Input-agnostic Certified Group Fairness via Gaussian Parameter Smoothing

Jiayin Jin[1]   **Zeru Zhang**[1]   **Yang Zhou**[1]   **Lingfei Wu**[2]

Abstract

Only recently, researchers attempt to provide classification algorithms with provable group fairness guarantees. Most of these algorithms suffer from harassment caused by the requirement that the training and deployment data follow the same distribution. This paper proposes an input-agnostic certified group fairness algorithm, FAIRSMOOTH, for improving the fairness of classification models while maintaining the remarkable prediction accuracy. A Gaussian parameter smoothing method is developed to transform base classifiers into their smooth versions. An optimal individual smooth classifier is learnt for each group with only the data regarding the group and an overall smooth classifier for all groups is generated by averaging the parameters of all the individual smooth ones. By leveraging the theory of nonlinear functional analysis, the smooth classifiers are reformulated as output functions of a Nemytskii operator. Theoretical analysis is conducted to derive that the Nemytskii operator is smooth and induces a Fréchet differentiable smooth manifold. We theoretically demonstrate that the smooth manifold has a global Lipschitz constant that is independent of the domain of the input data, which derives the input-agnostic certified group fairness.

1. Introduction

Machine learning algorithms that are deployed in human-facing applications (e.g., loan granting (Dedman, 1988), criminal recidivism (Northpointe, 2012), job screening (Miller, 2015), and predictive policing (Hvistendahl, 2016)) are often faced with unfair prediction results by discriminating against particular groups. Developing fair classification algorithms with respect to sensitive attributes has be-

come an important problem due to the growing importance of addressing social biases in machine learning (Celis et al., 2019). A fair classification model aims to build a classifier that not only makes accurate predictions but also prevents itself from acting against specific groups in a discriminatory way, i.e., the classifier should have similar performance for all groups (Iofinova et al., 2021).

To minimize discrimination effects caused by the classification models, a large number of fair classification techniques have been proposed to utilize various notions of group fairness to develop tailored classification algorithms that predict in the same way across different demographic groups (Zafar et al., 2017a;b; Goel et al., 2018; Krasanakis et al., 2018; Menon & Williamson, 2018; Mandal et al., 2020; Wang et al., 2020; Yang et al., 2020; Zhang et al., 2020a; Lahoti et al., 2020; Ben-Porat et al., 2021; Celis et al., 2021a; Gorantla et al., 2021; Chuang & Mroueh, 2021; Roh et al., 2021a; Shekhar et al., 2021; Qian et al., 2021; Liu & Wang, 2021; Maity et al., 2021; Celis et al., 2021b; Du et al., 2021; Roh et al., 2021a; Anonymous, 2022a; Zhu et al., 2022). Traditional group fairness techniques often follow manually-crafted heuristics to generate fair classifiers by solving non-convex optimization problems (Zafar et al., 2017a;b; Krasanakis et al., 2018). Such heuristic strategies plus non-convex optimization are hard to provide provable guarantees for group fairness certification.

Only recently, few studies have sought to provide provable group fairness guarantees for classification algorithms (Feldman et al., 2015; McNamara et al., 2017; Celis et al., 2019; Ustun et al., 2019; Gupta et al., 2021; Krishnaswamy et al., 2021; Segal et al., 2021; Celis et al., 2021a; Iofinova et al., 2021; Anonymous, 2022b). Despite achieving remarkable success, most of the above fairness guarantees rely on the assumption that the data the model being trained with and the data encountered after deployment follow the same distribution. However, this assumption is false for many real-world problems (Zhuang et al., 2021). It is demonstrated that these models often violate the fairness guarantees and exhibit unfair bias when evaluated on data from a different distribution (Anonymous, 2022b). Consider an example of job screening, applicants' demographics can shift over time, which results in the change of the distribution of applicants between training and deployment datasets. As a result, the derived fairness guarantees based on the training

data will not hold on the deployment data in practice, if the learning algorithms assume that the training and deployment distributions are the same. Shifty is the first classification strategy that provides high-confidence guarantees based on user-specified fairness constraints under demographic shift between training and deployment (Anonymous, 2022b). However, Shifty needs to know the old and new demographic proportions regarding the demographic shift at training, or the demographic proportions are bounded in known intervals in Shifty, which limits the applicability of Shifty in real-world scenarios.

To our best knowledge, this work is the first to certify the group fairness of classifiers with theoretical input-agnostic guarantees, while there is no need to know the shift between training and deployment datasets with respect to sensitive attributes, by leveraging the theory of nonlinear functional analysis, including Nemytskii operator and Fréchet differentiable smooth manifolds.

Randomized smoothing has achieved the state-of-the-art certified robustness guarantees against worst-case attacks by smoothing with isotropic Gaussian distribution (Lécuyer et al., 2019; Cohen et al., 2019; Lee et al., 2019; Li et al., 2019; Salman et al., 2019; Levine & Feizi, 2020; Zhai et al., 2020; Feng et al., 2020; Dvijotham et al., 2020; Hayes, 2020). One advantage of the randomized smoothing techniques is that they are agnostic to input data as well as network architectures (Cohen et al., 2019; Mohapatra et al., 2020; Blum et al., 2020; ?). BeFair presents a new notion of group fairness: best-effort fairness, in which the performance of an overall fair classifier $f(x)$ for all groups on each group G_k ($k = 1, 2, \cdots, K$) should be close to the optimal individual classifier $f_k(x)$ for each G_k (f_k is trained with only the data $x \in \mathbb{R}^n$ regarding G_k) (Krishnaswamy et al., 2021).

This motivates us to propose an ϵ-fairness metric that measures the gap between the predictions achieved by the overall and individual classifiers and to develop a Gaussian parameter smoothing framework for the certification of the input-agnostic group fairness.

First, a Gaussian parameter smoothing method is proposed to transform the base individual and overall classifiers $f(x; W_k)$ and $f(x; W)$ into their smooth versions below.

$$\hat{f}(x; W_k) = \mathbb{E}_{\Delta}\big(f(x; W_k + \Delta)\big), \text{ and}$$
$$\hat{f}(x; W) = \mathbb{E}_{\Delta}\big(f(x; W + \Delta)\big), \Delta \sim \mathcal{N}(0, \sigma^2 I) \tag{1}$$

where $\mathcal{N}(0, \sigma^2 I)$ is the isotropic Gaussian distribution. W_k and W are the parameters of $f(x; W_k)$ and $f(x; W)$.

Second, we train an optimal individual smooth classifier $\hat{f}(x; W_k^*)$ for each group G_k with only the data regarding G_k, where W_k^* is the parameter of $\hat{f}(x; W_k^*)$. An overall

classifier $\hat{f}(x; W^*)$ for all groups is generated by averaging the parameters of all the individual classifiers, i.e., $W^* = \frac{W_1^* + \cdots + W_K^*}{K}$. Ideally, when the parameters W_k^* of $\hat{f}(x; W_k^*)$ are close to each other, then the overall classifier $\hat{f}(x; W^*)$ with the parameter W^* is relatively fair to all groups.

Third, by leveraging the theory of nonlinear functional analysis, the smooth classifiers $\hat{f}(x; W)$ and $\hat{f}(x; W_k)$ are reformulated as as output functions of a Nemytskii operator $\hat{N}(W)(\cdot)$, which maps the parameter space to a function space. The theoretical analysis is conducted to derive that the Nemytskii operator $\hat{N}(W)(\cdot)$ is smooth and induces a Fréchet differentiable smooth manifold $\hat{N}(\mathbb{R}^m)$. We further prove that the smooth manifold has a global Lipschitz constant $\frac{1}{\sqrt{2\pi}\sigma}$, i.e. $\|\hat{N}(W_1)(x) - \hat{N}(W_2)(x)\| \leq \frac{\|W_1 - W_2\|_2}{\sqrt{2\pi}\sigma}$ for any $W_1, W_2 \in \mathbb{R}^m$. Moreover, the global Lipschitz constant $\frac{1}{\sqrt{2\pi}\sigma}$ is independent of the structure of classifiers (e.g., neural networks) and the domain of the input data. This global Lipschitz property of $\hat{N}(W)$ allows to measure the difference between two smoothed classifiers $\hat{f}(x; W_1)$ and $\hat{f}(x; W_2)$ with the distance $\|W_1 - W_2\|_2$ between their parameters. This is the reason why we perform the Gaussian parameter smoothing on classifiers. On the other hand, the base classifiers $f(x; W_k)$ and $f(x; W)$ are only locally Lipschitz regarding W_k and W respectively. Thus, the Lipschitz constants depend on the input data x.

Last but not least, with the global Lipschitz property of $\hat{N}(W)$, our training aims to minimize the training loss of each individual smooth classifier $\hat{N}(W_k)(x)$ for group G_k as well as the distance among the parameters of individual smooth classifiers, where a smaller distance indicates a better group fairness. We demonstrate the input-agnostic certified group fairness, i.e., the overall smooth classifier $\hat{N}(W^*)(x)$ has certified $\frac{(K-1)d}{\sqrt{2\pi}K\sigma}$-fairness, where $d = \max_{1 \leq i, j \leq K} \|W_i^* - W_j^*\|$. Namely, for $k = 1, \cdots, K$, it holds $|\hat{N}(W^*)(x) - \hat{N}(W_k^*)(x)| \leq \frac{(K-1)d}{\sqrt{2\pi}K\sigma}$ for any input $x \in G_k$. The certified $\frac{(K-1)d}{\sqrt{2\pi}K\sigma}$-fairness of $\hat{N}(W^*)(x)$ guarantees that the overall smooth classifier $\hat{N}(W^*)(x)$ is close to each optimal individual smooth one $\hat{N}(W_k^*)(x)$ within distance $\frac{(K-1)d}{\sqrt{2\pi}K\sigma}$.

In comparison with existing group fairness techniques with provable guarantees, our Gaussian parameter smoothing method based on the theory of nonlinear functional analysis exhibits two compelling advantages: (1) It is agnostic to the input data without the knowledge or assumption about the distribution of training and deployment data; (2) Producing the overall fair classifier from the optimal individual classifier for each group is able to recover the most accurate model for each group in an ideal setting, which is often ignored by other fair classification algorithms.

Empirical evaluation on real datasets demonstrates the superior performance of our Gaussian parameter smoothing model against several state-of-the-art group fairness techniques with provable guarantees. In addition, more experiments, implementation details, and hyperparameter selection and setting are presented in Appendices A.2-A.3.

2. Preliminaries

2.1. Problem Statement

Given a dataset $D = X \times Z \times Y$ where X, Z, and Y represent the feature, group attribute, and label spaces respectively. Each sample (x, z, y) is drawn from D where x is an n-dimensional feature vector $x = [x_1, \ldots, x_n] \in \mathbb{R}^n$, z is one of K sensitive attribute values that are relevant for fairness (e.g., gender, age, or race), i.e., $z \in Z = \{z_1, \cdots, z_K\}$, and y is a class label $y \in Y = \{0, 1\}$. Without loss of generality, we discuss the case where there is only one sensitive attribute Z in the paper. Our input-agnostic certified group fairness algorithm can be easily extended to the case of multiple sensitive attributes. Based on K sensitive attribute values, the dataset D is divided into K non-overlapping groups $G_k = \{(x, z, y) \in D | z = z_k\}$ and $G_k \cap G_j = \emptyset$ for all $1 \le j \ne k \le K$.

Let $f(x; W_k)$ denote a deterministic classifier for group G_k trained with only the data regarding G_k. Our Gaussian parameter smoothing method converts each individual base classifier $f(x; W_k)$ into an individual smooth one $\hat{f}(x; W_k)$.

$$\hat{f}(x; W_k) = \mathbb{E}_{\Delta}(f(x; W_k + \Delta)), \ \Delta \sim \mathcal{N}(0, \sigma^2 I) \quad (2)$$

where $\mathcal{N}(0, \sigma^2 I)$ is the isotropic Gaussian distribution. The training objective of classification is to make the parameters of all individual smooth classifiers close to each other, while achieving the optimal prediction accuracy by each $\hat{f}(x; W_k)$.

$$\min_{W_1, \cdots, W_K} \sum_{k=1}^{K} \mathbb{E}_{(x_i, z_i, y_i) \in G_k} \left(\mathcal{L}(\hat{f}(x_i; W_k), y_i) \right) + \\ \alpha \sum_{k=1}^{K} \sum_{l>k}^{K} \|W_k - W_l\|_2^2 \quad (3)$$

where $\mathcal{L}$ denotes the loss function of classification, e.g., cross-entropy. The second term is a parameter disparity term that forces all individual classifiers to approach each other, in order to alleviate the unfairness issue.

An ϵ-fairness metric will be introduced to measure the gap between the predictions achieved by the overall smooth classifier $\hat{f}(x; W^*)$ and the individual smooth ones $\hat{f}(x; W_k^*)$, where the parameter W^* of $\hat{f}(x; W^*)$ is generated by averaging the parameters of all $\hat{f}(x; W_k^*)$, i.e., $W^* = \frac{W_1^* + \cdots + W_K^*}{K}$.

