(U, V)$ is well-defined and:*

$$|HGR(U, V) - HGR_\Theta(U, V)| \leq \eta. \tag{A.4}$$

Proof. Let $\eta > 0$ and $\epsilon > 0$.

There exist functions f^*, g^* centered and standardized such that:

$$HGR(U, V) - \rho(f^*(U), g^*(V)) < \epsilon$$

Let f and g some functions with positive and finite variance and $\tilde{f} = \frac{(f-\mu_f)}{\sigma_f}$; $\tilde{g} = \frac{(g-\mu_g)}{\sigma_g}$, so that $\tilde{f}(U)$ and $\tilde{g}(V)$ are centered and standardized.

$$\begin{aligned} &HGR(U, V) - \rho(f(U), g(V)) \\ &\leq \epsilon + \rho(f^*(U), g^*(V)) - \rho(\tilde{f}(U), \tilde{g}(V)) \\ &= \epsilon + E(f^*(U)g^*(V)) - E(\tilde{f}(U)\tilde{g}(V)) \end{aligned} \tag{A.5a}$$

Using the Cauchy-Schwarz inequality:

$$\begin{aligned} &E(f^*(U)g^*(V)) - E(\tilde{f}(U)\tilde{g}(V)) \\ &= E\left((f^*(U) - \tilde{f}(U))g^*(V)\right) + E\left((g^*(V) - \tilde{g}(V))\tilde{f}(U)\right) \\ &\leq \sqrt{E\left((f^*(U) - \tilde{f}(U))^2\right)} + \sqrt{E\Big((g^*(V) - \tilde{g}(V))^2\Big)} \end{aligned} \tag{A.6a}$$

Let $||h||_2 = E(h(X)^2)^{1/2}$ with $X \sim U$ or $X \sim V$ depending on the context. The inequality becomes:

$$HGR(U, V) - \rho(f(U), g(V)) \leq \epsilon + ||f^* - \tilde{f}||_2 + ||g^* - \tilde{g}||_2 \tag{A.7a}$$

Let's find a bound of $||f^* - \tilde{f}||_2$ that depends on $||f^* - f||_2$:

$$||f^* - \tilde{f}||_2^2 = 2 - 2E\left(f^*(U)\left(\frac{f(U) - \mu_f}{\sigma_f}\right)\right) \tag{A.8a}$$

$$= 2 - 2\frac{E\left(f^*(U)f(U)\right)}{\sigma_f} \tag{A.8b}$$

$$= 2 + \frac{1}{\sigma_f}E\left((f^*(U) - f(U))^2 - 1 - \sigma_f^2 - \mu_f^2\right) \tag{A.8c}$$

$$\leq 2 + \frac{1}{\sigma_f}(||f^* - f||_2^2 - 1 - \sigma_f^2) \tag{A.8d}$$

$$= \frac{||f^* - f||_2^2}{\sigma_f} + 2 - (\frac{1}{\sigma_f} + \sigma_f) \tag{A.8e}$$

We bound the standard deviation error, using Cauchy-Schwarz inequality in (9c) and triangular inequality in (9d):

$$|1 - \sigma_f| \leq \sqrt{|1 - E(f(U)^2)|} + |E(f(U))| \tag{A.9a}$$

$$= \sqrt{\left|E\Big((f^*(U) - f(U))(f^*(U) + f(U))\Big)\right|}$$

$$+ \left|E\big(f(U) - f^*(U)\big)\right| \tag{A.9b}$$

$$\leq \sqrt{||f^* - f||_2||f^* + f||_2} + ||f^* - f||_2 \tag{A.9c}$$

$$\leq \sqrt{||f^* - f||_2(||f^* - f||_2 + 2||f^*||_2)} + ||f^* - f||_2 \tag{A.9d}$$

$$= \sqrt{||f^* - f||_2^2 + 2||f^* - f||_2} + ||f^* - f||_2 \tag{A.9e}$$

Using (A.9e), we have:

$$\frac{||f^* - f||_2^2}{\sigma_f} \leq \frac{||f^* - f||_2^2}{1 - |1 - \sigma_f|} \tag{A.10a}$$

$$\leq \frac{||f^* - f||_2^2}{1 - (\sqrt{||f^* - f||_2^2 + 2||f^* - f||_2} + ||f^* - f||_2)}$$

Combining this with (A.8e):

$$||f^* - \tilde{f}||_2^2 \leq \frac{||f^* - f||_2^2}{1 - (\sqrt{||f^* - f||_2^2 + 2||f^* - f||_2} + ||f^* - f||_2)}$$

$$+ 2 - (\frac{1}{\sigma_f} + \sigma_f) \tag{A.11a}$$

$$t : x \to \frac{x^2}{1 - (\sqrt{x^2 + 2x} + x)} \tag{A.12a}$$

is continuous at 0 so there exists $\gamma_1 > 0$ such that $|x| \leq \gamma_1 \Rightarrow t(x) \leq \frac{\eta^2}{8}$

$$r : x \to 2 - (\frac{1}{x} + x) \tag{A.13a}$$

is continuous at 1 so there exists $\gamma_2 > 0$ such that $|x - 1| \leq \gamma_2 \Rightarrow r(x) \leq \frac{\eta^2}{8}$

$$s : x \to \sqrt{x^2 + 2x} + x \tag{A.14a}$$

is continuous at 0 so there exists $\gamma_3 > 0$ such that $|x| \leq \gamma_3 \Rightarrow |s(x)| \leq \min(\gamma_2, \frac{1}{2})$

By the universal approximation theorem (see corollary 2.2 of (Hornik et al., 1989)) and knowing that U is bounded, we may choose a continuous feedforward network function f_{θ_f} such that:

$$||f^* - f_{\theta_f}||_2 \leq \min(\gamma_1, \gamma_3)$$

By construction of γ_1 and γ_3, f_{θ_f} has positive variance and: $||f^* - \tilde{f}_{\theta_f}||_2 \leq \sqrt{\frac{\eta^2}{8} + \frac{\eta^2}{8}} = \frac{\eta}{2}$

Similarly, we can choose a continuous feed-forward network function g_{θ_g} such that: $||g^* - \tilde{g}_{\theta_g}||_2 \leq \frac{\eta}{2}$

Therefore:

$$HGR(U, V) - \rho(f_{\theta_f}(U), g_{\theta_g}(V)) \leq \epsilon + \eta$$

Taking the limit as ϵ approaches 0:

$$HGR(U, V) - \rho(f_{\theta_f}(U), g_{\theta_g}(V)) \leq \eta$$

For Θ a given subset of $\mathbb{R}^k$ with k the number of coordinates in (θ_f, θ_g), we denote as F_Θ the family of neural networks with the same architecture as $(f_{\theta_f}, g_{\theta_g})$, parametrized by Θ.

We can find a compact set Θ containing (θ_f, θ_g) such that all the elements of F_Θ have positive and finite variance: while the finitude of the variance is due to the

boundedness of U, V and the continuity of the neural networks w.r.t the input, the positivity can be obtained by using the argument of the continuity of the variance w.r.t the parameters (due to the boundedness of U, V and the continuity of the neural networks w.r.t the parameters).

Choosing such a compact set Θ, we obtain the result:

$$|HGR(U,V) - HGR_\Theta(U,V)| \leq \eta. \tag{A.15}$$

□

Lemma A.1.2. *(estimation) Let $\eta > 0$, and F_Θ a family of continuous neural networks parametrized by a compact domain $\Theta \subset R^k$. There exists an $N \in \mathbb{N}$ such that:*

$$\forall n \geq N, |\widehat{HGR(U,V)}_n - HGR_\Theta(U,V)| \leq \eta, a.s. \tag{A.16}$$

Proof. To simplify notations, we will note f and g for $f(U)$ and $g(V)$ when there is no ambiguity.
Let $\eta > 0$. By triangular inequality:

$$\begin{aligned} &|\widehat{HGR(U,V)}_n - HGR_\Theta(U,V)| \\ \leq &\sup_{(\theta_f,\theta_g)\in\Theta} \left|\rho_n(f_{\theta_f}, g_{\theta_g}) - \rho(f_{\theta_f}, g_{\theta_g})\right| \end{aligned} \tag{A.17}$$

We denote E_n the empirical expectation, so that:

$$\rho_n(X,Y) = \frac{E_n(XY) - E_n(X)E_n(Y)}{\sqrt{E_n(X^2) - E_n(X)^2}\sqrt{E_n(Y^2) - E_n(Y)^2}} \tag{A.18}$$

The function $(\theta_f, \theta_g, u, v) \to (f_{\theta_f}(u), g_{\theta_g}(v))$ is continuous on a compact set, so it is bounded. The neural networks are, therefore, uniformly bounded. The compactness of Θ, along with the uniform boundedness argument and the continuity of the neural networks w.r.t their parameters, allows to use the uniform law of large numbers (Geer and van de Geer, 2000) to obtain the almost sure uniform convergence of all empirical expectations in ρ_n, to the corresponding expectations.

The almost sure uniform convergence is compatible with addition, subtraction, multiplication and division, so long as some hypotheses are verified. The compatibility with the first three operations can easily be demonstrated. As for division, we rely on the fact that we can find a uniform positive lower bound for $Var(f_{\theta_f})$ and $Var(g_{\theta_g})$. Indeed, these are positive and continuous functions w.r.t θ_f (*resp.* θ_g) on a compact

set. We can note that this uniform positive lower-bound for the variances, combined with the almost sure uniform convergence of the sample variances, allows us to state that, eventually, all sample variances are positive.

We deduce, by compatibility of operations with almost sure uniform convergence, the almost sure uniform convergence of $\rho_n(f_{\theta_f}, g_{\theta_g})$ to $\rho(f_{\theta_f}, g_{\theta_g})$.

Therefore, by combining the previous result with (A.17), we can find $N \in \mathbb{N}$ such that:

$$\forall n \geq N, |\widehat{HGR(U,V)}_n - HGR_\Theta(U,V)| \leq \eta, a.s. \tag{A.19}$$

□

Theorem A.1.3. *$\widehat{HGR(U,V)}_n$ is strongly consistent.*

Proof. This is a direct consequence of Lemma A.1.1 combined with Lemma A.1.2. □

B

Supplementary Material of Chapter 4

This is the supplementary material of the chapter 4, Ensuring Group Fairness for Neural Network predictors. This supplementary material is as follows. First, Section B.1 provides the proofs of our comparison with simple adversarial algorithms. Then, section B.3 provides further details about our experiments.

B.1 | Comparison With Simple Adversarial Algorithms

Theorem B.1.1. *If $E(Y|X)$ is constant, then $\sup_f \rho(f(X), Y) = 0$. Else, $f^* \in \arg\max_f \rho(f(X), Y)$ iff there exists $a, b \in \mathbb{R}$, with $a > 0$, such that:*

$$f^*(X) = aE(Y|X) + b \tag{B.1}$$

Proof. Let f a function with positive and finite variance w.r.t X.
Using the law of total expectation in (21b) and pulling out the known factor $f(X)$ in (21c):

$$\begin{aligned}
\mathrm{Cov}(f(X), Y) &= E(f(X)Y) - E(f(X))E(Y) && \text{(B.2a)}\\
&= E\Big(E(f(X)Y|X)\Big) - E(f(X))E(E(Y|X)) && \text{(B.2b)}\\
&= E\Big(f(X)E(Y|X)\Big) - E(f(X))E(E(Y|X)) && \text{(B.2c)}\\
&= \mathrm{Cov}(f(X), E(Y|X)) && \text{(B.2d)}
\end{aligned}$$

If $E(Y|X)$ is constant, $\text{Cov}(f(X), Y) = 0$ and therefore $\rho(f(X), Y) = 0$, so that $\sup_f \rho(f(X), Y) = 0$. Else, using B.2 and by the Cauchy-Schwarz inequality:

$$\rho(f(X), Y) = \frac{\text{Cov}(f(X), Y)}{\sigma_{f(X)}\sigma_Y} \tag{B.3a}$$

$$= \frac{\text{Cov}(f(X), E(Y|X))}{\sigma_{f(X)}\sigma_Y} \tag{B.3b}$$

$$\leq \frac{\sigma_{E(Y|X)}}{\sigma_Y} \tag{B.3c}$$

$$= \rho(E(Y|X), Y) \tag{B.3d}$$

The inequality above shows that any linear transformation of $E(Y|X)$ with positive slope maximizes $\rho(f(X), Y)$. Conversely, for $f^* \in \arg\max_f \rho(f(X), Y)$, B.3c is an equality, which gives $\rho(f^*(X), E(Y|X)) = 1$. This implies that there exists $a, b \in \mathbb{R}$, with $a > 0$, such that $f^*(X) = aE(Y|X) + b$. □

Note that a one-dimensional linear regression with $f^*(X)$ as input and Y as output allows to find $E(Y|X)$.

Proposition 2. *Given $Y \sim \mathcal{N}(\mu, \sigma^2)$, $X = \arctan(Y^2) + U\pi$, where $U \perp Y$ and U follows a Bernoulli distribution with $p = \frac{1}{2}$, we have:*

$$E(Y|X) = \tanh\left(\tfrac{\mu}{\sigma^2}\sqrt{\tan(X)}\right)\sqrt{\tan(X)}$$

Proof. We have $Y^2 = \tan(X)$, so that:

$$Y = (2\mathbb{1}_{\{Y>0\}} - 1)\sqrt{\tan(X)} \tag{B.4}$$

so it is sufficient to compute $E(\mathbb{1}_{\{Y>0\}}|X)$:

$$E(\mathbb{1}_{\{Y>0\}}|X) = E\Big(E(\mathbb{1}_{\{Y>0\}}|X, U)\Big|X\Big) \tag{B.5a}$$

$$= E\Big(E(\mathbb{1}_{\{Y>0\}}|\tan(X), U)\Big|X\Big) \tag{B.5b}$$

$$= E\Big(E(\mathbb{1}_{\{Y>0\}}|\tan(X))\Big|X\Big) \tag{B.5c}$$

$$= E(\mathbb{1}_{\{Y>0\}}|\tan(X)) \tag{B.5d}$$

$$= E(\mathbb{1}_{\{Y>0\}}|Y^2) \tag{B.5e}$$

Let $y > 0$ and $0 < \epsilon < y$:

$$
\begin{aligned}
&E(\mathbb{1}_{\{Y>0\}} | |Y^2 - y| < \epsilon) \\
&= \frac{\mathbb{P}(Y > 0, \sqrt{y-\epsilon} < Y < \sqrt{y+\epsilon})}{\mathbb{P}(\sqrt{y-\epsilon} < Y < \sqrt{y+\epsilon}) + \mathbb{P}(-\sqrt{y+\epsilon} < Y < -\sqrt{y-\epsilon})} && \text{(B.6a)} \\
&= \frac{\int_{\sqrt{y-\epsilon}}^{\sqrt{y+\epsilon}} P_Y(u)du}{\int_{\sqrt{y-\epsilon}}^{\sqrt{y+\epsilon}} P_Y(u)du + \int_{-\sqrt{y+\epsilon}}^{-\sqrt{y-\epsilon}} P_Y(u)du} && \text{(B.6b)} \\
&= \frac{\int_{\sqrt{y-\epsilon}}^{\sqrt{y+\epsilon}} P_Y(u)du}{\int_{\sqrt{y-\epsilon}}^{\sqrt{y+\epsilon}} (P_Y(u) + P_Y(-u))du} \xrightarrow[\epsilon \to 0]{} \frac{P_Y(\sqrt{y})}{P_Y(\sqrt{y}) + P_Y(-\sqrt{y})} && \text{(B.6c)}
\end{aligned}
$$

Therefore, knowing that $P_Y(y) = \frac{e^{-\frac{1}{2}(\frac{y-\mu}{\sigma})^2}}{\sqrt{2\pi}}$ we have:

$$
\begin{aligned}
2E(\mathbb{1}_{\{Y>0\}}|X) - 1 &= \frac{P_Y(|Y|) - P_Y(-|Y|)}{P_Y(|Y|) + P_Y(-|Y|)} && \text{(B.7a)} \\
&= \tanh\left(\frac{\mu}{\sigma^2}|Y|\right) && \text{(B.7b)} \\
&= \tanh\left(\frac{\mu}{\sigma^2}\sqrt{\tan(X)}\right) && \text{(B.7c)}
\end{aligned}
$$

Taking the conditional expectation in B.4 and plugging in B.7c, we obtain:

$$E(Y|X) = \tanh\left(\frac{\mu}{\sigma^2}\sqrt{\tan(X)}\right)\sqrt{\tan(X)}$$

□

Proposition 3. *With the same hypotheses as in proposition 2, and denoting $\alpha = \frac{\mu}{\sigma}$, we have:*

$$\sqrt{1 - e^{-\frac{\alpha^2}{2}}} \leq \rho(E(Y|X), Y) \leq \sqrt{1 - e^{-\frac{\alpha^2}{2}}(1+\alpha^2)^{-\frac{3}{2}}}$$

Proof. We first note that, knowing that $|Y| = \sqrt{\tan(X)}$ and with a parity argument:

$$E(Y|X) = \tanh\left(\frac{\mu}{\sigma^2}Y\right)Y \tag{B.8}$$

We have:

$$
\begin{aligned}
\rho(E(Y|X), Y)^2 &= \frac{\mathrm{Cov}(Y, E(Y|X))}{\sigma^2} && \text{(B.9a)} \\
&= \frac{\mathrm{Cov}(Y, Y) - \mathrm{Cov}(Y, Y - E(Y|X))}{\sigma^2} && \text{(B.9b)} \\
&= 1 - E\left(\left(\frac{Y}{\sigma}\right)^2\left(1 - \tanh\left(\frac{\mu}{\sigma^2}Y\right)\right)\right) && \text{(B.9c)}
\end{aligned}
$$

With a variable change ($y = \frac{y'}{\sigma}$), we obtain:

$$
\begin{aligned}
& E\left(\left(\frac{Y}{\sigma}\right)^2\left(1-\tanh\left(\frac{\mu}{\sigma^2}Y\right)\right)\right) \\
&= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \frac{2y^2 e^{-\alpha y}}{e^{\alpha y}+e^{-\alpha y}} e^{-\frac{1}{2}(y-\alpha)^2} dy \qquad \text{(B.10a)} \\
&= e^{-\frac{\alpha^2}{2}} \times \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \frac{y^2}{\cosh(\alpha y)} e^{-\frac{1}{2}y^2} dy \qquad \text{(B.10b)}
\end{aligned}
$$

We have, for all $y \in \mathbb{R}$, $1 \leq \cosh(\alpha y) \leq e^{\frac{\alpha^2 y^2}{2}}$. This gives:

$$
\begin{aligned}
\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} y^2 e^{-\frac{1}{2}y^2} dy &\leq \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \frac{y^2}{\cosh(\alpha y)} e^{-\frac{1}{2}y^2} dy \\
&\leq \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} y^2 e^{-\frac{1}{2}(1+\alpha^2)y^2} dy \qquad \text{(B.11)}
\end{aligned}
$$

i.e

$$
1 \leq \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \frac{y^2}{\cosh(\alpha y)} e^{-\frac{1}{2}y^2} dy \leq (1+\alpha^2)^{-\frac{3}{2}} \qquad \text{(B.12)}
$$

We combine B.9c, B.10b and B.12 to obtain the result:

$$
\sqrt{1-e^{-\frac{\alpha^2}{2}}} \leq \rho(E(Y|X), Y) \leq \sqrt{1-e^{-\frac{\alpha^2}{2}}(1+\alpha^2)^{-\frac{3}{2}}}
$$

□

Proposition 4. *We consider the global fairness objective of the prediction retreatment simple adversarial algorithm, with X the input data, Y the output data and S the sensitive attribute (with $\widehat{Y} = f(X)$):*

$$
\max_f \min_g E\Big(\big(S - g(f(X))\big)^2\Big) \qquad \text{(B.13)}
$$

whose optimum is achieved when $E(S|\widehat{Y}) = E(S)$, different from the demographic parity fairness objective $P(S|\widehat{Y}) = P(S)$ for continuous features.

Proof. We have:

$$
\max_f \min_g E\Big(\big(S - g(f(X))\big)^2\Big) = \max_f E\Big(\big(S - E(S|f(X))\big)^2\Big)
$$

Some algebraic manipulations with expectations give:

$$
\begin{aligned}
E\Big(\big(S - E(S|\widehat{Y})\big)^2\Big) &= E(S^2) - 2E\Big(SE(S|\widehat{Y})\Big) \\
&\quad + E(E(S|\widehat{Y})^2) && \text{(B.14a)} \\
&= E(S^2) - 2E\Big(E\big(SE(S|\widehat{Y})|\widehat{Y}\big)\Big) \\
&\quad + E(E(S|\widehat{Y})^2) && \text{(B.14b)} \\
&= E(S^2) - E(E(S|\widehat{Y})^2) && \text{(B.14c)} \\
&= (E(S^2) - E(S)^2) \\
&\quad - \Big(E(E(S|\widehat{Y})^2) - E(E(S|\widehat{Y}))^2\Big) && \text{(B.14d)} \\
&= \sigma_S^2 - \sigma^2_{E(S|\widehat{Y})} && \text{(B.14e)}
\end{aligned}
$$

Therefore, the global fairness objective is equivalent to

$$\min_f \sigma^2_{E(S|f(X))}$$

In the optimal case, we have $\sigma_{E(S|\widehat{Y})} = 0$, which corresponds to the case when $E(S|\widehat{Y})$ is constant equal to its expectation i.e:

$$E(S|\widehat{Y}) = E(S)$$

□

B.2 | Equalized Residuals

We propose an adversarial approach based on our *HGR* neural network estimation (Grari et al., 2020a) for enforcing equalized residuals. It uses an adversarial network that takes the form of two inter-connected neural networks for approximating the optimal transformations functions f and g for approximating the sensitive dependence by HGR.

$$\arg\min_{w_h} \Big\{ \max_{w_f, w_g} \big\{ \mathcal{L}(h_{w_h}(X), Y) + \lambda \mathbb{E}_{(X,S)\sim\mathcal{D}}(\widehat{f}_{w_f}(h_{w_h}(X) - Y)\widehat{g}_{w_g}(S) \big\} \Big\} \tag{B.15}$$

where $\mathcal{L}$ is the predictor loss function between the output $h_{w_h}(X) \in \mathbb{R}$ and the corresponding target Y. The second term, which corresponds to the expectation of the products of standardized outputs of both networks ($\hat{f}_{w_f}$ and $\hat{g}_{w_g}$), represents the HGR estimation between the residuals variable $h_{w_h}(X) - Y$ and the sensitive attribute S. The hyperparameter λ controls the impact of the dependence loss in the optimization.