In an ideal scenario where enough training data are given for each group, the training of individual smooth classifiers

would produce the accurate prediction for each group as well as achieve the remarkable fairness among different groups, which will be certified with the fairness guarantees.

2.2. Theory of Nonlinear Functional Analysis

In mathematics, nonlinear functional analysis aims to study nonlinear mappings (i.e., nonlinear operators) between infinite-dimensional vector spaces and certain classes of nonlinear spaces and their mappings (Riesz & Sz.-Nagy, 1955).

A Cauchy sequence is a sequence whose elements become arbitrarily close to each other as the sequence progresses (Lang, 1993).

Definition 1 (Cauchy Sequence). *Let $(X, \| \cdot \|)$ be a metric space. A sequence $\{x_i\}_{i=1}^{\infty} \subset X$ is called a Cauchy sequence if for any $\varepsilon > 0$, there exists an integer $I(\varepsilon) > 0$, such that for any integers $i, j > I(\varepsilon)$, it holds $\|x_i - x_j\| < \varepsilon$.*

A Banach space is a vector space with a metric that allows the computation of vector magnitude and distance between vectors and is complete in the sense that a Cauchy sequence of vectors always converges to a well-defined limit that is within the space (Riesz & Sz.-Nagy, 1955).

Definition 2 (Banach Space). *A Banach space is a complete metric space, where a metric space $(X, \| \cdot \|)$ is complete if any Cauchy sequence in X has a limit. That is, for any Cauchy sequence $\{x_i\}_{i=1}^{\infty} \subset X$, there exists $x \in X$, such that $\lim_{n \to \infty} \|x_i - x\| = 0$.*

Definition 3 (L^p Space). *$(\mathbb{R}^n, \| \cdot \|_{L^p})$ means the vector space $\mathbb{R}^n$ equipped with the L^p-norm, i.e.,*

$$\|(x_1, x_2, \cdots, x_n)\|_{L^p} = \left(\sum_{k=1}^{n} |x_n|^p \right)^{1/p} \text{for } 1 \le p < \infty; \\ \|(x_1, x_2, \cdots, x_n)\|_{L^\infty} = \sup_{1 \le k \le n} |x_k| \quad (4)$$

The L^p function spaces are defined using a natural generalization of the L^p-norm for finite-dimensional vector spaces.

Definition 4 ($L^p(\Omega)$ Space). *For $1 \le p < \infty$,*

$$L^p(\Omega) = \{f(x) : \Omega \subset \mathbb{R}^n \to \mathbb{R}, \int_{\Omega} |f(x)|^p dx < \infty\} \quad (5)$$

equipped with the norm $\|f(x)\|_{L^p} = (\int_{\Omega} |f(x)|^p dx)^{1/p}$;

For $p = \infty$,

$$L^\infty(\Omega) = \{f(x) : \Omega \subset \mathbb{R}^n \to \mathbb{R}, \sup_{x \in \Omega} |f(x)| < \infty\} \quad (6)$$

equipped with the norm $\|f(x)\|_{L^\infty} = \sup_{x \in \Omega} |f(x)|$.

Definition 5 (Norm of Linear Operators). *Let X, Y be two Banach spaces. The operator $O : X \to Y$ is linear if and only if $O(\alpha x_1 + \beta x_2) = \alpha O(x_1) + \beta O(x_2)$ for any $\alpha, \beta \in \mathbb{R}$ and $x_1, x_2 \in X$. The operator norm of O is defined by $\|O\|_{op} = \sup_{x \in X, x \ne 0} \frac{\|O(x)\|_Y}{\|x\|_X}$.*

Theorem 1. *Let $B(X;Y)$ be the spaces bounded linear operators, i.e., linear operators with bounded operator norm. Then $B(X;Y)$ is a Banach space.*

Proof. Please refer to the book (Conwa, 2007) for detailed proof. □

The following definitions describe the continuity and differentiability properties of operators between Banach spaces.

Definition 6 (Continuity). *Let X, Y be Banach spaces. An operator $g : X \to Y$ is continuous at $x_0 \in X$ if for any $\varepsilon > 0$, there exists $\delta > 0$ such that for any $x \in X$ satisfying $\|x - x_0\|_X < \delta$, it holds that $\|g(x) - g(x_0)\|_Y < \varepsilon$.*

Definition 7. *[Fréchet Derivative] Let X and Y be two Banach spaces, and g is an operator from X to Y. The operator g is called Fréchet differentiable at $x \in X$ if there exists a bounded linear operator $P : X \to Y$, such that*

$$\lim_{\|h\|_X \to 0} \frac{\|g(x+h) - g(x) - P(x)h\|_Y}{\|h\|_X} \to 0 \qquad (7)$$

The linear operator $P(x)$ is called the Fréchet derivative of O at x.

Definition 8 (Nemytskii Operator). *Let $\Omega \subset \mathbb{R}^n$ be a domain and Y be a Banach space. Given a functional $F : \Omega \times X \to \mathbb{R}$ for any $y \in Y$, a new functional $N(y) : \Omega \to \mathbb{R}$ is defined as $N(y)(x) = F(x;y)$. The operator N is called a Nemytskii operator.*

3. Certifying the Group Fairness

The idea of this work is to develop a Gaussian parameter smoothing method to transform the base classifiers $f(x;W)$ and $f(x;W_k)$ into their smooth versions $\hat{f}(x;W)$ and $\hat{f}(x;W_k)$ and train an optimal individual smooth classifier $\hat{f}(x;W_k)$ for each group G_k with only the data regarding G_k. By leveraging the theory of nonlinear functional analysis, we reformulate the smooth classifiers $\hat{f}(x;W)$ and $\hat{f}(x;W_k)$ as output functions of a Nemytskii operator $\hat{N}(W)(\cdot)$. We theoretically prove that the Nemytskii operator $\hat{N}(W)$ is smooth and therefore $\hat{N}(\mathbb{R}^m)$ is a Fréchet differentiable smooth manifold. We further prove that $\hat{N}(\mathbb{R}^m)$ has a global Lipschitz constant which is independent of the structure of classifiers (e.g., neural networks) and the domain of the input data. Based on global Lipschitz property of $\hat{N}(\mathbb{R}^m)$, we derive the input-agnostic certified group fairness.

First, we use linear functions as an example to demonstrate the concept of Nemytskii operators. In fact, linear functions can be viewed as linear functionals from Banach spaces to $\mathbb{R}$, as well as a Nemytskii operator from certain Banach spaces to the space of bounded linear functionals from L^p to $\mathbb{R}$, i.e., $B(L^p)$. Consider a linear function $g(x) = W \cdot x$,

where $W \in \mathbb{R}^n$ is a constant coefficient vector and $x \in \mathbb{R}^n$. If considering the domain of x as a Banach space $(\mathbb{R}^n, \|\cdot\|_{L^p})$, then $g(x)$ can be viewed as a functional from L^p to $\mathbb{R}$. In addition, by employing the Hölder Inequality, we have $|g(x)| = |W \cdot x| \leq \|W\|_{L^{p'}} \|x\|_{L^p}$, where p' is the conjugacy of p satisfying $\frac{1}{p} + \frac{1}{p'} = 1$. This implies that the operator norm $\|g\|_{op}$ of the linear functional $g(x)$ is $\|g\|_{op} = \sup_{x \in \mathbb{R}^n, x \neq 0} \frac{|g(x)|}{\|x\|_p} = \|W\|_{p'}$. Thus, $g(x)$ is a bounded linear functional from L^p to $\mathbb{R}$ for any $W \in L^{p'}$, i.e., $g(x) \in B(L^p)$ for any $W \in L^{p'}$. If we further consider W as a vector variable, the mapping $N(W):W \to g(x) \in S(L^p)$, i.e., $N(W)(x) = g(x)$, can be viewed as an operator from the Banach space $L^{p'}$ to the Banach space $S(L^p)$. The operator $N(W) : L^{p'} \to S(L^p)$ is a Nemytskii operator.

3.1. Fair Classification with Provable Guarantees

In this paper, we define the certified fairness metric as follows.

Definition 9 (ϵ-fairness in hypothesis space $\mathcal{H}$). *For $1 \leq p < \infty$ and a subset $\Omega \subset \mathbb{R}^n$, an overall classifier $h(x;W) \in \mathcal{H}$ for all groups has ϵ-fairness in $L^p(\Omega)$ space if*

$$|\Omega_k|^{-1}\|h(x;W) - h(x;W_k^*)\|_{L^p(\Omega_k)} \leq \epsilon, \; k = 1, \cdots, K; \; (8)$$

$h(x;W)$ has ϵ-fairness in $L^\infty(\Omega)$ space if

$$\|h(x;W) - h(x;W_k^*)\|_{L^\infty(\Omega_k)} \leq \epsilon, \; k = 1, \cdots, K \quad (9)$$

where $|\Omega_k|$ denotes the cardinality of the subset Ω_k, $h(x;W_k^*)$ is the optimal individual classifier for group G_k in $\mathcal{H}$ and $\Omega_k = \{x \in \Omega : z = z_k\}$.

Definition 10 (Certified ϵ-fairness in hypothesis space $\mathcal{H}$). *If a classifier $h(x) \in \mathcal{H}$ has ϵ-fairness in $L^p(\Omega)$ for any subset $\Omega \subset \mathbb{R}^n$, then $f(x)$ has certified ϵ-fairness in L^p for $1 \leq p \leq \infty$.*

From the above definition, we know that it is important to choose a suitable hypothesis space for certifying the ϵ-fairness. We define the hypothesis space $\mathcal{H}$ as a Banach manifold in L^p space. We reformulate the base classifiers $f(x;W)$ as output functions of a Nemytskii operator $N : \mathbb{R}^m \to L^p$ with $N(W)(x) = f(x;W)$.

We define another Nemytskii operator $\hat{N} : \mathbb{R}^m \to L^p$ with Gaussian parameter smoothing as follows.

$$\hat{N}(W)(\cdot) = \mathbb{E}\left(f(\cdot;W + \Delta)\right), \; \Delta \sim \mathcal{N}(0, \sigma^2 I) \qquad (10)$$

That is,

$$\hat{N}(W)(\cdot) = \frac{1}{(2\pi\sigma^2)^{m/2}} \int_{\mathbb{R}^m} N(W + \Delta)(\cdot) e^{-\frac{\|\Delta\|_2^2}{2\sigma^2}} d\Delta \quad (11)$$

Lemma 1. $\hat{N}$ *is Fréchet differentiable. Therefore,* $\hat{N}(\mathbb{R}^m) \subset L^p$ *is a smooth manifold.*

Proof. Please refer to Appendix A.1 for detailed proof of Lemma 1. $\square$

We define the hypothesis $\hat{\mathcal{H}}$ by $\hat{\mathcal{H}} = \hat{N}(\mathbb{R}^m)(\cdot)$. It is clear that $\hat{\mathcal{H}}$ is not a linear vector space, because $\hat{N}(W_1)(\cdot) + \hat{N}(W_2)(\cdot)$ may not be in $\hat{\mathcal{H}}$. However, $\hat{\mathcal{H}}$ is a smooth manifold which guarantees that the parameters W are trainable. Our goal is to certify ϵ-fairness in $\hat{\mathcal{H}}$ for the best possible ϵ. The key idea is that the smooth manifold $\hat{\mathcal{H}}$ has a global Lipschitz constant which is independent of the structure of classifiers (e.g., neural networks) and the domain of the input data, which is demonstrated in lemma 2 below. In contrast, the manifold $N(\mathbb{R}^m)(\cdot)$ without Gaussian parameter smoothing is only locally Lipschitz with respect to the parameter W and the Lipschitz constant is determined by the input data and the network structures. In addition, the Lipschitz constant regarding W could be rather large since it is hard to control the amplification of difference through propagation over neural networks and thus the Lipschitz constant keeps increasing with the number of layers, which prohibits the certification of group fairness.

Lemma 2. *For any* $W_1, W_2 \in \mathbb{R}^m$ *satisfying* $\|W_1 - W_2\|_2 \leq d$, *it holds the that*

$$|\Omega|^{-1}\|\hat{N}(W_1)(x) - \hat{N}(W_2)(x)\|_{L^p(\Omega)} \leq \frac{d}{\sqrt{2\pi}\sigma}$$

$$if \ 1 \leq p < \infty, \tag{12}$$

$$\|\hat{N}(W_1)(x) - \hat{N}(W_2)(x)\|_{L^\infty(\Omega)} \leq \frac{d}{\sqrt{2\pi}\sigma}$$

for any $\Omega \subset \mathbb{R}^n$.

Proof. Please refer to Appendix A.1 for detailed proof of Lemma 2. $\square$

According to the conclusion of Lemma 2, the following theorem derives the certified fairness by our proposed input-agnostic certified group fairness method and demonstrates that the certified fairness is irrelevant to the domain of the input data.