B.3 | Experiments

B.3.1 | Datasets

Our experiments on real-world data are performed on five data sets. First, we experiment with three data sets where the sensitive and the outcome true value are both continuous:

- The US Census demographic data set (US Census Bureau, 2019) is an extraction of the 2015 American Community Survey, with 37 features about 74,000 census tracts. The target is the percentage of children below the poverty line, the sensitive attribute is the percentage of women in the census tract.
- The Motor Insurance data set (The Institute of Actuaries of France, 2015) originates from a pricing game organized by The French Institute of Actuaries in 2015, with 15 attributes for 36,311 observations. The target is the average claim cost per policy, the sensitive attribute is the driver's age.
- The Crime data set is obtained from the UCI Machine Learning Repository (Dua and Graff, 2017), with 128 attributes for 1,994 instances. The target is the number of violent crimes per population, the sensitive attribute is the ratio of an ethnic group per population.

We experiment with two data sets with a binary classification task where the sensitive features are continuous:

- Compas: The COMPAS data set (Angwin et al., 2016) contains 13 attributes of about 7,000 convicted criminals with class labels that state whether or not the individual reoffended within 2 years of their most recent crime. Here, we use age as sensitive attribute.
- Default: The Default data set (Yeh and Lien, 2009) contains 23 features about 30,000 Taiwanese credit card users with class labels which state whether an individual will default on payments. As sensitive attribute, we use age.

B.3.2 | Experimental Parameters

For the reproducibility of the experimental results, we reported the deep learning architecture and the different hyperparameters chosen. For all data sets, we repeat

five experiments by randomly sampling two subsets, 80% for the training set and 20% for the test set.

Since the different data sets are not large, we train the different algorithms on a NVIDIA Titan Xp (12 Gb) GPU and we report the average runtime of each scenario (Runtime (s)). Note that we use an Adam optimization for each scenario.

Scenario	λ	Nb Epochs	Batch Size	Architecture h_{w_ψ}	Architecture ϕ_{w_ϕ}	Architecture f_{w_f} & g_{w_g}	Runtime (s)
Biased Model	0	200	2048	FC:16 R, FC:8 R, FC:2	FC:16 R, FC:8 R, FC:4 R, FC:1 Sig	FC:64 R, FC:64 R, FC:1	303
Biased Model	13	200	2048	FC:16 R, FC:8 R, FC:2	FC:16 R, FC:8 R, FC:4 R, FC:1 Sig	FC:64 R, FC:64 R, FC:1	287

Table B.1: Synthetic Scenario. FC stands for fully connected, R for the ReLU activation function and Sig for the Sigmoid activation function.

Scenario	λ	Nb Epochs	Batch Size	Architecture h_{w_ψ}	Architecture ϕ_{w_ϕ}	Architecture f_{w_f} & g_{w_g}	Runtime (s)
$\sigma \leq 0.03$	0.250	10	512	see Table B.4	FC:10 SM	FC:64 R, FC:64 R, FC:1	326
$\sigma > 0.04$	0.100	10	512	see Table B.4	FC:10 SM	FC:64 R, FC:64 R, FC:1	371

Table B.2: MNIST with Continuous Color Intensity. FC stands for fully connected, R for ReLU, SM for the Softmax activation function.

Scenario	λ	Nb Epochs	Batch Size	Architecture h_{w_ψ}	Architecture ϕ_{w_ϕ}	Architecture f_{w_f} & g_{w_g}	Runtime (s)
US Census	20	150	2048	FC:128 R, FC:64 R,FC:64	FC:128 R, FC:64 R,FC:16 R,FC:1	FC:64 R, FC:64 R, FC:1	1873
Motor	1.5	1000	2048	FC:128 R, FC:64 R,FC:64	FC:128 R, FC:64 T,FC:16 R,FC:1	FC:64 R, FC:64 T, FC:1	235
Crime	3	3000	512	FC:128 R, FC:64 R,FC:64	FC:128 R, FC:64 T,FC:16 R,FC:1	FC:64 R, FC:64 T, FC:1	1584
COMPAS	200	850	2048	FC:128 R, FC:64 R,FC:64	FC:128 R, FC:64 R,FC:16 R,FC:1 Sig	FC:64 R, FC:64 R, FC:1	1721
Default	100	400	2048	FC:128 R, FC:64 R,FC:64	FC:128 R, FC:64 R,FC:16 R,FC:1 Sig	FC:64 R, FC:64 R, FC:1	3378

Table B.3: Real-world Experiments. FC stands for fully connected, T for Tanh, R for the ReLU activation function and Sig for the Sigmoid activation function.

	Encoder MNIST h_{w_ψ}			
Layer	Number of outputs	Kernel size	Stride	Activation function
Input x	$3 * 28 * 28$			
Convolution	$64 * 26 * 26$	$5 * 5$	1	ReLU
MaxPooling	$64 * 13 * 13$	-	2	-
Convolution	$64 * 11 * 11$	$5 * 5$	1	ReLU
MaxPooling	$64 * 5 * 5$	-	2	-
Flatten	-	-	-	-
Fully-connected	512	-	-	ReLU
Fully-connected	64	-	-	None

Table B.4: Encoder h_{w_ψ} used for the MNIST Scenario with Continuous Color Intensity

C

Supplementary Material of Chapter 6

This is the supplementary material of chapter 6, Group Fairness without the sensitive attribute. This supplementary material is as follows. First, Section C.1 provides the proofs of our HGR theorem 6.2. Then, section C.2 provides further details about an extended causal graph. Finally, section C.3 provides further details about our experiments.

C.1 | Proof of the HGR Inequality

Theorem C.1.1. *For two nonempty index set S and Z such that $S \subset Z$ and $\hat{Y}$ the output prediction of a predictor model, we have :*

$$HGR(\hat{Y}, Z) \geq HGR(\hat{Y}, S) \tag{C.1}$$

Proof. Let's assume that the set Z^c represent all the elements of Z private of S: $Z^c = Z \backslash S$, Following the definition of the Hirschfeld-Gebelein-Renyi Maximum Correlation Coefficient, we have:

$$HGR(\hat{Y}, Z) = \sup_{f:\mathcal{U}\to\mathbb{R}, g:\mathcal{V}\to\mathbb{R}} \rho(f(\hat{Y}), g(Z)) \tag{C.2}$$

By the Cauchy inequality and by setting $f(Z) = E[g(\hat{Y})|Z]$, we can show the

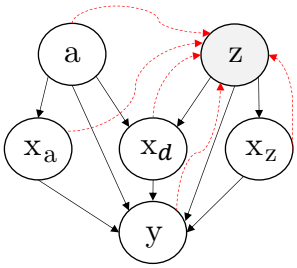


Figure C.1: Causal graph of the SRCVAE: The following causal graph represents the more general representation where x is mapped into four components x_a, x_d, x_z and a;

equivalence characterization of the HGR:

$$HGR(\hat{Y}, Z) = \sup_g \sqrt{\frac{var(E(g(\hat{Y})|Z)])}{var(g(\hat{Y}))}} \tag{C.3}$$

$$= \sup_g \sqrt{\frac{E(E[g(\hat{Y})|S, Z^c]^2) + E(g(\hat{Y})|S, Z^c)^2}{var(g(\hat{Y}))}} \tag{C.4}$$

$$= \sup_g \sqrt{\frac{E(E(E[g(\hat{Y})|S, Z^c]^2|S)) + E(g(\hat{Y})|S, Z^c)^2)}{var(g(\hat{Y}))}} \tag{C.5}$$

$$\geq \sup_g \sqrt{\frac{E(E[g(\hat{Y})|S]^2) + E(g(\hat{Y})|S, Z^c)^2}{var(g(\hat{Y}))}} = HGR(\hat{Y}, S) \tag{C.6}$$

We use in (6) the Jensen inequality for conditional expectation. □

C.2 | An Extended Causal Graph

As explained in the paper, our work relies on the assumption of underlying causal graphs. In the figure C.1 we present a more general graph, where parents of the output y are split in five components x_a, x_d, x_z, x_c and z, where x_c contains only variables not caused by the sensitive attributes in z. Although this graph is more general, we have considered, for a sake of presentation, a simplified version in our paper, which is enough to capture most dependencies in most settings. We denote by x_a (resp., x_z), the set only caused by x_c (resp., z). The variables subset x_d is both caused by the sensitive information z and x_c. In our setting, we assume that we observe $x = (x_c, x_a, x_d, x_z)$, but variables in z remain hidden from the learner.

The decoder distribution, $p_\theta(x_c, x_a, x_d, x_z, y|z)$, can be factorized as below:

$$p_\theta(x_c, x_a, x_d, x_z, y|z) = p(x_c)p(x_a|x_c)p_\theta(x_d|x_c, z)p_\theta(x_z|z) \\ p_\theta(y|x_c, x_a, x_d, x_z, z) \tag{C.7}$$

Given an approximate posterior $q_\phi(z|x_c, x_a, x_d, x_z, y)$, we obtain the variational lower bound as Eq. C.8

$$\begin{aligned} log(p_\theta(x_c, x_a, x_d, x_z, y)) \geq \mathbb{E}_{\substack{(x_c,x_a,x_d,x_z,y)\sim\mathcal{D}, \\ z\sim q_\phi(z|x_c,x_a,x_d,x_z,y)}} [\log p_\theta(x_d|x_c, z) \\ + \log p_\theta(y|x_c, x_a, x_d, x_z, z) + \log p_\theta(x_z|z) \\ - D_{KL}(q_\phi(z|x_c, x_a, x_d, x_z, y)||p(z))] \\ =: -\mathcal{L}_{ELBO} \end{aligned} \tag{C.8}$$

where D_{KL} denotes the Kullback-Leibler divergence of the posterior $q_\phi(z|x_c, x_a, x_d, x_z, y)$ from a prior $p(z)$, typically a standard Gaussian distribution $\mathcal{N}(0, I)$. The posterior $q_\phi(z|x_c, x_a, x_d, x_z, y, z)$ is represented by a deep neural network with parameters ϕ, which typically outputs the mean μ_ϕ and the variance σ_ϕ of a diagonal Gaussian distribution $\mathcal{N}(\mu_\phi, \sigma_\phi I)$. The likelihood term factorizes as $p_\theta(x_d, x_z, y|x_c, x_a, x_d, x_z, z) = p_\theta(x_d|x_c, z)p_\theta(x_z|z)p_\theta(y|x_c, x_a, x_d, x_z, z)$, are defined as neural networks with parameters w_h. The maximization of this marginal log-likelihood is realized by the minimization of the negative lower bound, mentioned as $\mathcal{L}_{ELBO}$. Since attracted by a standard prior, the posterior is supposed to remove probability mass for any features of z that are not involved in the reconstruction of x_d, x_z and y. Since x_c is given together with z as input of the likelihoods, all the information from x_c should be removed from the posterior distribution of z. In practice, we use the neural network layers to infer the parameters of a Gaussian distribution over the joint space of z. To obtain the posterior distribution of q_ϕ, $p(z)$ is the prior distributions following the Gaussian distribution, and we utilize the reparametrization, accordingly. Notice that the complementary set x_c is not involved in any specific reconstruction.