Theorem 2. *Let* $\hat{N}(W_k^*)$ *be the optimal individual classifier for group* G_k *in the hypothesis space* $\hat{\mathcal{H}}$ ($k = 1, 2, \cdots, K$). *Let* $W^* = \frac{W_1^* + \cdots + W_K^*}{K}$ *and* $\hat{N}(W^*)$ *be the overall fair classifier. If* $\max_{1 \leq k, l \leq K} \|W_k - W_l\|_2 = d$, *then for any* $\Omega \subset \mathbb{R}^n$, *it holds that*

$$|\Omega_k|^{-1}\|\hat{N}(W^*)(x) - \hat{N}(W_k^*)(x)\|_{L^p(\Omega)} \leq \frac{(K-1)d}{\sqrt{2\pi}K\sigma}$$

$$if \ 1 \leq p < \infty, \tag{13}$$

$$\|\hat{N}(W^*)(x) - \hat{N}(W_k^*)(x)\|_{L^\infty(\Omega_k)} \leq \frac{(K-1)d}{\sqrt{2\pi}K\sigma}$$

Therefore, $\hat{N}(W^*)(x)$ *has certified* $\frac{(K-1)d}{\sqrt{2\pi}K\sigma}$-*fairness in* L^p *space for any* $1 \leq p \leq \infty$.

Proof. Please refer to Appendix A.1 for detailed proof of Lemma 2. $\square$

Notice that Theorem 2 is satisfied for any $\Omega \subset \mathbb{R}^n$. Therefore, it provides an input-agnostic certified group fairness guarantee.

In Theorem 2, the standard deviation σ in Gaussian noise serves as a tradeoff hyperparameter to well balance the fairness and accuracy achieved by the smoothed classifier. A larger σ results in better classification fairness, while a smaller σ leads to better accuracy. Especially, when $\sigma \to 0$, the smooth classifier $\hat{N}(W)(x)$ converges to the base classifier $N(W)(x)$, which is verified by Lemma 3.

Theorem 3 (Dominated Convergence Theorem). *Let* $f_n(x)$ *be a sequence of measurable functions. Suppose* $f_n(x) \to f(x)$ *almost everywhere and there exists a integrable function* $g(x)$ *such that* $|f_n(x)| \leq g(x)$ *for any* n *and* x. *Then*

$$\lim_{n\to\infty} \int f_n(x)d\mu = \int f(x)d\mu \tag{14}$$

Proof. Please refer to the book (Conwa, 2007) for detailed proof. $\square$

We will leverage Lebesgue's Dominated Convergence Theorem to prove the following lemma.

Lemma 3. *For any* $x \in \mathbb{R}^n$, $\hat{N}(W)(x) \to N(W)(x)$ *as* $\sigma \to 0$.

Proof. Please refer to Appendix A.1 for detailed proof of Lemma 3. $\square$

Based on Theorem 2, we modify the training objective of fair classification as follows.

$$\min_{W_1, \cdots, W_K} \sum_{k=1}^{K} \underset{(x_i, z_i, y_i) \in G_k}{\mathbb{E}} \left(\mathcal{L}\big(\hat{N}(W_k)(x_i), y_i\big)\right) +$$
$$\alpha \sum_{k=1}^{K} \sum_{l>k}^{K} \|W_k - W_l\|_2^2 \tag{15}$$

After the training, the overall fair classifier $\hat{N}(W^*)(x)$ for all groups is generated by averaging the parameters of the smooth classifier $\hat{N}(W_k^*)(x)$ for each group G_k, i.e., $W^* = \frac{W_1^* + \cdots + W_K^*}{K}$. α is a hyperparameter to balance the fairness and accuracy by $\hat{N}(W^*)(x)$.

By reducing the loss function to local chart of the manifold $\hat{\mathcal{H}}$, i.e. the parameter space, we can view the loss function

as a function of W_k. By changing the notation slightly, the loss function can be rewritten as follows.

$$\min_{W_1,\cdots,W_K} \sum_{k=1}^{K} \mathcal{L}_k(W_k) + \alpha \sum_{k=1}^{K} \sum_{l>k}^{K} \|W_k - W_l\|_2^2 \quad (16)$$

The training process of each individual classifier is given below.

$$
\begin{aligned}
&W_k(t+1)\\
&= W_k(t) - \eta\nabla\mathcal{L}_k(W_k(t)) -\\
&\quad 2\alpha\eta\Big(\sum_{l=1}^{k-1}(W_k(t) - W_l(t)) + \sum_{l=k+1}^{K}(W_k(t) - W_l(t))\Big)
\end{aligned}
\quad (17)
$$

where η is the learning rate and t is the iteration step.

3.2. Computation of Numerical Integrals

Since the dimension of the parameter space of the classifiers, especially deep neural networks, is often pretty high, it is not practical to use deterministic numerical integration methods due to the high computational cost. In our implementation, we utilize the Monte-Carlo integration technique to compute the numerical integrals.

$$\hat{N}(W)(\cdot) = \frac{1}{(2\pi\sigma^2)^{m/2}} \int_{\mathbb{R}^m} N(W+\Delta)(\cdot)e^{-\frac{\|\Delta\|_2^2}{2\sigma^2}} d\Delta \quad (18)$$

Specifically, we sample $\Delta_j \sim \mathcal{N}(0, \sigma^2 I)$, $j = 1, \cdots, 10^5$, i.e., the sample size is $100,000$, where $\mathcal{N}(0, \sigma^2 I)$ is the isotropic Gaussian distribution. We then calculate the following Monte-Carlo integration.

$$\hat{N}(W)(x) \approx 10^{-5} \sum_{j=1}^{10^5} N(W+\Delta_j)(x) \quad (19)$$

By using the standard error analysis of Monte-Carlo integration, the probability of the error of the above Monte-Carlo integral lying between the following range is equal to $\frac{1}{2\pi}\int_{-a}^{a} e^{-t^2/2}dt$.

$$\left(-a\mathbb{V}(N(W+\Delta)(x))/\sqrt{10^5}, a\mathbb{V}(N(W+\Delta)(x))/\sqrt{10^5}\right) \quad (20)$$

where $\mathbb{V}(N(W+\Delta)(x))$ is the variance of $N(W+\Delta)(x)$ for $\Delta \sim \mathcal{N}(0, \sigma^2 I)$.

Since the classifier $|N(W)(x)| \leq 1$ for any W and x, $\mathbb{V}(N(W+\Delta)(x)) \leq 1$. Taking $a = 3$ as an example in Eq.(20), we obtain that the probability of the error of the numerical integral in Eq.(19) being less than 0.01 is about 99.7%. This implies that our input-agnostic certified group fairness technique has great potential to achieve the superior certified fairness guarantees for any input data, which is desirable in practice.

4. Experiments

In this section, we have evaluated the certified fairness of our FairSmooth model and other comparison methods over two benchmark datasets from the fair classification: Adult from the UCI Machine Learning Repository (UCI) and COMPAS introduced by the ProPublica (ProPublica). Both datasets have binary labels and a mixture of numerical and categorical features. The Adult dataset contains 48,842 individuals with 18 binary features and a label indicating whether the income is greater than 50,000 USD or not. The COMPAS dataset consists of 6,172 individuals with 10 binary features and a label that takes value 1 if the individual does not re-offend and 0 otherwise. In both datasets, We use *Sex* and *Race* as the protected attributes respectively. The experiments exactly follow the same settings described by other group fairness methods with provable guarantees (Ustun et al., 2019; Krishnaswamy et al., 2021; Celis et al., 2021a; Iofinova et al., 2021). The datasets are preprocessed by using AIF360 toolkit (Bellamy et al., 2018).

Baselines. We compare the FairSmooth model with ten state-of-the-art group fairness models, including six regular group fairness and three provable group fairness guarantee approaches. **ApxFair** is a classifier that are fair not only with respect to the training distribution, but also for a class of distributions that are weighted perturbations of the training samples (Mandal et al., 2020). **ARL** is an adversarially reweighted learning model that aims to improve the utility for worst-off protected groups, without access to protected features at training or inference time (Lahoti et al., 2020). **Fair Mixup** is a data augmentation strategy that improves the generalization of group fairness metrics (Chuang & Mroueh, 2021). **FairBatch** adaptively selects minibatch sizes for the purpose of improving model fairness based on various prominent fairness measures (Roh et al., 2021b). **fair-robust-selection (FRS)** is a sample selection-based algorithm for fair and robust training by solving a combinatorial optimization problem for the unbiased selection of samples (Roh et al., 2021c). **Implicit** derives an implicit path alignment algorithm to learn a fair representation by encouraging the invariant optimal predictors on the top of data representation (Anonymous, 2022c). **Group-Fair** is a meta-algorithm for provable group fairness guarantees by approximately reducing classification problems with general types of fairness constraints to ones with convex constraints (Celis et al., 2019). **FCRL** provides theoretical guarantees on the parity of downstream classifiers by limiting the mutual information between representations and protected attributes (Gupta et al., 2021). **DLR** is an optimization framework for learning a fair classifier with provable guarantees in the presence of noisy perturbations in the protected attributes. (Celis et al., 2021a). To our best knowledge, this work is the first to certify the group fairness of classifiers with theoretical input-agnostic guarantees, while there is no need to know the shift between training and deployment datasets with respect to sensitive attributes.

Table 1: Accuracy and Fairness with *Sex* Attribute

Dataset	Adult			COMPAS		
Protected Attribute	Sex			Sex		
Metric	**Acc.**	Δ_{DP}	Δ_{EO}	**Acc.**	Δ_{DP}	Δ_{EO}
ApxFair	0.7	0.24	0.18	0.52	0.21	0.58
ARL	**0.84**	0.13	0.1	0.49	0.17	0.28
Fair Mixup	0.76	**0.05**	0.27	0.66	0.17	0.29
FairBatch	**0.84**	0.19	0.19	0.65	0.34	0.64
FRS	0.44	0.11	0.16	0.31	0.13	**0.19**
Implicit	0.8	0.11	0.15	0.64	0.17	0.3
Group-Fair	0.76	0.33	0.51	0.44	0.13	0.25
FCRL	0.81	**0.05**	0.22	0.56	0.19	0.39
DLR	0.5	0.13	0.22	0.49	0.28	0.46
FairSmooth	**0.84**	**0.05**	**0.08**	**0.67**	**0.11**	**0.19**

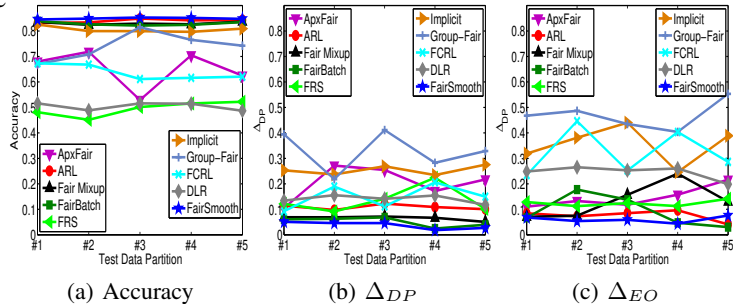(a) Accuracy   (b) Δ_{DP} (c) Δ_{EO}

Figure 1: Performance with Varying Test Data Partitions on Adult

Variants of FairSmooth model. We evaluate three versions of FairSmooth to show the strengths of different techniques. By leveraging the theory of nonlinear functional analysis, FairSmooth utilizes our proposed Gaussian parameter smoothing method with parameter disparity term to produce a fair classifier with input-agnostic certified group fairness guarantees. FairSmooth-S makes use of the Gaussian parameter smoothing only, while FairSmooth-D employs the parameter disparity term only.

Evaluation metrics. We use two popular measures in fair machine learning to verify the fairness of different methods: **Demographic Parity Disparity** (Δ_{DP}) and **Equalized Odds Disparity** (Δ_{EO}) (Cho et al., 2020; Chuang & Mroueh, 2021; Roh et al., 2021a;a; Gupta et al., 2021; Pan et al., 2021; Anonymous, 2022b). Demographic Parity defines fairness as an equal probability of being classified with the positive label, i.e., each group has the same probability of being classified with the positive outcome (Caton & Haas, 2020). Equalized Odds defines fairness as an equal probability if the true positive rate and false positive rate of a classifier is the same across different groups (Caton & Haas, 2020). Δ_{DP} or Δ_{EO} represent the difference of Demographic Parity or Equalized Odds between sensitive groups respectively. A smaller Δ_{DP} or Δ_{EO} score indicates a better fairness. In addition, we use **Accuracy** to evaluate the quality of the fair classification algorithms.

Performance with *Sex* as protected attribute. Table 1 exhibits the *Accuracy*, Δ_{DP}, and Δ_{EO} scores of ten fair classification algorithms over two datasets of Adult and COMPAS. It is observed that among ten fair classification methods our FairSmooth method achieve the highest *Accuracy*, the smallest Δ_{DP}, and the lowest Δ_{EO} on two datasets in most experiments, showing the superior quality and fairness of FairSmooth against protected attributes. Compared to the fair classification results by other models, FairSmooth, on average, achieves 18.7% *Accuracy* boost, 48.2% Δ_{DP} improvement, and 49.1% Δ_{EO} decrease on two datasets. In addition, the promising performance of FairSmooth over

both datasets implies that FairSmooth has great potential as a general fair classification solution to other datasets, which is desirable in practice.

Performance with varying test data partitions on Adult. Figures 1 presents the fair classification performance by varying the distributions of test data on Adult. We randomly divide the test data into five partitions, which sample 5%, 10%, 15%, 30%, and 40% of the entire test data respectively. Therefore, these five partitions have quite different distributions from each other. In addition, we allow the same samples to occur in multiple partitions. It is obvious that the performance curves by other classification algorithms oscillate up and down with varying test data partitions. This phenomenon indicates that not only the Accuracy but also the Fairness of current fair classification methods are sensitive to test data distributions. However, the curves by our FairSmooth keep relatively stable. This demonstrates that our FairSmooth is able to produce the input-agnostic fair classification results, without the knowledge or assumption about the distribution of training and deployment data. In addition, FairSmooth still achieves the highest *Accuracy* values (> 0.83), the smallest Δ_{DP} scores (< 0.11), and the lowest Δ_{EO} values (< 0.10), which are better than other nine methods in most tests.

Ablation study. Table 2 presents the performance of fair classification on two datasets with three variants of our FairSmooth model. We observe the complete FairSmooth achieves the highest *Accuracy* values (> 0.66), the smallest Δ_{DP} scores (< 0.05), and the lowest Δ_{EO} values (< 0.08), which are obviously better than other versions. Compared with FairSmooth-S, FairSmooth performs well in most experiments. A reasonable explanation is that the parameter disparity term forces all individual classifiers to approach each other, which helps recover the most accurate model for each group as well as produce more fair overall classifier. In addition, FairSmooth achieves the better performance than FairSmooth-D. A rational guess is that the parameter disparity term only fails to provide a tight fairness guarantee without help of Gaussian parameter smoothing.

Table 2: Accuracy and Fairness of FairSmooth variants

Dataset	Adult			COMPAS		
Protected Attribute	Sex			Race		
Metric	Acc.	Δ_{DP}	Δ_{EO}	Acc.	Δ_{DP}	Δ_{EO}
FairSmooth-S	**0.84**	0.2	0.22	0.59	0.15	0.25
FairSmooth-D	0.81	0.27	0.27	0.56	0.07	0.1
FairSmooth	**0.84**	**0.05**	**0.08**	**0.66**	**0.03**	**0.01**

Table 3: Fairness Certificate

Dataset	Adult	COMPAS
Average Distance d	0.0074	0.0057
Fairness Certificate $\frac{(K-1)d}{\sqrt{2\pi}K\sigma}$	0.0147	0.0114

Figure 2: Performance with Varying Group Ratios
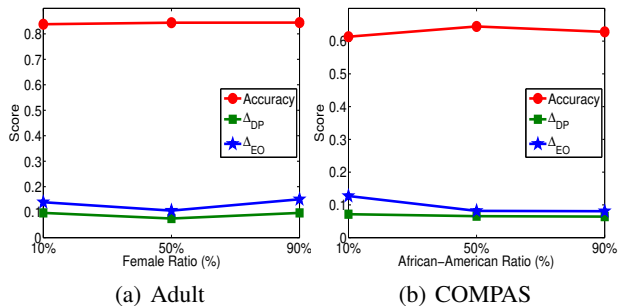

(a) σ (b) α

Figure 3: Performance with Varying Parameters on Adult

Impact of group ratios. Figure 2 shows the quality of ten fair classification algorithms on Adult (and COMPAS) with *Sex* (and *Race*) as protected attribute by varying the *Female* (and *African-American*) ratio of training data from 10% to 90% respectively. We randomly reduce the number of *Female* (or *African-American*) training samples for decreasing the ratio or reduce the number of *Male* (or *Other*) training samples for increasing the ratio. We average metrics over 10 trials of randomly selected groups for imbalanced experiments. The manually imbalanced datasets are used to benchmark fair classification performance. We make the following observations on the performances by our fair classification algorithm: (1) The accuracy (or fairness) curves initially keep increasing (or fairness) and then slightly drop (or raise) when the ratio increases; (2) Although we see that there is a little oscillation in two figures, the performance curves still keep relative stable. A reasonable explanation is that the group imbalance degree of training data definitely affects the performance of our fair classification method. However, our FairSmooth method is still robust to imbalanced training data with the help of Gaussian parameter smoothing and parameter disparity term.

Impact of σ and α on Adult. Figures 3 (a) and (b) show the impact of the standard deviation σ in Gaussian distribution and the weight α of parameter disparity term in our FairSmooth model over the Adult dataset. The performance curves initially improve and finally become worse when two parameters increase. This demonstrates that there must exist an optimal σ and α for the Gaussian parameter smoothing and the parameter disparity term. A too large σ may introduce too much Gaussian noise and thus it is hard to
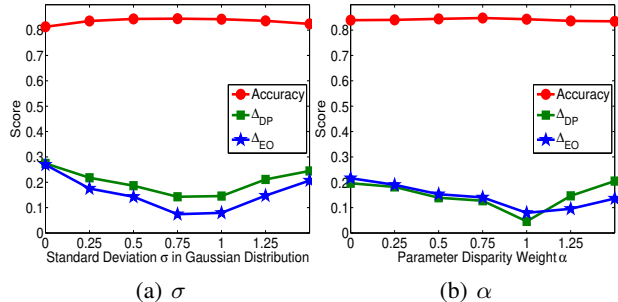
guarantee the performance of Gaussian parameter smoothing, while a too small σ may loss the strength of Gaussian parameter smoothing. Similarly, a too large α may make all individual classifiers become the same and thus fails to ensure the classification accuracy, while a too small α may make all individual classifiers far away from each other and thus fails to provide the fairness guarantees. Thus, it is important to choose the appropriate parameters for well balancing the accuracy and fairness.

Fairness certificate computation. Table 3 presents the average distance between the parameters of the overall and individual classifiers on Adult and COMPAS. We also measure the fairness certificates $\frac{(K-1)d}{\sqrt{2\pi}K\sigma}$ over two datasets. In the current experiments, the calculated average distances and fairness certificates on two datasets are very small with order 10^{-3} and 10^{-2} respectively. This implies that introducing the Gaussian parameter smoothing method and parameter disparity term to the training objective is able to help the overall classifier achieve high accuracy, which providing good fairness guarantees.

5. Related Work

Fair Classification. Trustworthy machine learning has attracted active research in recent years (Palanisamy et al., 2018; Lei et al., 2019; Zhou et al., 2020; Zhang et al., 2020b; Zhou & Liu, 2019; Wu et al., 2021; Zhou et al., 2021b; Zhao et al., 2021; Ren et al., 2021; Zhou et al., 2021a; Ma et al., 2021; Zhang et al., 2021). Fair classification techniques aim to guarantee the learnt classifiers that are not only accurate but also fair with respect to sensitive attributes (Caton & Haas, 2020; Mehrabi et al., 2021). Existing techniques on fair classification can be broadly classified into three categories based on which stage of the machine learning pipeline they target: (1) Pre-processing techniques try to

pre-process the original dataset for reducing bias present in the training data and achieve an unbiased dataset for learning (Kamiran & Calders, 2011; Thanh et al., 2011; Zemel et al., 2013; Feldman et al., 2015; Edwards & Storkey, 2016; McNamara et al., 2017; Krasanakis et al., 2018; Madras et al., 2018; Song et al., 2019; Cisse & Koyejo, 2019; Gupta et al., 2021). The pre-processing methods are beneficial when the onus of fairness is on a third party or the data controller (McNamara et al., 2017; Cisse & Koyejo, 2019; Gupta et al., 2021). One of the main objectives of the pre-processing methods is to ensure that any downstream classifiers can freely use the sanitized datasets for fair machine learning (Edwards & Storkey, 2016; McNamara et al., 2017; Song et al., 2019; Madras et al., 2018). However, their operationalization often leads to an approach that may not ensure fairness guarantees (Gupta et al., 2021); (2) In-processing approaches aim to regularize the classifier during training to ensure the learning fairness (e.g. Zafar et al. (2017)). The in-processing algorithms can be categorized into two core mechanisms. Penalty-based methods add constraints (Hardt et al., 2016; Zafar et al., 2017a;b; Zhang et al., 2018; Nabi & Shpitser, 2018; Goel et al., 2018; Menon & Williamson, 2018; Narasimhan, 2018; Kim et al., 2018; Celis et al., 2019; Chierichetti et al., 2019; Nabi et al., 2019; Cotter et al., 2019) or a regularizer (Kamishima et al., 2012; Goel et al., 2018; Heidari et al., 2018; Aghaei et al., 2019; Huang & Vishnoi, 2019; Jiang et al., 2019; Mandal et al., 2020; Saralajew et al., 2020; Chuang & Mroueh, 2021) to the learning objective that penalize or prevent parameter choices that lead to unfair decisions. Adversarial learning-based methods train an adversary in parallel to the classifier that tries to predict the protected attribute from the model outputs (Beutel et al., 2017; Wadsworth et al., 2018; Zhang et al., 2018; Lahoti et al., 2020). If this cannot be done better than chance level, the fairness is achieved. However, many of the aforementioned algorithms do not come with theoretical fairness guarantees; perhaps because the manually-crafted heuristics are used to generate fair classifiers and the resulting optimization problem is non-convex; (3) Post-processing algorithms formulate fair classification problems as constrained optimization problems: first train an unconstrained optimal classifier and then shift the classifiers' decision boundary according to the fairness requirement (Fish et al., 2016; Hardt et al., 2016; Goh et al., 2016; Woodworth et al., 2017; Pleiss et al., 2017; Dwork et al., 2018; Chzhen et al., 2020). This is a simple, reliable and often effective method, but it requires the protected attribute to be available at prediction time.

Group Fairness with Provable Guarantees. Group-based fair learning models aim to reduce the performance gaps between different protected groups with sensitive attributes (Zafar et al., 2017a;b; Goel et al., 2018; Krasanakis et al., 2018; Menon & Williamson, 2018; Mandal et al., 2020; Wang et al., 2020; Yang et al., 2020; Zhang et al., 2020a; Lahoti et al., 2020; Ben-Porat et al., 2021; Celis et al., 2021a; Gorantla et al., 2021; Chuang & Mroueh, 2021; Roh et al., 2021a; Shekhar et al., 2021; Qian et al., 2021; Liu & Wang, 2021; Maity et al., 2021; Celis et al., 2021b; Du et al., 2021; Roh et al., 2021a; Anonymous, 2022a; Zhu et al., 2022). However, most of existing studies fail to provide a provable fairness guarantee, because the manually-crafted heuristics are used to generate fair classifiers and the resulting optimization problems are non-convex (Zafar et al., 2017a;b; Krasanakis et al., 2018). Only recently, researchers attempt to theoretically provide different classification algorithms with provable fairness guarantees (Feldman et al., 2015; McNamara et al., 2017; Celis et al., 2019; Ustun et al., 2019; Gupta et al., 2021; Krishnaswamy et al., 2021; Segal et al., 2021; Celis et al., 2021a; Iofinova et al., 2021; Anonymous, 2022b).

Most of the above fairness guarantees are based on an assumption that the training and deployment data follow the same distribution. However, this assumption is false for many real-world problems (Zhuang et al., 2021). A recent study reports that the models based on the above assumption often violate these guarantees and exhibit unfair bias when evaluated on data from a different distribution (Anonymous, 2022b). Shifty is the first method to provide provable guarantees under a demographic shift between training and deployment data (Anonymous, 2022b). However, the fact that the old and new demographic proportions should be known in Shifty, which limits the applicability of Shifty in real-world scenarios.

To our best knowledge, this work is the first to certify the group fairness of classifiers with theoretical input-agnostic guarantees, while there is no need to know the shift between training and deployment datasets with respect to sensitive attributes, by leveraging the theory of nonlinear functional analysis, including Nemytskii operator and smooth manifold.

6. Conclusions

In this work, we proposed an input-agnostic certified group fairness algorithm for improving the fairness of classification models while maintaining the remarkable prediction accuracy. First, a Gaussian parameter smoothing method is proposed to transform base classifiers into their smooth versions by averaging its prediction over Gaussian perturbations of the former's parameter within its neighborhood. Second, an overall smooth classifier for all groups is generated by averaging the parameters of all the optimal individual smooth classifiers. Finally, the theoretical analysis is conducted to verify that the smooth classifier is able to achieve the certified fairness guarantees for any input data.

References

Aghaei, S., Azizi, M. J., and Vayanos, P. Learning optimal and fair decision trees for non-discriminative decision-making. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pp. 1418–1426, 2019.

Anonymous. Is fairness only metric deep? evaluating and addressing subgroup gaps in deep metric learning. In *Under Review. 10th International Conference on Learning Representations, ICLR 2022, Online, April 25-29, 2022, Conference Track Proceedings*, 2022a.