In addition, we employ in this paper a variant of the ELBO optimization as done in (Pfohl et al., 2019), where the $D_{KL}(q_\phi(z|x_c, x_a, x_d, x_z, y)||p(z))$ term is replaced by a MMD term $\mathcal{L}_{MMD}(q_\phi(z)||p(z))$ between the aggregated posterior $q_\phi(z)$ and the prior. This has been shown more powerful than the classical D_{KL} for ELBO optimization in (Zhao et al., 2017), as the latter can reveal as too restrictive (uninformative latent code problem) (Chen et al., 2016; Bowman et al., 2015; Sønderby et al., 2016) and can also tend to overfit the data (Variance Over-estimation in Feature Space).

This inference must however ensure that no dependence is created between x_c and z (no arrow from x_c to z in the graph from C.1), unless preventing the generation of proper sensitive proxy which is not linked to the complementary. However, by optimizing this ELBO optimization, some dependence can still be observed empirically between x_c and z. Some information from x_c leaks in the inferred z. In order to ensure some minimum independence level we add a dependence penalisation term

in this loss function. Leveraging the last research for mitigating the dependence with continuous multidimensional space we extend the main idea of (Grari et al., 2021b) by adapting this penalization in the variational autoencoder case. Originally, this paper used an HGR estimation in a minmax game to penalize the intrinsic bias in a multi dimensional latent representation for deterministic autoencoder. They have showed that a neural HGR-based approach presents a very competitive results in the continuous case by identifying some optimal transformations for multidimensional features. Finally, the inference of our SRCVAE with the more general representation is optimized by a mini-max game as follows:

$$\begin{aligned}\arg\min_{w_h,\phi}\max_{w_f,w_g} - \mathbb{E}_{\substack{(x_c,x_a,x_d,x_z,y)\sim\mathcal{D},\\ z\sim q_\phi(z|x_c,x_a,x_d,x_z,y)}} & [[\log p_\theta(x_d|x_c,z) \\ & + \log p_\theta(y|x_c,x_a,x_d,x_z,z) + \log p_\theta(x_z|z)] \\ & + \lambda_{mmd}\mathcal{L}_{MMD}(q_\phi(z)||p(z)) \\ & + \lambda_{inf}\mathbb{E}_{\substack{(x_c,x_a,x_d,x_z,y)\sim\mathcal{D},\\ z\sim q_\phi(z|x_c,x_a,x_d,x_z,y)}} (\widehat{f}_{w_f}(x_c)\widehat{g}_{w_g}(z))]\end{aligned}$$

where λ_{mmd}, λ_{inf} are scalar hyperparameters. The additional MMD objective can be interpreted as minimizing the distance between all moments of each aggregated latent code distribution and the prior distribution. Note that the use of y as input for our generic inference scheme $q(z|x_c,x_a,x_d,x_z,y)$ is allowed since z is only used during training for learning a fair predictive model and is not used at deployment time.

The complementary set x_c is the only input given to the adversarial f and the continuous latent space z as input for the adversarial g. In that case, we only capture for each gradient iteration the estimated HGR between the complementary set and the generated proxy latent space. The algorithm takes as input a training set from which it samples batches of size b at each iteration. At each iteration it first standardize the output scores of networks f_{w_f} and g_{w_g} to ensure 0 mean and a variance of 1 on the batch. Then it computes the objective function to maximize to estimate the HGR score and the global variational reconstruction objective. At the end of each iteration, the algorithm updates the parameters of the encoders parameters w_h as well as the decoder parameters ϕ by one step of gradient descent. Concerning the HGR adversary, the backpropagation of the parameters w_f and w_g is performed by multiple steps of gradient ascent. This allows us to optimize a more accurate estimation of the HGR at each step, leading to a greatly more stable learning process.

C.3 | Experiments

We present in this following section the Impact of the Mitigation hyper-parameters, the choice of our HGR compared to other penalties and the scenario of proxy dimensions on Default Dataset.

C.3.1 | Impact of the Mitigation Hyper-Parameters

We present in Figure C.2 the dynamics of the adversarial training with different hyperparameters λ_{DP}, optimized for demographic parity. The choice of this value depends on the main objective, resulting in a trade-off between accuracy and fairness. We represent the accuracy of the model (top), the P-rule metric between the prediction and the real sensitive S (middle), and the HGR between the prediction and the latent space Z (bottom).

As desired, higher values of λ_{DP} produce fairer predictions, the P-rule (assessed with the real sensitive) increase, while λ_{DP} near 0 allows to only focus on optimizing the classifier predictor.

C.3.2 | Mitigation With Other Penalties

In the following subsection, we assume have first pretrained a bayesian inference model q_ϕ for reconstructing a sensitive proxy Z via our SRCVAE architecture (via the simplified graph in our paper). We illustrate, for our specific case, the interest of adopting an HGR-based approach inspired for mitigating unwanted biases. For this purpose, we extend and compare different common state-of-the-art mitigating algorithms for the demographic parity task. We focus on in-processing fairness, which proves to be the most powerful framework for settings where acting on the training process is an option.

We extend for each of these state-of-the-art algorithms the mitigation of our sensitive proxy representation z.

SA_PR: Simple Adversarial via Prediction Retreatment We extend the idea of (Zhang et al., 2018) by proposing a novel adversarial cost for fairness without demographics via prediction retreatment.

$$\begin{aligned}&\arg\min_{w_h}\max_{w_f}\mathcal{L}(h_{w_h}(x_c,x_d),y)\\&-\lambda_{DP}\mathbb{E}_{(x_c,x_d,y)\sim\mathcal{D}}[\mathbb{E}_{z\sim q_\phi(z|x_c,x_d,y)}(f_{w_f}(h_{w_h}(x_c,x_d))-z)^2]\end{aligned}$$

where $\mathcal{L}$ is the predictor loss function (the log-loss function in our experiments) between the output $h_{w_h}(x_c, x_d) \in \mathbb{R}$ and the corresponding target y, with h_{w_h} a neural network with parameters w_h which takes as input the complementary set x_c and the descendant attribute x_d. The hyperparameter λ_{DP} controls the impact of the dependence loss in the optimization. Following the causal model learned at step 1, z is obtained by inferring from the posterior distribution $q_\phi(z|x_{d_i}, x_{c_i}, y_i)$. We generate for each observation (x_{c_i}, x_{d_i}, y_i) multiple latent representation z_i^k (k-ith generation), which allows to have a better understanding of the specific sensitive distribution for each individual. In practice, for the i-th individual of the training set we generate K samples from the causal model (200 in our experiment). The adversarial f with parameters w_f takes as input the prediction $h_{w_h}(x_{d_i}, x_{c_i}) = p_{w_h}(y_i = 1|x_{d_i}, x_{c_i})$ and tries to predict the sensitive proxy z. A mean square loss function is applied for this objective. The algorithm takes as input a training set from which it samples batches of size b at each iteration. The backpropagation of the adversary f with parameters ω_f is performed by multiple steps of gradient descent (50 in our experiments). This allows us to optimize a more accurate estimation of the biases at each step, leading to a greatly more stable learning process.

SA_FR: Simple Adversarial via Fair Representation: We extend the idea of (Adel et al., 2019) by proposing a novel adversarial representation cost for fairness without demographics via fair representation.

$$\arg\min_{w_h, w_g} \max_{w_f} \mathcal{L}(h_{w_h}(g_{w_g}(x_c, x_d)), y)$$
$$- \lambda_{DP} \mathbb{E}_{(x_c, x_d, y)\sim\mathcal{D}}[\mathbb{E}_{z\sim q_\phi(z|x_c, x_d, y)}(f_{w_f}(g_{w_g}(x_c, x_d))) - z)^2]$$

where g_{w_g} with parameters w_g is the encoder which takes as input the complementary set x_c and the descendant attribute x_d, $\mathcal{L}$ is the predictor loss function (the log-loss function in our experiments) between the output $h_{w_h}(g_{w_g}(x_c, x_d)), y) \in \mathbb{R}$ and the corresponding target y, with h_{w_h} a neural network with parameters w_h which takes as input the representation of the encoder g_{w_g}. The hyperparameter λ_{DP} controls the impact of the dependence loss in the optimization. Following the causal model learned at step 1, z is obtained by inferring from the posterior distribution $q_\phi(z|x_{d_i}, x_{c_i}, y_i)$. We generate for each observation (x_{c_i}, x_{d_i}, y_i) multiple latent representation z_i^k (k-ith generation), which allows to have a better understanding of the specific sensitive distribution for each individual. In practice, for the i-th individual of the training set we generate K samples from the causal model (200 in our experiment). The representation $g_{w_g}(x_c, x_d)$ is the only input given to the adversarial f which aims to predict the continuous latent space z . A mean square loss function is applied for this objective.

The algorithm takes as input a training set from which it samples batches of size b at each iteration. As in the inference phase, the backpropagation of the adversary f with parameters ω_f is performed by multiple steps of gradient descent (50 steps). This allows us to optimize a more accurate estimation of the unwanted biases at each step, leading to a greatly more stable learning process.