Anonymous. Fairness guarantees under demographic shift. In *Under Review. 10th International Conference on Learning Representations, ICLR 2022, Online, April 25-29, 2022, Conference Track Proceedings*, 2022b.

Anonymous. Fair representation learning through implicit path alignment. In *Under Review. 10th International Conference on Learning Representations, ICLR 2022, Online, April 25-29, 2022, Conference Track Proceedings*, 2022c.

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J. T., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., and Zhang, Y. AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *CoRR*, abs/1810.01943, 2018.

Ben-Porat, O., Sandomirskiy, F., and Tennenholtz, M. Protecting the protected group: Circumventing harmful fairness. 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*, pp. 5176–5184. AAAI Press, 2021.

Beutel, A., Chen, J., Zhao, Z., and Chi, E. H. Data decisions and theoretical implications when adversarially learning fair representations. In *Proceedings of the 2017 Conference on Fairness, Accountability and Transparency (FAccT)*, 2017.

Blum, A., Dick, T., Manoj, N., and Zhang, H. Random smoothing might be unable to certify l∞ robustness for high-dimensional images. *J. Mach. Learn. Res.*, 21:211:1–211:21, 2020.

Caton, S. and Haas, C. Fairness in machine learning: A survey. *CoRR*, abs/2010.04053, 2020.

Celis, L. E., Huang, L., Keswani, V., and Vishnoi, N. K. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pp. 319–328, 2019.

Celis, L. E., Huang, L., Keswani, V., and Vishnoi, N. K. Fair classification with noisy protected attributes: A framework with provable guarantees. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1349–1361. PMLR, 2021a.

Celis, L. E., Mehrotra, A., and Vishnoi, N. K. Fair classification with adversarial perturbations. In *The 35th Conference on Neural Information Processing Systems, (NeurIPS'21)*, Online, December 6-14 2021b.

Chierichetti, F., Kumar, R., Lattanzi, S., and Vassilvitskii, S. Matroids, matchings, and fairness. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, pp. 2212–2220, 2019.

Cho, J., Hwang, G., and Suh, C. A fair classifier using kernel density estimation. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Chuang, C. and Mroueh, Y. Fair mixup: Fairness via interpolation. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Chzhen, E., Denis, C., Hebiri, M., Oneto, L., and Pontil, M. Fair regression with wasserstein barycenters. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Cisse, M. and Koyejo, S. Fairness and representation learning. In *Invited Talk, Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 2019.

Cohen, J. M., Rosenfeld, E., and Kolter, J. Z. Certified adversarial robustness via randomized smoothing. In

Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, pp. 1310–1320, 2019.

Conwa, J. B. *A Course in Functional Analysis*. Springer, 2007.

Cotter, A., Jiang, H., Gupta, M. R., Wang, S., Narayan, T., You, S., and Sridharan, K. Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *J. Mach. Learn. Res.*, 20: 172:1–172:59, 2019.

Dedman, B. The color of money, 1988.

Du, M., Mukherjee, S., Wang, G., Tang, R., Awadallah, A., and Hu, X. Fairness via representation neutralization. In *The 35th Conference on Neural Information Processing Systems, (NeurIPS'21)*, Online, December 6-14 2021.

Dvijotham, K. D., Hayes, J., Balle, B., Kolter, J. Z., Qin, C., György, A., Xiao, K., Gowal, S., and Kohli, P. A framework for robustness certification of smoothed classifiers using f-divergences. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.

Dwork, C., Immorlica, N., Kalai, A. T., and Leiserson, M. D. M. Decoupled classifiers for group-fair and efficient machine learning. In Friedler, S. A. and Wilson, C. (eds.), *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA*, volume 81 of *Proceedings of Machine Learning Research*, pp. 119–133. PMLR, 2018.

Edwards, H. and Storkey, A. J. Censoring representations with an adversary. In Bengio, Y. and LeCun, Y. (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and removing disparate impact. In Cao, L., Zhang, C., Joachims, T., Webb, G. I., Margineantu, D. D., and Williams, G. (eds.), *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*, pp. 259–268. ACM, 2015.

Feng, H., Wu, C., Chen, G., Zhang, W., and Ning, Y. Regularized training and tight certification for randomized smoothed classifier with provable robustness. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 3858–3865, 2020.

Fish, B., Kun, J., and Lelkes, Á. D. A confidence-based approach for balancing fairness and accuracy. In Venkatasubramanian, S. C. and Jr., W. M. (eds.), *Proceedings of the 2016 SIAM International Conference on Data Mining, Miami, Florida, USA, May 5-7, 2016*, pp. 144–152. SIAM, 2016.

Goel, N., Yaghini, M., and Faltings, B. Non-discriminatory machine learning through convex fairness criteria. In McIlraith, S. A. and Weinberger, K. Q. (eds.), *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 3029–3036, 2018.

Goh, G., Cotter, A., Gupta, M. R., and Friedlander, M. P. Satisfying real-world goals with dataset constraints. In Lee, D. D., Sugiyama, M., von Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 2415–2423, 2016.

Gorantla, S., Deshpande, A., and Louis, A. On the problem of underranking in group-fair ranking. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3777–3787. PMLR, 2021.

Gupta, U., Ferber, A. M., Dilkina, B., and Steeg, G. V. Controllable guarantees for fair outcomes via contrastive information estimation. 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*, pp. 7610–7619. AAAI Press, 2021.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In Lee, D. D., Sugiyama, M., von Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 3315–3323, 2016.

Hayes, J. Extensions and limitations of randomized smoothing for robustness guarantees. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*

Workshops 2020, Seattle, WA, USA, June 14-19, 2020, pp. 3413–3421, 2020.

Heidari, H., Ferrari, C., Gummadi, K. P., and Krause, A. Fairness behind a veil of ignorance: A welfare analysis for automated decision making. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 1273–1283, 2018.

Huang, L. and Vishnoi, N. K. Stable and fair classification. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pp. 2879–2890, 2019.

Hvistendahl, M. Can 'predictive policing' prevent crime before it happens, 2016.

Iofinova, E., Konstantinov, N., and Lampert, C. H. FLEA: provably fair multisource learning from unreliable training data. *CoRR*, abs/2106.11732, 2021.

Jiang, R., Pacchiano, A., Stepleton, T., Jiang, H., and Chiappa, S. Wasserstein fair classification. In *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2019, Tel Aviv, Israel, July 22-25, 2019*, pp. 862–872, 2019.

Kamiran, F. and Calders, T. Data preprocessing techniques for classification without discrimination. *Knowl. Inf. Syst.*, 33(1):1–33, 2011.

Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. Fairness-aware classifier with prejudice remover regularizer. In Flach, P. A., Bie, T. D., and Cristianini, N. (eds.), *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II*, volume 7524 of *Lecture Notes in Computer Science*, pp. 35–50. Springer, 2012.

Kim, M. P., Reingold, O., and Rothblum, G. N. Fairness through computationally-bounded awareness. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 4847–4857, 2018.

Krasanakis, E., Xioufis, E. S., Papadopoulos, S., and Kompatsiaris, Y. Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In Champin, P., Gandon, F., Lalmas, M., and Ipeirotis, P. G. (eds.), *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018*, pp. 853–862. ACM, 2018.

Krishnaswamy, A. K., Jiang, Z., Wang, K., Cheng, Y., and Munagala, K. Fair for all: Best-effort fairness guarantees for classification. In Banerjee, A. and Fukumizu, K. (eds.), *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pp. 3259–3267. PMLR, 2021.

Lahoti, P., Beutel, A., Chen, J., Lee, K., Prost, F., Thain, N., Wang, X., and Chi, E. Fairness without demographics through adversarially reweighted learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Lang, S. *Algebra*. Addison-Wesley, 1993.

Lécuyer, M., Atlidakis, V., Geambasu, R., Hsu, D., and Jana, S. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy, SP 2019, San Francisco, CA, USA, May 19-23, 2019*, pp. 656–672, 2019.

Lee, G., Yuan, Y., Chang, S., and Jaakkola, T. S. Tight certificates of adversarial robustness for randomly smoothed classifiers. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pp. 4911–4922, 2019.

Lei, Q., Wu, L., Chen, P., Dimakis, A., Dhillon, I. S., and Witbrock, M. J. Discrete adversarial attacks and submodular optimization with applications to text classification. In Talwalkar, A., Smith, V., and Zaharia, M. (eds.), *Proceedings of Machine Learning and Systems 2019, MLSys 2019, Stanford, CA, USA, March 31 - April 2, 2019*, 2019.

Levine, A. and Feizi, S. Robustness certificates for sparse adversarial attacks by randomized ablation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 4585–4593, 2020.

Li, B., Chen, C., Wang, W., and Carin, L. Certified adversarial robustness with additive noise. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pp. 9459–9469, 2019.

Liu, Y. and Wang, J. Can less be more? when increasing-to-balancing label noise rates considered beneficial. In

The 35th Conference on Neural Information Processing Systems, (NeurIPS'21), Online, December 6-14 2021.

Ma, Y., Wang, S., Derr, T., Wu, L., and Tang, J. Graph adversarial attack via rewiring. In Zhu, F., Ooi, B. C., and Miao, C. (eds.), *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, pp. 1161–1169, 2021.

Madras, D., Creager, E., Pitassi, T., and Zemel, R. S. Learning adversarially fair and transferable representations. In Dy, J. G. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3381–3390. PMLR, 2018.

Maity, S., Mukherjee, D., Yurochkin, M., and Sun, Y. Does enforcing fairness mitigate biases caused by subpopulation shift? In *The 35th Conference on Neural Information Processing Systems, (NeurIPS'21)*, Online, December 6-14 2021.

Mandal, D., Deng, S., Jana, S., Wing, J. M., and Hsu, D. J. Ensuring fairness beyond the training data. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

McNamara, D., Ong, C. S., and Williamson, R. C. Provably fair representations. *CoRR*, abs/1710.04394, 2017.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6):115:1–115:35, 2021.

Menon, A. K. and Williamson, R. C. The cost of fairness in binary classification. In Friedler, S. A. and Wilson, C. (eds.), *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA*, volume 81 of *Proceedings of Machine Learning Research*, pp. 107–118. PMLR, 2018.

Miller, C. C. Can an algorithm hire better than a human?, 2015.

Mohapatra, J., Ko, C., Weng, T., Chen, P., Liu, S., and Daniel, L. Higher-order certification for randomized smoothing. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Nabi, R. and Shpitser, I. Fair inference on outcomes. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 1931–1940, 2018.

Nabi, R., Malinsky, D., and Shpitser, I. Learning optimal fair policies. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pp. 4674–4682, 2019.

Narasimhan, H. Learning with complex loss functions and constraints. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, pp. 1646–1654, 2018.

Northpointe. Compas risk and needs assessment system, 2012.

Palanisamy, B., Liu, L., Zhou, Y., and Wang, Q. Privacy-preserving publishing of multilevel utility-controlled graph datasets. *ACM Transactions on Internet Technology (TOIT)*, 18(2):24:1–24:21, 2018.

Pan, W., Cui, S., Bian, J., Zhang, C., and Wang, F. Explaining algorithmic fairness through fairness-aware causal path decomposition. In Zhu, F., Ooi, B. C., and Miao, C. (eds.), *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, pp. 1287–1297. ACM, 2021.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J. M., and Weinberger, K. Q. On fairness and calibration. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5680–5689, 2017.

ProPublica. Compas recidivism risk score data and analysis. https://github.com/propublica/compas-analysis/.

Qian, S., Pham, H., Lutellier, T., Hu, Z., Kim, J., Tan, L., Yu, Y., Chen, J., and Shah, S. Are my deep learning systems fair? an empirical study of fixed-seed training. In *The 35th Conference on Neural Information Processing Systems, (NeurIPS'21)*, Online, December 6-14 2021.

Ren, J., Zhang, Z., Jin, J., Zhao, X., Wu, S., Zhou, Y., Shen, Y., Che, T., Jin, R., and Dou, D. Integrated defense for resilient graph matching. In *Proceedings of the 38th International Conference on Machine Learning, (ICML'21)*, pp. 8982–8997, Virtual Event, July 18-24 2021.

Riesz, F. and Sz.-Nagy, B. *Functional Analysis*. Dover Publications, 1955.

Roh, Y., Lee, K., Whang, S., and Suh, C. Sample selection for fair and robust training. In *The 35th Conference on Neural Information Processing Systems, (NeurIPS'21)*, Online, December 6-14 2021a.

Roh, Y., Lee, K., Whang, S. E., and Suh, C. Fairbatch: Batch selection for model fairness. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021b.

Roh, Y., Lee, K., Whang, S. E., and Suh, C. Sample selection for fair and robust training. In *The 35th Conference on Neural Information Processing Systems, (NeurIPS'21)*, Online, December 6-14 2021c.

Salman, H., Li, J., Razenshteyn, I. P., Zhang, P., Zhang, H., Bubeck, S., and Yang, G. Provably robust deep learning via adversarially trained smoothed classifiers. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pp. 11289–11300, 2019.

Saralajew, S., Holdijk, L., and Villmann, T. Fast adversarial robustness certification of nearest prototype classifiers for arbitrary seminorms. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Segal, S., Adi, Y., Pinkas, B., Baum, C., Ganesh, C., and Keshet, J. Fairness in the eyes of the data: Certifying machine-learning models. In Fourcade, M., Kuipers, B., Lazar, S., and Mulligan, D. K. (eds.), *AIES '21: AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event, USA, May 19-21, 2021*, pp. 926–935. ACM, 2021.

Shekhar, S., Ghavamzadeh, M., and Javidi, T. Adaptive sampling for minimax fair classification. In *The 35th Conference on Neural Information Processing Systems, (NeurIPS'21)*, Online, December 6-14 2021.

Song, J., Kalluri, P., Grover, A., Zhao, S., and Ermon, S. Learning controllable fair representations. In Chaudhuri, K. and Sugiyama, M. (eds.), *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pp. 2164–2173. PMLR, 2019.

Thanh, B. L., Ruggieri, S., and Turini, F. k-nn as an implementation of situation testing for discrimination discovery and prevention. In Apté, C., Ghosh, J., and Smyth, P. (eds.), *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, August 21-24, 2011*, pp. 502–510. ACM, 2011.

UCI. Uci machine learning repository. http://archive.ics.uci.edu/ml/index.php.

Ustun, B., Liu, Y., and Parkes, D. C. Fairness without harm: Decoupled classifiers with preference guarantees. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6373–6382. PMLR, 2019.

Wadsworth, C., Vera, F., and Piech, C. Achieving fairness through adversarial learning: an application to recidivism prediction. In *Proceedings of the 2018 Conference on Fairness, Accountability and Transparency (FAccT)*, 2018.

Wang, S., Guo, W., Narasimhan, H., Cotter, A., Gupta, M. R., and Jordan, M. I. Robust optimization for fairness with noisy protected groups. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Woodworth, B. E., Gunasekar, S., Ohannessian, M. I., and Srebro, N. Learning non-discriminatory predictors. In Kale, S. and Shamir, O. (eds.), *Proceedings of the 30th Conference on Learning Theory, COLT 2017, Amsterdam, The Netherlands, 7-10 July 2017*, volume 65 of *Proceedings of Machine Learning Research*, pp. 1920–1953. PMLR, 2017.

Wu, S., Li, Y., Zhang, D., Zhou, Y., and Wu, Z. Topicka: Generating commonsense knowledge-aware dialogue responses towards the recommended topic fact. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, (IJCAI'20)*, pp. 3766–3772, Online, January 7-15 2021.

Yang, F., Cisse, M., and Koyejo, O. Fairness with overlapping groups; a probabilistic perspective. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., and Gummadi, K. P. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In Barrett, R., Cummings, R., Agichtein,

E., and Gabrilovich, E. (eds.), *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*, pp. 1171–1180. ACM, 2017a.

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., and Gummadi, K. P. Fairness constraints: Mechanisms for fair classification. In Singh, A. and Zhu, X. J. (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, volume 54 of *Proceedings of Machine Learning Research*, pp. 962–970. PMLR, 2017b.

Zemel, R. S., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, pp. 325–333. JMLR.org, 2013.

Zhai, R., Dan, C., He, D., Zhang, H., Gong, B., Ravikumar, P., Hsieh, C., and Wang, L. MACER: attack-free and scalable robust training via maximizing certified radius. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.

Zhang, B. H., Lemoine, B., and Mitchell, M. Mitigating unwanted biases with adversarial learning. In Furman, J., Marchant, G. E., Price, H., and Rossi, F. (eds.), *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2018, New Orleans, LA, USA, February 02-03, 2018*, pp. 335–340. ACM, 2018.

Zhang, X., Tu, R., Liu, Y., Liu, M., Kjellström, H., Zhang, K., and Zhang, C. How do fair decisions fare in long-term qualification? In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020a.

Zhang, Z., Zhang, Z., Zhou, Y., Shen, Y., Jin, R., and Dou, D. Adversarial attacks on deep graph matching. In *The 34th Conference on Neural Information Processing Systems (NeurIPS'20)*, Virtual, December 6-12 2020b.

Zhang, Z., Zhang, Z., Zhou, Y., Wu, L., Wu, S., Han, X., Dou, D., Che, T., and Yan, D. Adversarial attack against cross-lingual knowledge graph alignment. In *Proceedings of the 26th Conference on Empirical Methods in Natural Language Processing, (EMNLP'21)*, pp. 5320–5337, Virtual Event / Punta Cana, Dominican Republic, November 7-11 2021.

Zhao, X., Zhang, Z., Zhang, Z., Wu, L., Jin, J., Zhou, Y., Jin, R., Dou, D., and Yan, D. Expressive 1-lipschitz neural networks for robust multiple graph learning against adversarial attacks. In *Proceedings of the 38th International Conference on Machine Learning, (ICML'21)*, pp. 12719–12735, Virtual Event, July 18-24 2021.

Zhou, Y. and Liu, L. Approximate deep network embedding for mining large-scale graphs. In *Proceedings of the 2019 IEEE International Conference on Cognitive Machine Intelligence (CogMI'19)*, pp. 53–60, Los Angeles, CA, December 12-14 2019.

Zhou, Y., Ren, J., Dou, D., Jin, R., Zheng, J., and Lee, K. Robust meta network embedding against adversarial attacks. In *Proceedings of the 20th IEEE International Conference on Data Mining (ICDM'20)*, pp. 1448–1453, Sorrento, Italy, November 17-20 2020.

Zhou, Y., Ren, J., Jin, R., Zhang, Z., Zheng, J., Jiang, Z., Yan, D., and Dou, D. Unsupervised adversarial network alignment with reinforcement learning. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(3): 1–29, 2021a.

Zhou, Y., Zhang, Z., Wu, S., Sheng, V., Han, X., Zhang, Z., and Jin, R. Robust network alignment via attack signal scaling and adversarial perturbation elimination. In *Proceedings of the 30th Web Conference (WWW'21)*, pp. 3884–3895, Virtual Event / Ljubljana, Slovenia, April 19-23 2021b.

Zhu, Z., Luo, T., and Liu, Y. The rich get richer: Disparate impact of semi-supervised learning. In *10th International Conference on Learning Representations, ICLR 2022, Online, April 25-29, 2022, Conference Track Proceedings*, 2022.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. A comprehensive survey on transfer learning. *Proc. IEEE*, 109(1):43–76, 2021.

A. Appendix

A.1. Theoretical Proof

Lemma 1. $\hat{N}$ *is Fréchet differentiable. Therefore,* $\hat{N}(\mathbb{R}^m) \subset L^p$ *is a smooth manifold.*

Proof. Recall the definition of Fréchet derivative in Definition 7. Generally speaking, the proof of an operator being Fréchet differentiable consists of two steps: 1) Find a candidate $P(x)$ for the Fréchet derivative; 2) Show that $L(x)$ satisfies the limit condition in the definition. We first show $\hat{N}(W)$ is Fréchet differentiable. Without loss of generality, we prove it for $\sigma = 1$. By the change of variable, we rewrite $\hat{\phi}(W)$ as

$$\hat{N}(W)(\cdot) = \frac{1}{(2\pi)^{m/2}} \int_{\mathbb{R}^m} N(\Delta)(\cdot) e^{-\frac{\|W-\Delta\|_2^2}{2}} d\Delta \tag{21}$$

We define a linear operator $P(W)(\delta) : \mathbb{R}^m \to L^p$ by

$$P(W)(\delta)(\cdot) = -\frac{1}{(2\pi)^{m/2}} \int_{\mathbb{R}^m} N(\Delta)(\cdot) e^{-\frac{\|W-\Delta\|_2^2}{2}} (W-\Delta) \cdot \delta d\Delta \tag{22}$$

where $P(W)$ is the candidate for the Fréchet derivative of $\hat{N}(W)$.

Next, we show that $P(W)$ is indeed the Fréchet derivative. We compute

$$\begin{aligned}
&\|\hat{N}(W+\delta)(\cdot) - \hat{N}(W)(\cdot) - P(W)\delta(\cdot)\|_{L^p(\Omega)} \\
=&\left\| \frac{1}{(2\pi)^{m/2}} \int_{\mathbb{R}^m} N(\Delta)(\cdot) \left(e^{-\frac{\|W+\delta-\Delta\|_2^2}{2}} - e^{-\frac{\|W-\Delta\|_2^2}{2}} + e^{-\frac{\|W-\Delta\|_2^2}{2}} (W-\Delta)\cdot\delta \right) d\Delta \right\|_{L^p(\Omega)}
\end{aligned} \tag{23}$$

By direct calculations, we have

$$\frac{d}{ds} e^{-\frac{\|W+s\delta-\Delta\|_2^2}{2}} = -e^{-\frac{\|W+s\delta-\Delta\|_2^2}{2}} (W+s\delta-\Delta) \cdot \delta \tag{24}$$

By integrating the above identity from 0 to 1, we obtain

$$e^{-\frac{\|W+\delta-\Delta\|_2^2}{2}} - e^{-\frac{\|W-\Delta\|_2^2}{2}} = -\int_0^1 e^{-\frac{\|W+s\delta-\Delta\|_2^2}{2}} (W+s\delta-\Delta) \cdot \delta ds \tag{25}$$

By plugging Eq.(25) into Eq.(23), we have

$$\begin{aligned}
&\|\hat{N}(W+\delta)(\cdot) - \hat{N}(W)(\cdot) - P(W)\delta(\cdot)\|_{L^p(\Omega)} \\
=&\left\| \frac{1}{(2\pi)^{m/2}} \int_{\mathbb{R}^m} N(\Delta)(\cdot) \left(\int_0^1 e^{-\frac{\|W+s\delta-\Delta\|_2^2}{2}} (W+s\delta-\Delta) \cdot \delta - e^{-\frac{\|W-\Delta\|_2^2}{2}} (W-\Delta) \cdot \delta ds \right) d\Delta \right\|_{L^p(\Omega)} \\
=&\left\| \frac{1}{(2\pi)^{m/2}} \int_{\mathbb{R}^m} N(\Delta)(\cdot) \left(\int_0^1 (e^{-\frac{\|W+s\delta-\Delta\|_2^2}{2}} - e^{-\frac{\|W-\Delta\|_2^2}{2}})(W-\Delta) \cdot \delta + e^{-\frac{\|W+s\delta-\Delta\|_2^2}{2}} s\delta^2 ds \right) d\Delta \right\|_{L^p(\Omega)}
\end{aligned} \tag{26}$$

Based on the Fubini's Theorem, we derive that

$$\begin{aligned}
&\|\hat{N}(W+\delta)(\cdot) - \hat{N}(W)(\cdot) - P(W)\delta(\cdot)\|_{L^p(\Omega)} \\
=&\left\| \frac{1}{(2\pi)^{m/2}} \int_0^1 \left(\int_{\mathbb{R}^m} N(\Delta)(\cdot)(e^{-\frac{\|W+s\delta-\Delta\|_2^2}{2}} - e^{-\frac{\|W-\Delta\|_2^2}{2}})(W-\Delta) \cdot \delta + e^{-\frac{\|W+s\delta-\Delta\|_2^2}{2}} s\delta^2 d\Delta \right) ds \right\|_{L^p(\Omega)}
\end{aligned} \tag{27}$$

It is clear that

$$\frac{1}{(2\pi)^{m/2}} \int_{\mathbb{R}^m} e^{-\frac{\|W+s\delta-\Delta\|_2^2}{2}} \delta^2 d\Delta = \frac{1}{(2\pi)^{m/2}} \int_{\mathbb{R}^m} e^{-\frac{\|\Delta\|_2^2}{2}} \delta^2 d\Delta = \|\delta\|_2^2 \tag{28}$$

Notice that $|N(\Delta)(x)| \leq 1$ for any Δ and x. Therefore,

$$\left\|\frac{1}{(2\pi)^{m/2}} \int_0^1 \int_{\mathbb{R}^m} N(\Delta)(\cdot) e^{-\frac{\|W+s\delta-\Delta\|_2^2}{2}} s\delta^2 d\Delta ds\right\|_{L^p(\Omega)} \leq \|1\|_{L^p(\Omega)} \|\delta\|_2^2 \tag{29}$$

Moreover, similar to Eq.(25), one can check that

$$e^{-\frac{\|W+s\delta-\Delta\|_2^2}{2}} - e^{-\frac{\|W-\Delta\|_2^2}{2}} = -\int_0^1 e^{-\frac{\|W+st\delta-\Delta\|_2^2}{2}} (W+s\delta-\Delta) \cdot \delta dt \tag{30}$$