RA_PR: Rényi Adversarial via Prediction Retreatment: This is the proposed mitigation used in the paper where we extend the idea of (Grari et al., 2020a) by proposing a novel HGR adversarial cost for fairness without demographics via prediction retreatment.

$$\arg\min_{w_h} \max_{\psi_f, \psi_g} \mathcal{L}(h_{w_h}(x_c, x_d), y) + \lambda_{DP} \underset{\substack{(x_c, x_d, y)\sim\mathcal{D}, \\ z\sim q_\phi(z|x_c, x_d, y)}}{\widehat{HGR}_{\psi_f, \psi_g}} (h_{w_h}(x_c, x_d), z)$$

where $\mathcal{L}$ is the predictor loss function (the log-loss function in our experiments) between the output $h_{w_h}(x_c, x_d) \in \mathbb{R}$ and the corresponding target y, with h_{w_h} a neural network with parameters w_h which takes as input the complementary set x_c and the descendant attribute x_d. The hyperparameter λ_{DP} controls the impact of the dependence loss in the optimization. Following the causal model learned at step 1, z is obtained by inferring from the posterior distribution $q_\phi(z|x_{d_i}, x_{c_i}, y_i)$. We generate for each observation (x_{c_i}, x_{d_i}, y_i) multiple latent representation z_i^k (k-ith generation), which allows to have a better understanding of the specific sensitive distribution for each individual. In practice, for the i-th individual of the training set we generate K samples from the causal model (200 in our experiment). The adversarial ψ_f takes as input the prediction $h_{w_h}(x_{d_i}, x_{c_i}) = p_{w_h}(y_i = 1|x_{d_i}, x_{c_i})$ and the adversarial ψ_g takes as input the continuous latent space z_i^k (k-ith generation). In that case, we only capture for each gradient iteration the estimated HGR between the prediction and the proxy latent space. The algorithm takes as input a training set from which it samples batches of size b at each iteration. At each iteration, it first standardizes the output scores of networks ψ_f and ψ_g to ensure 0 mean and a variance of 1 on the batch. Then it computes the objective function to maximize the estimated HGR score and the global regression objective. As in the inference phase, the backpropagation of the HGR adversary with parameters f and g is performed by multiple steps of gradient ascent. This allows us to optimize a more accurate estimation of the HGR at each step, leading to a greatly more stable learning process.

RA_FR: Rényi Adversarial via Fair Representation: We extend the idea of (Grari et al., 2021b) by proposing a novel adversarial cost for fairness without demographics via fair representation.

$$\arg\min_{w_h, w_v} \max_{\psi_f, \psi_g} \mathcal{L}(h_{w_h}(v_{w_v}(x_c, x_d)), y) + \lambda_{DP} \underset{\substack{(x_c, x_d, y)\sim\mathcal{D}, \\ z\sim q_\phi(z|x_c, x_d, y)}}{\widehat{HGR}_{\psi_f, \psi_g}} (h_{w_h}(v_{w_v}(x_c, x_d)), z)$$

where v_{w_v} with parameters w_v is the encoder which takes as input the complementary set x_c and the descendant attribute x_d, $\mathcal{L}$ is the predictor loss function (the log-loss function in our experiments) between the output $h_{w_h}(v_{w_v}(x_c, x_d)) \in \mathbb{R}$ and the corresponding target y, with h_{w_h} a neural network with parameters w_h which takes as input the representation of the encoder v_{w_v}. The hyperparameter λ_{DP} controls the impact of the dependence loss in the optimization. Following the causal model learned at step 1, z is obtained by inferring from the posterior distribution $q_\phi(z|x_{d_i}, x_{c_i}, y_i)$. We generate for each observation (x_{c_i}, x_{d_i}, y_i) multiple latent representation z_i^k (k-ith generation), which allows to have a better understanding of the specific sensitive distribution for each individual. In practice, for the i-th individual of the training set we generate K samples from the causal model (200 in our experiment). The adversarial ψ_f with parameters f takes as input the representation $v_{w_v}(x_{d_i}, x_{c_i})$ and the adversarial ψ_g takes as input the continuous latent space z_i^k (k-ith generation). In that case, we only capture for each gradient iteration the estimated HGR between the latent representation and the proxy latent space. The algorithm takes as input a training set from which it samples batches of size b at each iteration. At each iteration, it first standardizes the output scores of networks ψ_f and ψ_g to ensure 0 mean and a variance of 1 on the batch. Then it computes the objective function to maximize the estimated HGR score and the global regression objective. As in the inference phase, the backpropagation of the HGR adversary with parameters f and g is performed by multiple steps of gradient ascent. This allows us to optimize a more accurate estimation of the HGR at each step, leading to a greatly more stable learning process.

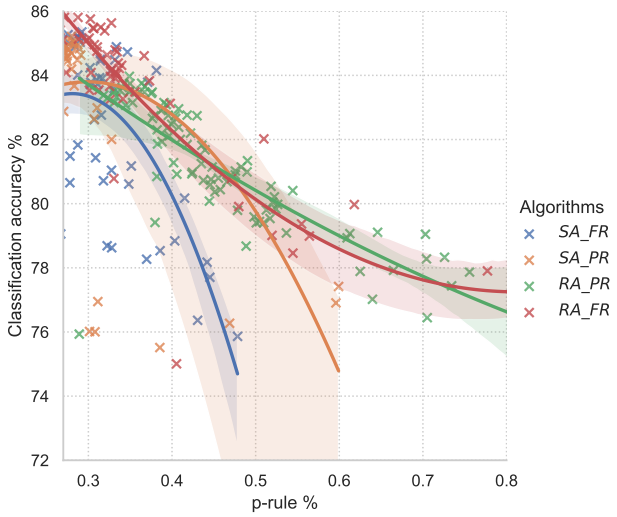
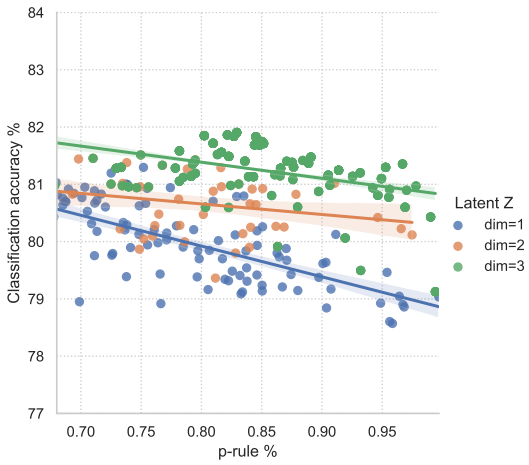


Figure C.3: Demographic Parity task - Adversarial comparison for different level of fairness

We plot the performance of these different approaches by displaying the Accuracy against the P-rule for Demographic Parity (Figure C.3) on Adult UCI data set. We clearly observe for all algorithms that the Accuracy, or predictive performance, decreases when fairness increases. We note that, for all levels of fairness (controlled by the mitigation weight in every approach), Rényi architectures (RA_FR and RA_PR) outperforms simple adversarial architectures (SA_FR and SA_PR) for fairness tasks (except some points for very low levels of fairness, at the left of the curves for SA_PR). We attribute this to the ability of the Rényi to capture more complex - non linear dependencies. We also observe, for these Rényi penalizations, that the architecture with fair Representation is relatively similar to Prediction Retreatment. For the sake of presentation and lower complexity in model we have adopted the RA_PR version in our paper. It has the advantage of containing one less encoder network for quite comparable results.

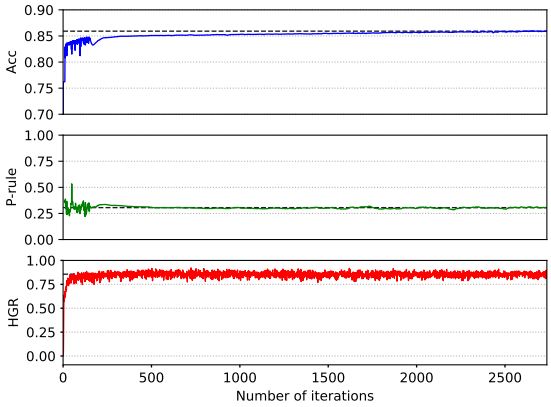
C.3.3 | Proxy Dimensions



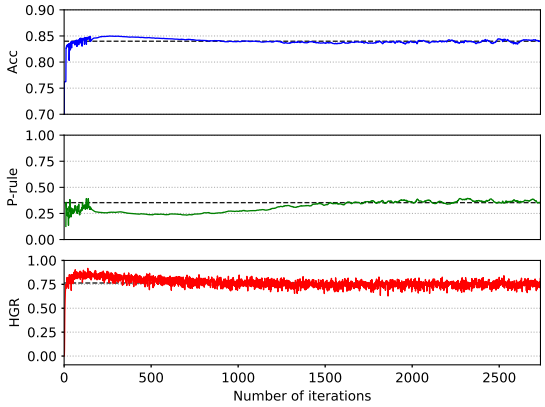
(a) Default Data set

Figure C.4: Additional Experiments

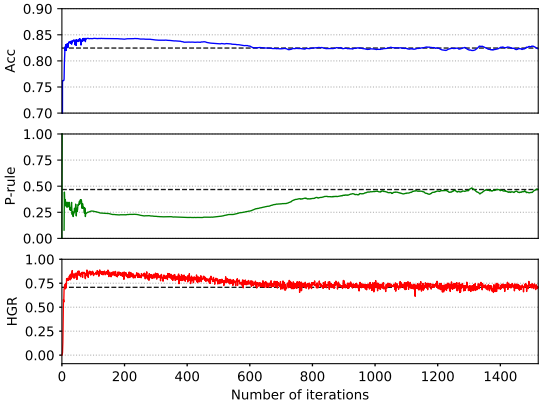
In figure C.4, we perform an additional experiment on the sensitive proxy for th Default data set. We observe that increasing z dimensions results in increased accuracy. Increasing the dimensions to 3 allows to obtain better results in terms of accuracy and this for all levels of P-rule. We claim that mitigating biases in larger spaces allows better generalisation abilities at test time, as already observed in another context in (Grari et al., 2021b). It supports the choice of considering a multivariate sensitive proxy z, rather than directly acting on a reconstruction of s as a univariate variable.



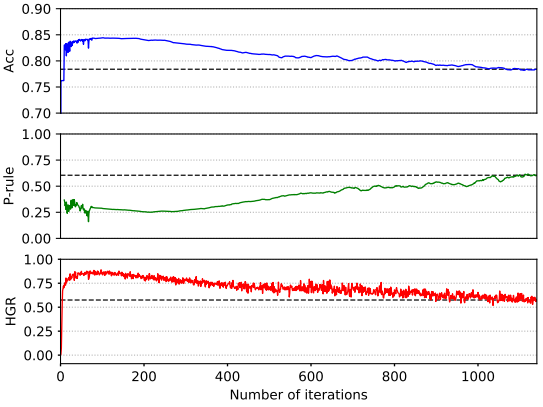
(a) $\lambda = 0.00$; $P - rule = 29.5\%$



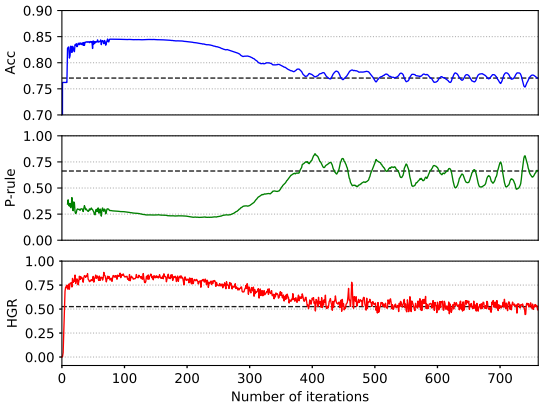
(b) $\lambda = 0.24$; $P - rule = 32.1\%$