It follows that

$$\begin{aligned}
&\left\|\frac{1}{(2\pi)^{m/2}} \int_0^1 \int_{\mathbb{R}^m} N(\Delta)(\cdot) (e^{-\frac{\|W+s\delta-\Delta\|_2^2}{2}} - e^{-\frac{\|W-\Delta\|_2^2}{2}})(W-\Delta) \cdot \delta d\Delta ds\right\|_{L^p(\Omega)} \\
=&\left\|\frac{1}{(2\pi)^{m/2}} \int_0^1 \int_0^1 \int_{\mathbb{R}^m} N(\Delta)(\cdot) e^{-\frac{\|W+st\delta-\Delta\|_2^2}{2}} (W+s\delta-\Delta) \cdot \delta(W-\Delta) \cdot \delta d\Delta ds dt\right\|_{L^p(\Omega)} \\
\leq&\left\|\frac{1}{(2\pi)^{m/2}} \int_0^1 \int_0^1 \int_{\mathbb{R}^m} e^{-\frac{\|W+st\delta-\Delta\|_2^2}{2}} \|W+s\delta-\Delta\|_2 \|W-\Delta\|_2 \|\delta\|^2 d\Delta ds dt\right\|_{L^p(\Omega)}
\end{aligned} \tag{31}$$

Based on Dominant Convergence Theorem, as $\delta \to 0$, we have

$$\int_{\mathbb{R}^m} e^{-\frac{\|W+st\delta-\Delta\|_2^2}{2}} \|W+s\delta-\Delta\|_2 \|W-\Delta\|_2 \|\delta\|^2 d\Delta \to \int_{\mathbb{R}^m} e^{-\frac{\|W-\Delta\|_2^2}{2}} \|W-\Delta\|_2^2 d\Delta \tag{32}$$

Therefore, as $\delta \to 0$, we have

$$\begin{aligned}
&\left\|\frac{1}{(2\pi)^{m/2}} \int_0^1 \int_{\mathbb{R}^m} N(\Delta)(\cdot) (e^{-\frac{\|W+s\delta-\Delta\|_2^2}{2}} - e^{-\frac{\|W-\Delta\|_2^2}{2}})(W-\Delta) \cdot \delta d\Delta ds\right\|_{L^p(\Omega)} \\
\leq&\left\|\frac{1}{(2\pi)^{m/2}} \int_{\mathbb{R}^m} e^{-\frac{\|W'\|_2^2}{2}} \|W'\|_2^2 dW'\right\|_{L^p(\Omega)} \|\delta\|^2
\end{aligned} \tag{33}$$

Combining Eq.(29) and Eq.(33), we obtain

$$\hat{N}(W+\delta)(\cdot) - \hat{N}(W)(\cdot) - P(W)\delta(\cdot)\|_{L^p(\Omega)} = O(\delta^2) \tag{34}$$

This implies that $P(W)$ is Fréchet derivative of $\hat{N}$. By similar arguments, one can show that $\hat{N}(W)$ is infinitely Fréchet differentiable. Note that $\hat{N}(W)$ can be viewed as a global chart for the manifold $\hat{N}(\mathbb{R}^m)$. Therefore, $\hat{N}(\mathbb{R}^m)$ is a smooth manifold. $\qquad\square$

Lemma 2. *For any $W_1, W_2 \in \mathbb{R}^m$ satisfying $\|W_1 - W_2\|_2 \leq d$, it holds the that*

$$\begin{aligned}
&|\Omega|^{-1} \|\hat{N}(W_1)(x) - \hat{N}(W_2)(x)\|_{L^p(\Omega)} \leq \frac{d}{\sqrt{2\pi}\sigma} \\
&\textit{if } 1 \leq p < \infty, \\
&\|\hat{N}(W_1)(x) - \hat{N}(W_2)(x)\|_{L^\infty(\Omega)} \leq \frac{d}{\sqrt{2\pi}\sigma}
\end{aligned} \tag{35}$$

for any $\Omega \subset \mathbb{R}^n$.

Proof. Since $\mathcal{N}(0, \sigma^2 I) = \sigma^{-m}\mathcal{N}(0, I)$, it suffices to prove the lemma for $\sigma = 1$. Rewrite $\hat{\phi}(W)$ as

$$\hat{N}(W)(\cdot) = \frac{1}{(2\pi)^{m/2}} \int_{\mathbb{R}^m} N(\Delta)(\cdot) e^{-\frac{\|W-\Delta\|_2^2}{2}} d\Delta \tag{36}$$

By direct calculations, it holds that

$$
\begin{aligned}
&\|\hat{N}(W_1)(\cdot) - \hat{N}(W_2)(\cdot)\|_{L^p(\Omega)} \\
&= \left\| \frac{1}{(2\pi)^{m/2}} \int_{\mathbb{R}^m} N(\Delta)(\cdot) \left(e^{-\frac{\|W_1 - \Delta\|_2^2}{2}} - e^{-\frac{\|W_2 - \Delta\|_2^2}{2}} \right) d\Delta \right\|_{L^p(\Omega)}
\end{aligned}
\tag{37}
$$

One can check that

$$
\frac{d}{ds} e^{-\frac{\|W_2 - \Delta + s(W_1 - W_2)\|_2^2}{2}} = -e^{-\frac{\|W_2 - \Delta + s(W_1 - W_2)\|_2^2}{2}} (W_2 - \Delta + s(W_1 - W_2)) \cdot (W_1 - W_2)
\tag{38}
$$

By integrating the above identity from 0 to 1, we obtain

$$
e^{-\frac{\|W_1 - \Delta\|_2^2}{2}} - e^{-\frac{\|W_2 - \Delta\|_2^2}{2}} = -\int_0^1 e^{-\frac{\|W_2 - \Delta + s(W_1 - W_2)\|_2^2}{2}} (W_2 - \Delta + s(W_1 - W_2)) \cdot (W_1 - W_2) ds
\tag{39}
$$

By plugging Eq.(39) into Eq.(37), we have

$$
\begin{aligned}
&\|\hat{N}(W_1)(\cdot) - \hat{N}(W_2)(\cdot)\|_{L^p(\Omega)} \\
&= \left\| \frac{1}{(2\pi)^{m/2}} \int_{\mathbb{R}^m} N(\Delta)(\cdot) \left(\int_0^1 e^{-\frac{\|W_2 - \Delta + s(W_1 - W_2)\|_2^2}{2}} (W_2 - \Delta + s(W_1 - W_2)) \cdot (W_1 - W_2) ds \right) d\Delta \right\|_{L^p(\Omega)} \\
&= \left\| \frac{1}{(2\pi)^{m/2}} \int_{\mathbb{R}^m} \int_0^1 N(\Delta)(\cdot) e^{-\frac{\|W_2 - \Delta + s(W_1 - W_2)\|_2^2}{2}} (W_2 - \Delta + s(W_1 - W_2)) \cdot (W_1 - W_2) ds d\Delta \right\|_{L^p(\Omega)}
\end{aligned}
\tag{40}
$$

By applying Fubini's Theorem to the above inequality, we have

$$
\begin{aligned}
&\|\hat{N}(W_1)(\cdot) - \hat{N}(W_2)(\cdot)\|_{L^p(\Omega)} \\
&= \left\| \frac{1}{(2\pi)^{m/2}} \int_0^1 \int_{\mathbb{R}^m} N(\Delta)(\cdot) e^{-\frac{\|W_2 - \Delta + s(W_1 - W_2)\|_2^2}{2}} (W_2 - \Delta + s(W_1 - W_2)) \cdot (W_1 - W_2) d\Delta ds \right\|_{L^p(\Omega)}
\end{aligned}
\tag{41}
$$

For convenience, we let $\widetilde{W} = W_2 + s(W_1 - W_2)$ and $\bar{W} = W_1 - W_2$. Then, we rewrite

$$
\begin{aligned}
&\|\hat{N}(W_1)(\cdot) - \hat{N}(W_2)(\cdot)\|_{L^p(\Omega)} \\
&= \left\| \frac{1}{(2\pi)^{m/2}} \int_0^1 \int_{\mathbb{R}^m} N(\Delta)(\cdot) e^{\frac{-\|\widetilde{W} - \Delta\|_2^2}{2}} (\widetilde{W} - \Delta) \cdot \bar{W} d\Delta ds \right\|_{L^p(\Omega)} \\
&= \left\| \frac{1}{(2\pi)^{m/2}} \int_0^1 \int_{\mathbb{R}^m} N(\Delta)(\cdot) e^{\frac{-\|\widetilde{W} - \Delta\|_2^2}{2}} (\widetilde{W} - \Delta) \cdot \frac{\bar{W}}{\|\bar{W}\|_2} d\Delta ds \right\|_{L^p(\Omega)} \|\bar{W}\|_2
\end{aligned}
\tag{42}
$$

Notice that $0 \leq N(W)(x) \leq 1$ for any W and x and $\frac{\bar{W}}{\|\bar{W}\|_2}$ is a unit vector. By using standard Gaussian integrals, for any unit vector u, we obtain

$$
\begin{aligned}
&\left\| \frac{1}{(2\pi)^{m/2}} \int_0^1 \int_{\mathbb{R}^m} N(\Delta)(\cdot) e^{\frac{-\|\widetilde{W} - \Delta\|_2^2}{2}} (\widetilde{W} - \Delta) \cdot u d\Delta ds \right\|_{L^p(\Omega)} \\
&\leq \left\| \frac{1}{(2\pi)^{m/2}} \int_0^1 \int_{\mathbb{R}^m} e^{\frac{-\|\widetilde{W} - \Delta\|_2^2}{2}} |(\widetilde{W} - \Delta) \cdot u| d\Delta ds \right\|_{L^p(\Omega)} \\
&= \left\| \frac{1}{(2\pi)^{m/2}} \int_0^1 \int_{\mathbb{R}^m} e^{\frac{-\|\Delta\|_2^2}{2}} |\Delta \cdot u| d\Delta ds \right\|_{L^p(\Omega)}
\end{aligned}
\tag{43}
$$

We rotate the coordinate system such that the unit vector u coincides with any coordinate axis. Let $u, v_1, \cdots, v_{m-1}$ be the standard basis of the rotated coordinate system. In other words, for any $\Delta \in \mathbb{R}^m$, we decompose Δ into

$$\Delta = (\Delta \cdot u)u + \sum_{l=1}^{m-1} (\Delta \cdot v_{m-1})v_{m-1} \tag{44}$$

Since the Jacobian of rotation is 1, by utilizing standard Gaussian integrals, one can check that

$$\int_{\mathbb{R}^m} e^{\frac{-\|\Delta\|_2^2}{2}} |\Delta \cdot u| d\Delta = \int_{\mathbb{R}^m} e^{\frac{-\sum_{l=1}^{m-1} |\Delta \cdot v_l|_2^2}{2}} e^{\frac{-|\Delta \cdot u|_2^2}{2}} |\Delta \cdot u| d\Delta$$
$$= \int_{\mathbb{R}^m} e^{\frac{-\sum_{l=1}^{m-1} \tau_l^2}{2}} e^{\frac{-|\tau_m|^2}{2}} d\tau_1 \cdots d\tau_m = 2(2\pi)^{\frac{m-1}{2}} \tag{45}$$

Consequently, we have

$$\left\| \frac{1}{(2\pi)^{m/2}} \int_0^1 \int_{\mathbb{R}^m} N(\Delta)(\cdot) e^{\frac{-\|\widetilde{W} - \Delta\|_2^2}{2}} (\widetilde{W} - \Delta) \cdot u d\Delta ds \right\|_{L^p(\Omega)} = \left\| \frac{1}{(2\pi)^{1/2}} \right\|_{L^p(\Omega)} \tag{46}$$

Therefore, it follows that for $1 \le p < \infty$,

$$\|\hat{N}(W_1)(\cdot) - \hat{N}(W_2)(\cdot)\|_{L^p(\Omega)} \le \frac{1}{(2\pi)^{1/2}} |\Omega| \|W_1 - W_2\|_2 \tag{47}$$

$$\|\hat{N}(W_1)(\cdot) - \hat{N}(W_2)(\cdot)\|_{L^\infty(\Omega)} \le \frac{1}{(2\pi)^{1/2}} \|W_1 - W_2\|_2 \tag{48}$$

The proof is finished. $\qquad \square$

Theorem 2. *Let $\hat{N}(W_k^*)$ be the optimal individual classifier for group G_k in the hypothesis space $\hat{\mathcal{H}}$ $(k = 1, 2, \cdots, K)$. Let $W^* = \frac{W_1^* + \cdots + W_K^*}{K}$ and $\hat{N}(W^*)$ be the overall fair classifier. If $\max\limits_{1 \le k, l \le K} \|W_k - W_l\|_2 = d$, then for any $\Omega \subset \mathbb{R}^n$, it holds that*

$$|\Omega_k|^{-1} \|\hat{N}(W^*)(x) - \hat{N}(W_k^*)(x)\|_{L^p(\Omega)} \le \frac{(K-1)d}{\sqrt{2\pi}K\sigma}$$
$$\text{if } 1 \le p < \infty, \tag{49}$$
$$\|\hat{N}(W^*)(x) - \hat{N}(W_k^*)(x)\|_{L^\infty(\Omega_k)} \le \frac{(K-1)d}{\sqrt{2\pi}K\sigma}$$

Therefore, $\hat{N}(W^*)(x)$ has certified $\frac{(K-1)d}{\sqrt{2\pi}K\sigma}$-fairness in L^p for any $1 \le p \le \infty$.

Proof. By direct calculations, we have that for any $k = 1, \cdots K$, $W^* - W_k^* = \frac{\sum_{l \ne k} W^* - W_k^*}{K}$, which implies that $\|W^* - W_k^*\| \le \frac{(K-1)d}{K}$. Then the desired results follows directly from Lemma 2. $\qquad \square$

Lemma 3. *For any $x \in \mathbb{R}^n$, $\hat{N}(W)(x) \to N(W)(x)$ as $\sigma \to 0$.*

Proof. Through a change of variable $\Delta = \sigma\tau$, we rewrite

$$\hat{N}(W)(x) = \frac{1}{(2\pi)^{m/2}} \int_{\mathbb{R}^m} N(W + \sigma\tau)(x) e^{-\frac{\|\tau\|_2^2}{2}} d\tau \tag{50}$$

Table 4: Accuracy and Fairness with *Race* Attribute

Dataset	Adult			COMPAS		
Protected Attribute	Race			Race		
Metric	**Acc.**	Δ_{DP}	Δ_{EO}	**Acc.**	Δ_{DP}	Δ_{EO}
ApxFair	0.72	0.21	1.19	0.6	0.17	0.22
ARL	**0.84**	0.18	0.19	0.5	0.05	0.17
Fair Mixup	0.83	0.09	0.14	0.65	**0.03**	0.05
FairBatch	**0.84**	0.15	0.35	0.65	0.31	0.59
FRS	0.65	0.15	0.17	0.31	0.04	0.13
Implicit	0.67	0.18	0.2	0.6	0.11	0.3
Group-Fair	0.73	0.12	**0.12**	0.49	0.10	0.15
FCRL	0.81	0.12	0.16	0.49	0.08	0.16
DLR	0.76	0.09	0.18	0.41	0.12	0.21
FairSmooth	**0.84**	**0.04**	0.13	**0.66**	**0.03**	**0.01**

It is clear that $|N(W + \sigma\tau)(x)e^{-\frac{\|\tau\|_2^2}{2}}| \leq e^{-\frac{\|\tau\|_2^2}{2}}$ and $e^{-\frac{\|\tau\|_2^2}{2}}$ is integrable. Moreover, by the continuity of $N(W + \sigma\tau)(x)$, one has $N(W + \sigma\tau)(x) \to N(W)(x)$ as $\sigma \to 0$. Therefore, by Dominated Convergence Theorem, we obtain that as $\sigma \to 0$, it holds

$$\frac{1}{(2\pi)^{m/2}} \int_{\mathbb{R}^m} N(W + \sigma\tau)(x)e^{-\frac{\|\tau\|_2^2}{2}} d\tau \to \frac{1}{(2\pi)^{m/2}} \int_{\mathbb{R}^m} N(W)(x)e^{-\frac{\|\tau\|_2^2}{2}} d\tau \tag{51}$$

This completes the proof. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad\square$

A.2. Additional Experiments

In this section, we conduct more experiments to validate the accuracy and fairness of our proposed FairSmoothmethod and evaluate the sensitivity of various parameters in our input-agnostic certified group fairness framework for the fair classification task.

Performance with *Race* as protected attribute. Table 4 presents the *Accuracy*, Δ_{DP}, and Δ_{EO} scores of ten fair classification algorithms over two datasets of Adult and COMPAS. Similar trends are observed for the classification performance comparison: FairSmooth achieves the highest *Accuracy* values (> 0.66), the smallest Δ_{DP} scores (< 0.04), and the lowest Δ_{EO} values (< 0.15) respectively, which are obviously better than all other methods. Especially, as shown the experiment on Adult, compared to the best competitors among ten group fairness algorithms, the Δ_{DP} and Δ_{EO} scores achieved by FairSmooth averagely achieves 10.0% and 36.4% improvement respectively. This demonstrates that the integration of the Gaussian parameter smoothing for providing a tight fairness guarantee and the parameter disparity term for helping recover the most accurate model for each group is able to make the classification results achieved by our input-agnostic certified group fairness method be accurate as well as fair.

Performance with varying test data partitions on COMPAS. Figure 4 exhibits the fair classification performance with different test data on Adult. We have observed similar trends: the performance (both accuracy and fairness) achieved by our FairSmooth is insensitive to test data distributions. This is because the smooth classifiers are reformulated as output functions of a Nemytskii operator, which induces a Fréchet differentiable smooth manifold. The smooth manifold has a global Lipschitz constant that is independent of the domain of the input data, which derives the input-agnostic certified group fairness. In addition, FairSmooth still outperforms other fair classification methods in most experiments. The exceptional performance of FairSmooth over different datasets implies that our input-agnostic certified group fairness technique has great potential to achieve the superior certified fairness guarantees for any input data, which is desirable in practice.

Impact of σ and α on COMPAS. Figure 5 presents the influence of the standard deviation σ in Gaussian distribution and the weight α of parameter disparity term in our FairSmooth model over the COMPAS dataset. We have witnessed the accuracy (or fairness) curves initially increase (or decrease) quickly and then become stable or even drop (or raise) when the parameters continuously increase. Initially, a large σ can help leverage the strength of Gaussian parameter smoothing to produce a fair classifier. Later on, when σ continues to increase and goes beyond some thresholds, too much Gaussian
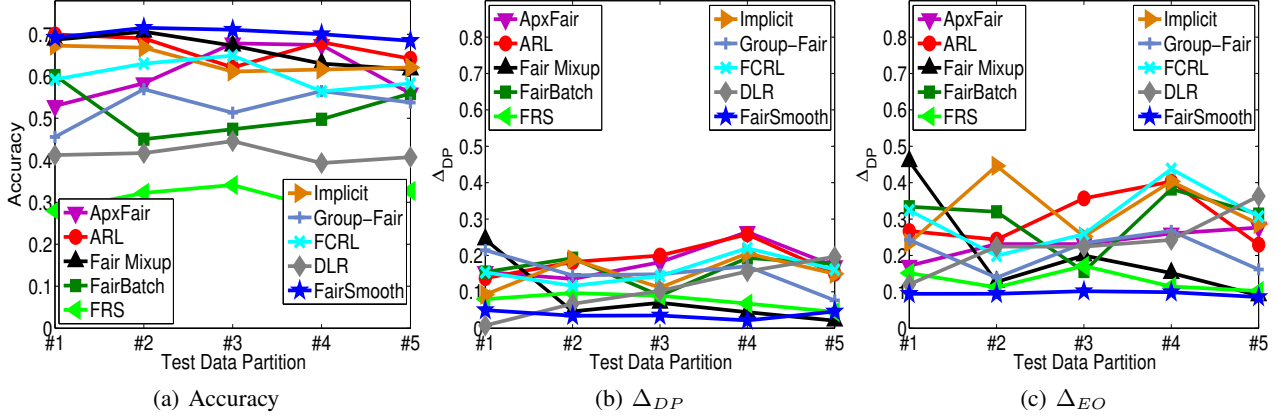
(a) Accuracy

(b) Δ_{DP}

(c) Δ_{EO}

Figure 4: Performance with Varying Test Data Partitions on COMPAS
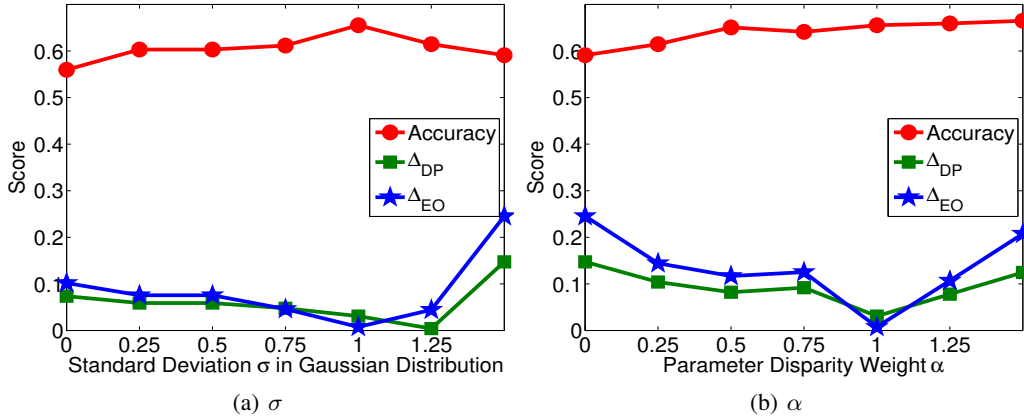

(a) σ

(b) α

Figure 5: Performance with Varying Parameters on COMPAS

noise may ruin the prediction by the smooth classifier. In addition, α serves as a tradeoff hyperparameter to well balance the fairness and accuracy. Therefore, choosing a suitable parameter is important for the fair classifiers.

A.3. Experimental Details

Environment. Our experiments were conducted on a compute server running on Red Hat Enterprise Linux 7.2 with 2 CPUs of Intel Xeon E5-2650 v4 (at 2.66 GHz) and 8 GPUs of NVIDIA GeForce GTX 2080 Ti (with 11GB of GDDR6 on a 352-bit memory bus and memory bandwidth in the neighborhood of 620GB/s), 256GB of RAM, and 1TB of HDD. Overall, our experiments took about 2 days in a shared resource setting. We expect that a consumer-grade single-GPU machine (e.g., with a 1080 Ti GPU) could complete our full set of experiments in around tens of hours, if its full resources were dedicated. The codes were implemented in Python 3.7.3 and PyTorch 1.0.14. We also employ Numpy 1.16.4 and Scipy 1.3.0 in the implementation. Since the datasets used are all public datasets and the hyperparameter settings are explicitly described, our experiments can be easily reproduced on top of a GPU server. We promise to release our open-source codes on GitHub and maintain a project website with detailed documentation for long-term access by other researchers and end-users after the paper is accepted.

Datasets. The COMPAS dataset was introduced by ProPublica. It contains data from the US criminal justice system and was obtained by a public records request. The dataset contains personal information. To mitigate negative side effects, we delete the name, first, last and dob (date of birth) entries from the dataset before processing it further. We then exclude entries that do not fit the problem setting of predicting two year recidivism, following the steps of the original analysis (ProPublica). Specifically, this means keeping only cases from Broward county, Florida, for which data has been entered within 30 days of the arrest. Traffic offenses and cases with insufficient information are also excluded. This steps leave 6,172 examples out of the original 7,214 cases. The Adult dataset is available in the UCI data repository as well as multiple other online

sources (UCI). We use the dataset in unmodified form, except for binning some of the feature values.

Implementation. For six regular group fairness models of ApxFair [1], ARL [2], Fair Mixup [3], FairBatch [4], fair-robust-selection (FRS) [5], and Implicit [6], we used the open-source implementation and default parameter settings by the original authors for our experiments. For three provable group fairness guarantee approaches of Group-Fair [7], FCRL [8], and DLR [9], we also utilized the same model architecture as the official implementation provided by the original authors and used the same datasets to validate the fairness of these fair classification models in all experiments. All hyperparameters are standard values from reference codes or prior works. The above open-source codes from the GitHub are licensed under the MIT License, which only requires preservation of copyright and license notices and includes the permissions of commercial use, modification, distribution, and private use.

For our proposed input-agnostic certified group fairness algorithm, we performed hyperparameter selection by performing a parameter sweep on sample numbers $N \in \{500, 1,000, 5,000, 10,000, 50,000, 100,000\}$ in the integral calculation, parameter disparity weight $\alpha \in \{0.25, 0.5, 0.75, 1, 1.25, 1.5\}$ in the FairSmooth-S model, training epochs of the smooth classifiers $\in \{10, 20, 40, 80, 160, 320\}$, batch size for training the smooth classifiers $\in \{16, 32, 48, 64, 128, 256\}$, and learning rate $\in \{0.0005, 0.001, 0.005, 0.01, 0.05, 0.1\}$. In this paper we smooth the base classifier with Gaussian parameter augmentation (i.e., perturbations) at variance σ^2. We select σ from a range $\sigma \in \{0.25, 0.5, 1, 1.25, 1.5, 1.75, 2\}$. We select the best parameters over 50 epochs of training and evaluate the model at test time. After the hyperparameter selection, the model was trained for 320 iterations, with a batch size of 128, and a learning rate of 0.05.

---

[1]https://github.com/essdeee/Ensuring-Fairness-Beyond-the-Training-Data

[2]https://github.com/google-research/google-research/tree/master/group_agnostic_fairness

[3]https://github.com/chingyaoc/fair-mixup

[4]https://github.com/yuji-roh/fairbatch

[5]https://github.com/yuji-roh/fair-robust-selection

[6]https://openreview.net/forum?id=pkh8bwJbUbL

[7]https://github.com/vijaykeswani/FairClassification

[8]https://github.com/umgupta/fairness-via-contrastive-estimation

[9]https://github.com/vijaykeswani/Noisy-Fair-Classification