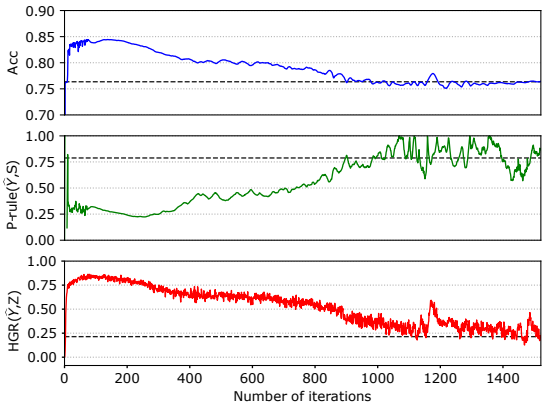
(c) $\lambda = 0.35$; $P - rule = 49.5\%$



(d) $\lambda = 0.45$; $P - rule = 58.9\%$



(e) $\lambda = 0.48$; $P - rule = 65.4\%$



(f) $\lambda = 0.5$; $P - rule = 89.5\%$

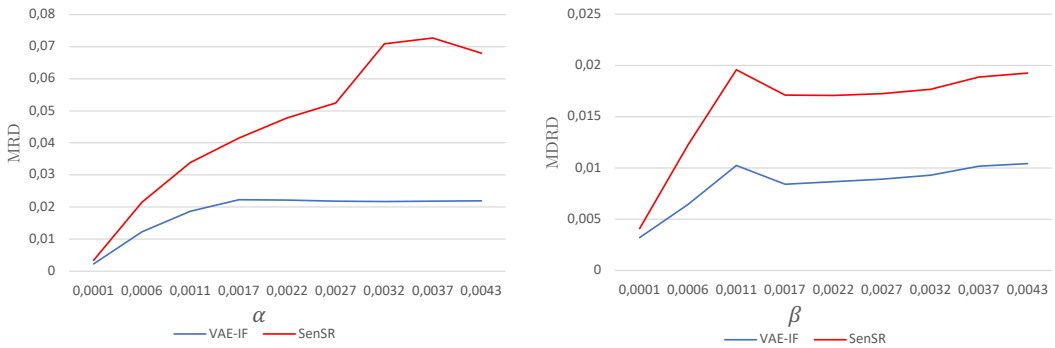
Figure C.2: Training curves with different hyperparameters λ_{DP}

D

Supplementary Material of Chapter 8

D.1 | Threshold Choice for Individual Fairness Metric

We plot in Figure D.1, for the Adult UCI data set, the performance of our proposed algorithm (section 8.1) against the SensR approach (Yurochkin et al., 2019) in terms of MRD and MDRD for different levels of threshold (i.e., α, β). Note that for the left graph, we have fixed a specific $\alpha = 0.0005$. We observe that increasing the different levels of threshold results in increasing the level of unfairness. Also, for all threshold levels, our method outperforms the SenSR method. For the results in tables 8.1 and 8.2, we select a level of $\alpha = 0.0005$ and $\beta = 0.001$.



(a) MRD

(b) MDRD

Figure D.1: Impact of α and β (Adult UCI data set)

D.2 | Details on Artificial Datasets

Additionally to the 6 real-world datasets, we consider a synthetic scenario, that allows us to perform a further analysis of the relative performances of the approaches. The synthetic scenario subject is a pricing algorithm for a fictional car insurance policy. The objective of this scenario is to achieve a counterfactual fair predictor which estimates the average cost history of insurance customers. We suppose 5 unobserved variables (Aggressiveness, Inattention, Restlessness, Reckless and Overreaction) which corresponds to a 5 dimensional confounder U. The input X is composed of four explicit variables: vehicle age, speed average, horsepower and average kilometers per year. We consider the policyholder's age as sensitive attribute A. The input X and the average cost variable Y are sampled from U and A as depicted in figure 1 from the main paper. We propose both a binary and a continuous version of this scenario. For both of them, 5000 individuals are sampled.

We report below details on the distributions for the discrete setting (binary A and Y) of the synthetic scenario:

$$U \sim \mathcal{N}\left[\begin{pmatrix}0\\0.5\\1\\1.5\\2\end{pmatrix}, \begin{pmatrix}1&0&0&0&0\\0&4&0&0&0\\0&0&2&0&0\\0&0&0&3&0\\0&0&0&0&2\end{pmatrix}\right]$$

$$X1 \sim \mathcal{N}(10 + 2 * A_i + U_1 + U_2 + U_3, 1);$$
$$X2 \sim \mathcal{N}(100 + 3 * A + U_2^2, 10);$$
$$X3 \sim \mathcal{N}(200 + 4 * A + 5 * U_3, 20);$$
$$X4 \sim \mathcal{N}((10^4 + 5 * A + U_4 + U_5, 1000)$$

$$X \sim [X1, X2, X3, X4];$$
$$A \sim \mathcal{B}ernouilli(0.5);$$
$$Y \sim \mathcal{B}ernouilli(sigmoid(5 * (70 * A + 20 * \sum_j U_j)))$$

References

Tameem Adel, Isabel Valera, Zoubin Ghahramani, and Adrian Weller. One-network adversarial fairness. In *AAAI'19*, volume 33, pages 2412–2420, 2019.

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *ICML'18*, pages 60–69, 2018.

Sina Aghaei, Andrés Gómez, and Phebe Vayanos. Strong optimal classification trees. *arXiv preprint arXiv:2103.15965*, 2021.

Guilherme Alves, Fabien Bernier, Miguel Couceiro, Karima Makhlouf, Catuscia Palamidessi, and Sami Zhioua. Survey on Fairness Notions and Related Tensions. working paper or preprint, December 2021. URL https://hal.archives-ouvertes.fr/hal-03484009.

Mohsan Alvi, Andrew Zisserman, and Christoffer Nellåker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. ProPublica, May 23, 2016, 2016.

Shahab Asoodeh, Fady Alajaji, and Tamás Linder. On maximal correlation, mutual information and data privacy. In *2015 IEEE 14th Canadian Workshop on Information Theory (CWIT)*, pages 27–31. IEEE, 2015.

Francis R Bach and Michael I Jordan. Kernel independent component analysis. *Journal of machine learning research*, 3(Jul):1–48, 2002.

Sina Baharlouei, Maher Nouiehed, and Meisam Razaviyayn. Rényi fair inference. *CoRR*, abs/1906.12005, 2019.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. *Nips tutorial*, 1:2, 2017.

James Barrett and Daniel Lampard. An expansion for some second-order probability distributions and its application to noise problems. *IRE Transactions on Information Theory*, 1(1): 10–15, 1955. doi: 10.1109/TIT.1955.1055122.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: Mutual information neural estimation, 2018.

Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John T. Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *CoRR*, abs/1810.01943, 2018.

G Beraud-Sudreau. Construction d'un zonier en mrh à l'aide d'outils de data-science. *Institut des Actuaires*, 2, 2017.

H.J. Bierens. The Nadaraya-Watson Kernel regression function estimator. Serie Research Memoranda 0058, VU University Amsterdam, Faculty of Economics, Business Administration and Econometrics, 1988. URL https://ideas.repec.org/p/vua/wpaper/1988-58.html.

Christopher Blier-Wong, Jean-Thomas Baillargeon, Hélène Cossette, Luc Lamontagne, and Etienne Marceau. Rethinking representations in p&c actuarial science with deep neural networks. *arXiv*, 2102.05784, 2021a.

Christopher Blier-Wong, Hélène Cossette, Luc Lamontagne, and Etienne Marceau. Geographic ratemaking with spatial embeddings. *ASTIN Bulletin: The Journal of the IAA*, pages 1–31, 2021b.

John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 120–128, 2006.

Walter Block, Nicholas Snow, and Edward Stringham. Banks, insurance companies, and discrimination 1. *Business and Society Review*, 113(3):403–419, 2008.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357, 2016.

Login Nikolaevich Bol'shev and M Mirvaliev. Chi square goodness-of-fit test for the poisson, binomial and negative binomial distributions. *Theory of Probability & Its Applications*, 23(3): 461–474, 1979.

Michael Boskov and Richard Verrall. Premium rating by geographic area using spatial models. *ASTIN Bulletin*, 24(1):131–143, 1994. doi: 10.2143/AST.24.1.2005085.

B Le Boucher. Tarification iard et open data. *Institut des Actuaires*, 2, 2016.

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. *arXiv:1511.06349*, 2015.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

M. Broniatowski and S. Leorato. An estimation method for the neyman chi-square divergence with application to test of hypotheses. *J. Multivar. Anal.*, 97(6):1409–1436, July 2006. ISSN 0047-259X. doi: 10.1016/j.jmva.2006.02.001.

Andreas Buja. Remarks on Functional Canonical Variates, Alternating Least Squares Methods and Ace. *The Annals of Statistics*, 18(3):1032 – 1069, 1990. doi: 10.1214/aos/1176347739.

Alessandro Castelnovo, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, 12(1):1–21, 2022.

L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 319–328, 2019.

Arthur Charpentier. *Computational actuarial science with R*. CRC press, 2014.

Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 339–348, 2019.

Xi Chen, Diederik P Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. Variational lossy autoencoder. *arXiv preprint arXiv:1611.02731*, 2016.

Silvia Chiappa. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7801–7808, 2019.

Alexandre Corradin, Michel Denuit, Marcin Detyniecki, Vincent Grari, Matteo Sammarco, and Julien Trufin. Joint modeling of claim frequencies and behavioral signals in motor insurance. *ASTIN Bulletin: The Journal of the IAA*, pages 1–22, 2022.

Amanda Coston, Karthikeyan Natesan Ramamurthy, Dennis Wei, Kush R Varshney, Skyler Speakman, Zairah Mustahsan, and Supriyo Chakraborty. Fair transfer learning with missing protected attributes. In *AAAI*, 2019.

Rachel Cummings, Varun Gupta, Dhamma Kimpara, and Jamie Morgenstern. On the compatibility of privacy and fairness. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, pages 309–315, 2019.

Dwork Cynthia. Differential privacy. *International Colloquium on Automata, Languages, and Programming*, pages 1–12, 2006.

Hal Daume III and Daniel Marcu. Domain adaptation for statistical classifiers. *Journal of artificial Intelligence research*, 26:101–126, 2006.

Jacques Dauxois and Guy Martial Nkiet. Nonlinear canonical analysis and independence tests. *The Annals of Statistics*, 26(4):1254 – 1278, 1998. doi: 10.1214/aos/1024691242.

Virginie Do, Sam Corbett-Davies, Jamal Atif, and Nicolas Usunier. Two-sided fairness in rankings via lorenz dominance. *Advances in Neural Information Processing Systems*, 34, 2021.

Chris Dolman, Edward Jed Frees, and Fei Huang. Multidisciplinary collaboration on discrimination–not just "nice to have". *Annals of Actuarial Science*, 15(3):485–487, 2021.

M. Donini, S. Ben-David, M. Pontil, and J. Shawe-Taylor. An efficient method to impose fairness in linear models. In *NIPS Workshop on Prioritising Online Content*, 2017.

Flávio du Pin Calmon, Dennis Wei, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. Optimized data pre-processing for discrimination prevention. *CoRR*, abs/1704.03354, 2017.

Dheeru Dua and Casey Graff. UCI ml repository. http://archive.ics.uci.edu/ml, 2017.

Christophe Dutang and Arthur Charpentier. Casdatasets r package vignette. *Reference Manual, November*, 13(2019):1–0, 2019.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *ITCS'12*, pages 214–226, 2012.

Paul Embrechts and Mario V Wüthrich. Recent challenges in actuarial science. *Annual Review of Statistics and Its Application*, 9, 2022.

Vitalii Emelianov, George Arvanitakis, Nicolas Gast, Krishna P. Gummadi, and Patrick Loiseau. The price of local fairness in multistage selection. In Sarit Kraus, editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5836–5842. ijcai.org, 2019. doi: 10.24963/ijcai.2019/809. URL https://doi.org/10.24963/ijcai.2019/809.

European Commission. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance), 2016. URL https://eur-lex.europa.eu/eli/reg/2016/679/oj.

Ludwig Fahrmeir, Stefan Lang, and Friedemann Spies. Generalized geoadditive models for insurance claims data. *Blätter der DGVFM*, 26(1):7–23, 2003.

Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In Longbing Cao, Chengqi Zhang, Thorsten Joachims, Geoffrey I. Webb, Dragos D. Margineantu, and Graham Williams, editors, *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*, pages 259–268. ACM, 2015. doi: 10.1145/2783258.2783311. URL https://doi.org/10.1145/2783258.2783311.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.*, 20(177):1–81, 2019.

Edward W Frees and Fei Huang. The discriminating (pricing) actuary. *North American Actuarial Journal*, pages 1–23, 2021.

Edward W Frees and Gee Lee. Rating endorsements using generalized linear models. *Variance*, 10(1):51–74, 2015.

Batya Friedman and Helen Nissenbaum. Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, 14(3):330–347, 1996.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

Kazuto Fukuchi and Jun Sakuma. Fairness-aware learning with restriction of universal dependency using f-divergences. *arXiv preprint arXiv:1506.07721*, 2015.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. *arXiv preprint arXiv:1409.7495*, 2014.

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

Hans Gebelein. Das statistische problem der korrelation als variations- und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung. *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik*, 21(6):364–379, 1941. doi: https://doi.org/10.1002/zamm.19410210604.

Sara A Geer and Sara van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

Stephen Gillen, Christopher Jung, Michael Kearns, and Aaron Roth. Online learning with an unknown fairness metric. In *Advances in neural information processing systems*, pages 2600–2609, 2018.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Ian Goodfellow, Patrick McDaniel, and Nicolas Papernot. Making machine learning robust against adversarial inputs. *Communications of the ACM*, 61(7):56—66, 2018. ISSN 0001-0782. doi: 10.1145/3134599.

Priya Goyal, Adriana Romero Soriano, Caner Hazirbas, Levent Sagun, and Nicolas Usunier. Fairness indicators for systematic assessments of visual feature extractors. *arXiv preprint arXiv:2202.07603*, 2022.

Vincent Grari, Boris Ruf, Sylvain Lamprier, and Marcin Detyniecki. Fair adversarial gradient tree boosting. In Jianyong Wang, Kyuseok Shim, and Xindong Wu, editors, *2019 IEEE International Conference on Data Mining, ICDM 2019, Beijing, China, November 8-11, 2019*, pages 1060–1065. IEEE, 2019. doi: 10.1109/ICDM.2019.00124. URL https://doi.org/10.1109/ICDM.2019.00124.

Vincent Grari, Sylvain Lamprier, and Marcin Detyniecki. Fairness-aware neural rényi minimization for continuous features. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 2262–2268. ijcai.org, 2020a. doi: 10.24963/ijcai.2020/313. URL https://doi.org/10.24963/ijcai.2020/313.

Vincent Grari, Sylvain Lamprier, and Marcin Detyniecki. Adversarial learning for counterfactual fairness. *arXiv preprint arXiv:2008.13122*, 2020b.

Vincent Grari, Boris Ruf, Sylvain Lamprier, and Marcin Detyniecki. Achieving fairness with decision trees: An adversarial approach. *Data Sci. Eng.*, 5(2):99–110, 2020c. doi: 10.1007/s41019-020-00124-2.

Vincent Grari, Oualid El Hajouji, Sylvain Lamprier, and Marcin Detyniecki. Enforcing individual fairness via rényi variational inference. In *International Conference on Neural Information Processing*, pages 608–616. Springer, 2021a. doi: 0.1007/978-3-030-92307-5_71. URL https://doi.org/10.1007/978-3-030-92307-5_71.

Vincent Grari, Oualid El Hajouji, Sylvain Lamprier, and Marcin Detyniecki. Learning unbiased representations via rényi minimization. In Nuria Oliver, Fernando Pérez-Cruz, Stefan Kramer, Jesse Read, and José Antonio Lozano, editors, *Machine Learning and Knowledge Discovery in Databases. Research Track - European Conference, ECML PKDD 2021, Bilbao,*

Spain, September 13-17, 2021, Proceedings, Part II, volume 12976 of *Lecture Notes in Computer Science*, pages 749–764. Springer, 2021b. doi: 10.1007/978-3-030-86520-7_46. URL https://doi.org/10.1007/978-3-030-86520-7_46.

Vincent Grari, Sylvain Lamprier, and Marcin Detyniecki. Fairness without the sensitive attribute via causal variational autoencoder. *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, (still waiting DOI)*, 2109.04999, 2021c.

Vincent Grari, Arthur Charpentier, Sylvain Lamprier, and Marcin Detyniecki. A fair pricing model via adversarial learning. *arXiv:2202.12008*, 2022.

Arthur Gretton, Ralf Herbrich, Alexander Smola, Olivier Bousquet, and Bernhard Schölkopf. Kernel methods for measuring independence. *J. Mach. Learn. Res.*, 6:2075–2129, December 2005. ISSN 1532-4435.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5), 2018.

Maya Gupta, Andrew Cotter, Mahdi Milani Fard, and Serena Wang. Proxy fairness. *arXiv preprint arXiv:1806.11212*, 2018.

David R Hardoon and John Shawe-Taylor. Convergence analysis of kernel canonical correlation analysis: theory and practice. *Machine learning*, 74(1):23–38, 2009.

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

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938. PMLR, 2018.

J Henry Hinnefeld, Peter Cooman, Nat Mammo, and Rupert Deese. Evaluating fairness metrics in the presence of dataset bias. *arXiv preprint arXiv:1809.09245*, 2018.

Hermann Otto Hirschfeld. A connection between correlation and contingency. *Mathematical Proceedings of the Cambridge Philosophical Society*, 31(4):520–524, 1935. doi: 10.1017/S0305004100013517.

Kurt Hornik, Maxwell Stinchcombe, Halbert White, et al. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

Charles Huff and Joel Cooper. Sex bias in educational software: The effect of designers' stereotypes on the software they design 1. *Journal of Applied Social Psychology*, 17(6):519–532, 1987.

Christina Ilvento. Metric learning for individual fairness. *arXiv preprint arXiv:1906.00250*, 2019.

Donald Jensen and Lawrence Mayer. Some Variational Results and Their Applications in Multiple Inference. *The Annals of Statistics*, 5(5):922 – 931, 1977. doi: 10.1214/aos/1176343948.

Christopher Jung, Michael Kearns, Seth Neel, Aaron Roth, Logan Stapleton, and Zhiwei Steven Wu. Eliciting and enforcing subjective individual fairness. *arXiv preprint arXiv:1905.10660*, 2019.

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, Oct 2012. ISSN 0219-3116. doi: 10.1007/s10115-011-0463-8.

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In Peter A. Flach, Tijl De Bie, and Nello Cristianini, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 35–50, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-33486-3.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *arXiv preprint arXiv:1711.05144*, 2017.

Niki Kilbertus, Adrià Gascón, Matt Kusner, Michael Veale, Krishna Gummadi, and Adrian Weller. Blind justice: Fairness with encrypted sensitive attributes. In *ICML*, 2018.

Niki Kilbertus, Philip J Ball, Matt J Kusner, Adrian Weller, and Ricardo Silva. The sensitivity of counterfactual fairness to unmeasured confounding. In *Uncertainty in artificial intelligence*, pages 616–626. PMLR, 2020.

Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Colored mnist dataset. https://github.com/feidfoe/learning-not-to-learn/tree/master/dataset/colored_mnist, 2019a.

Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9012–9020, 2019b.

Hyemi Kim, Seungjae Shin, JoonHo Jang, Kyungwoo Song, Weonyoung Joo, Wanmo Kang, and Il-Chul Moon. Counterfactual fairness with disentangled causal effect variational autoencoder. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third*

Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021, pages 8128–8136. AAAI Press, 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/16990.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Murat Kocaoglu, Christopher Snyder, Alexandros G. Dimakis, and Sriram Vishwanath. Causalgan: Learning causal implicit generative models with adversarial training. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=BJE-4xW0W.

Peter Kontschieder, Madalina Fiterau, Antonio Criminisi, and Samuel Rota Bulo. Deep neural decision forests. In *Proceedings of the IEEE international conference on computer vision*, pages 1467–1475, 2015.

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

L Krolikowski. Modélisation de l'élasticité au prix à la souscription en assurance automobile dans le cadre d'une optimisation tarifaire. *Institut des Actuaires*, 2, 2021.

Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4066–4076, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html.

Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed H Chi. Fairness without demographics through adversarially reweighted learning. *arXiv preprint arXiv:2006.13114*, 2020.

Romain F Laine, Guillaume Jacquemet, and Alexander Krull. Imaging in focus: An introduction to denoising bioimages in the era of deep learning. *The international journal of biochemistry & cell biology*, 140:106077, 2021.

Anja Lambrecht and Catherine E. Tucker. Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads. *SSRN Electronic Journal*, 2016. doi: 10.2139/ssrn.2852260.

Henry Oliver Lancaster. The Structure of Bivariate Distributions. *The Annals of Mathematical Statistics*, 29(3):719 – 736, 1958. doi: 10.1214/aoms/1177706532.

Yann LeCun, Corinna Cortes, and Chris Burges. Mnist handwritten digit database, 2010.

Joshua Ka-Wing Lee. *Maximal Correlation Feature Selection and Suppression With Applications*. PhD thesis, Massachusetts Institute of Technology, 2021.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017.

Xiaoxiao Li, Ziteng Cui, Yifan Wu, Lin Gu, and Tatsuya Harada. Estimating and improving fairness with adversarial learning. *arXiv preprint arXiv:2103.04243*, 2021.

Mathias Lindholm, Ronald Richman, Andreas Tsanakas, and Mario V Wüthrich. Discrimination-free insurance pricing. *ASTIN Bulletin: The Journal of the IAA*, 52(1):55–89, 2022.

David Lopez-Paz, Philipp Hennig, and Bernhard Schölkopf. The randomized dependence coefficient. In *Advances in neural information processing systems*, pages 1–9, 2013.

Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pages 6446–6456, 2017.

Gilles Louppe, Michael Kagan, and Kyle Cranmer. Learning to pivot with adversarial networks. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 981–990, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/48ab2f9b45957ab574cf005eb8a76760-Abstract.html.

H. Läuter. Silverman, b. w.: Density estimation for statistics and data analysis. chapman & hall, london – new york 1986, 175 pp., £12.—. *Biometrical Journal*, 30(7):876–877, 1988. doi: 10.1002/bimj.4710300745.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3384–3393. PMLR, 2018.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Fairness through causal awareness: Learning causal latent-variable models for biased data. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 349–358, 2019.

Jeremie Mary, Clément Calauzènes, and Noureddine El Karoui. Fairness-aware learning for continuous attributes and treatments. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4382–4391, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/mary19a.html.

Julien Mathis. Elaboration d'un zonier en assurance de véhicules par des méthodes de lissage spatial basées sur des simulations mcmc. *Institut des Actuaires*, 2, 2007.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6): 1–35, 2021.

Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pages 263–275. IEEE, 2017.

Alan Mishler and Niccolò Dalmasso. Fair when trained, unfair when deployed: Observable fairness measures are unstable in performative prediction settings. *arXiv preprint arXiv:2202.05049*, 2022.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *International Conference on Machine Learning*, 2019.

James H Moor. What is computer ethics? *Metaphilosophy*, 16(4):266–275, 1985.

Sérgio Moro, Paulo Cortez, and Paulo Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62, 06 2014.

Daniel Moyer, Shuyang Gao, Rob Brekelmans, Aram Galstyan, and Greg Ver Steeg. Invariant representations without adversarial training. In *Advances in Neural Information Processing Systems*, pages 9084–9093, 2018.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pages 10–18, 2013.

Debarghya Mukherjee, Mikhail Yurochkin, Moulinath Banerjee, and Yuekai Sun. Two simple ways to learn individual fairness metrics from data. *arXiv preprint arXiv:2006.11439*, 2020.

Johannes Paefgen, Thorsten Staake, and Frédéric Thiesse. Evaluation and aggregation of pay-as-you-drive insurance rate factors: A classification analysis approach. *Decision Support Systems*, 56, 2013.

Judea Pearl et al. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.

Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08*, page 560, 2008. ISSN 0309-0167 (Print). doi: 10.1145/1401890.1401959.

Stephen Pfohl, Tony Duan, Daisy Yi Ding, and Nigam H Shah. Counterfactual reasoning for fair clinical risk prediction. *arXiv preprint arXiv:1907.06260*, 2019.

Barnabás Póczos, Zoubin Ghahramani, and Jeff Schneider. Copula-based kernel dependency measures. In *Proceedings of the 29th International Conference on International Conference on Machine Learning*, ICML'12, pages 1635–1642, USA, 2012. Omnipress. ISBN 978-1-4503-1285-1.

Anya E.R. Prince and Daniel Schwarcz. Proxy discrimination in the age of artificial intelligence and big data. *Iowa Law Review*, 105:1257, 2019.

David Pujol, Ryan McKenna, Satya Kuppam, Michael Hay, Ashwin Machanavajjhala, and Gerome Miklau. Fair decision making using privacy-protected data. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 189–199, 2020.

Ruggero Ragonesi, Riccardo Volpi, Jacopo Cavazza, and Vittorio Murino. Learning unbiased representations via mutual information backpropagation. *arXiv preprint arXiv:2003.06430*, 2020.

Srinivasan Ravichandran, Drona Khurana, Bharath Venkatesh, and Narayanan Unny Edakunni. Fairxgboost: Fairness-aware classification in xgboost. *arXiv preprint arXiv:2009.01442*, 2020.

Alfréd Rényi. On measures of dependence. *Acta mathematica hungarica*, 10(3-4):441–451, 1959.

Luc Rocher, Julien M Hendrickx, and Yves-Alexandre De Montjoye. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature communications*, 10(1):1–9, 2019.

Chris Russell, Matt J Kusner, Joshua Loftus, and Ricardo Silva. When worlds collide: integrating different counterfactual assumptions in fairness. In *Advances in Neural Information Processing Systems*, pages 6414–6423, 2017.

S. Said. Refonte des modèles de tarification de lássurance automobile et création de zoniers tarifaires. *Institut des Actuaires*, 2, 2016.

Marco Scarsini. On measures of concordance. *Stochastica*, 8(3):201–218, 1984.

Hato Schmeiser, Tina Störmer, and Joël Wagner. Unisex insurance pricing: Consumers' perception and market implications. *The Geneva Papers on Risk and Insurance - Issues and Practice*, 39(2):322–350, 2014.

Candice Schumann, Xuezhi Wang, Alex Beutel, Jilin Chen, Hai Qian, and Ed H Chi. Transfer of machine learning fairness across domains. *arXiv preprint arXiv:1906.09688*, 2019.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *ICML'17*, pages 3076–3085, 2017.

Peng Shi and Kun Shi. Nonlife insurance risk classification using categorical embedding. *SSRN*, 3777526, 2021.

Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. In *NIPS'16*, pages 3738–3746, 2016.

Harini Suresh and John V Guttag. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, 2, 2019.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2014.

Gábor J Székely, Maria L Rizzo, et al. Brownian distance covariance. *The annals of applied statistics*, 3(4):1236–1265, 2009.

Tsubasa Tamba, Hirokazu Odaka, Aya Bamba, Hiroshi Murakami, Koji Mori, Kiyoshi Hayashida, Yukikatsu Terada, Tsunefumi Mizuno, and Masayoshi Nobukawa. Simulation-based spectral analysis of x-ray ccd data affected by photon pile-up. *Publications of the Astronomical Society of Japan*, 74(2):364–383, 2022.

Greg Taylor. Use of spline functions for premium rating by geographic area. *ASTIN Bulletin*, 19(1):91–122, 1989. doi: 10.2143/AST.19.1.2014917.

Stan Developent Team et al. Rstan: the r interface to stan. *R package version*, 2(1), 2016.

The French Institute of Actuaries. Pricing game 2017. ttps://actinfo.hypotheses.org/86, 2017. Online; accessed 22 August 2021.

The Institute of Actuaries of France. Pricing game 2015. https://freakonometrics.hypotheses.org/20191, 2015. Online; accessed 14 August 2019.

Guy Thomas. Non-risk price discrimination in insurance: market outcomes and public policy. *The Geneva Papers on Risk and Insurance-Issues and Practice*, 37(1):27–46, 2012.

Dag Tjøstheim, Håkon Otneim, and Bård Støve. Statistical dependence: Beyond pearson's ρ. *Statistical Science*, 37(1):90–109, 2022.

Nenad Tomasev, Kevin R McKee, Jackie Kay, and Shakir Mohamed. Fairness for unobserved characteristics: Insights from technological impacts on queer communities. *arXiv preprint arXiv:2102.04257*, 2021.

Britta N Torgrimson and Christopher T Minson. Sex and gender: what is the difference?, 2005.

Oskar Tufvesson, Johan Lindström, and Erik Lindström. Spatial statistical modelling of insurance risk: a spatial epidemiological approach to car insurance. *Scandinavian Actuarial Journal*, 2019(6):508–522, 2019. doi: 10.1080/03461238.2019.1576146.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7167–7176, 2017.

US Census Bureau. Us census demographic data. https://data.census.gov/cedsci/, 2019. Online; accessed 03 April 2019.

Michael Veale and Reuben Binns. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, 4(2):2053951717743530, 2017.

Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*, pages 1–7. IEEE, 2018.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Why fairness cannot be automated: Bridging the gap between eu non-discrimination law and ai. *Computer Law & Security Review*, 41:105567, 2021.

Christina Wadsworth, Francesca Vera, and Chris Piech. Achieving fairness through adversarial learning: an application to recidivism prediction. *CoRR*, abs/1807.00199, 2018.

Felix Wagner. Nonlinear pile-up separation with lstm neural networks for cryogenic particle detectors. *arXiv preprint arXiv:2112.06792*, 2021.

Chun Wang, Elizabeth D Schifano, and Jun Yan. Geographical ratings with spatial random effects in a two-part model. *Variance*, 13(1):20, 2017.

Zhuo Wang, Wei Zhang, Ning Liu, and Jianyong Wang. Transparent classification with multilayer logical perceptrons and random binarization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

Hans S Witsenhausen. On sequences of pairs of dependent random variables. *SIAM Journal on Applied Mathematics*, 28(1):100–113, 1975.

Mario V. Wuthrich and Michael Merz. Statistical foundations of actuarial learning and its applications. *SSRN*, 3822407, 2021.

Depeng Xu, Yongkai Wu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Achieving causal fairness through generative adversarial networks. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019.

Longfei Yan, W Bastiaan Kleijn, and Thushara Abhayapala. A linear-time independence criterion based on a finite basis approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 202–212. PMLR, 2020a.

Shen Yan, Hsien-te Kao, and Emilio Ferrara. Fair class balancing: enhancing model fairness without observing sensitive attributes. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020b.

I-Cheng Yeh and Che-hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Syst. Appl.*, 36(2):2473–2480, March 2009. ISSN 0957-4174. doi: 10.1016/j.eswa.2007.12.020.

Mikhail Yurochkin and Yuekai Sun. Sensei: Sensitive set invariance for enforcing individual fairness. *arXiv preprint arXiv:2006.14168*, 2020.

Mikhail Yurochkin, Amanda Bower, and Yuekai Sun. Training individually fair ml models with sensitive subspace robustness. In *International Conference on Learning Representations*, 2019.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In Rick Barrett, Rick Cummings, Eugene Agichtein, and Evgeniy Gabrilovich, editors, *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*, pages 1171–1180. ACM, 2017a. doi: 10.1145/3038912.3052660. URL https://doi.org/10.1145/3038912.3052660.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, Krishna P. Gummadi, and Adrian Weller. From parity to preference-based notions of fairness in classification. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 229–239, 2017b. URL https://proceedings.neurips.cc/paper/2017/hash/82161242827b703e6acf9c726942a1e4-Abstract.html.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P. Gummadi. Fairness Constraints: Mechanisms for Fair Classification. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 962–970, Fort Lauderdale, FL, USA, 20–22 Apr 2017c. PMLR. URL http://proceedings.mlr.press/v54/zafar17a.html.

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

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *AAAI'18*, pages 335–340, 2018.

Chongsheng Zhang, Changchang Liu, Xiangliang Zhang, and George Almpanidis. An up-to-date comparison of state-of-the-art classification algorithms. *Expert Systems with Applications*, 82, 04 2017. doi: 10.1016/j.eswa.2017.04.003.

Shengjia Zhao, Jiaming Song, and Stefano Ermon. Infovae: Information maximizing variational autoencoders. *arXiv preprint arXiv:1706.02262*, 2017.

Tianxiang Zhao, Enyan Dai, Kai Shu, and Suhang Wang. You can still achieve fairness without sensitive attributes: Exploring biases in non-sensitive features. *arXiv preprint arXiv:2104.14537*, 2021.

Indre Zliobaite. A survey on measuring indirect discrimination in machine learning. *arXiv: 1511.00148*, 2015.