Individual Fairness Revisited: Transferring Techniques from Adversarial Robustness

Samuel Yeom and **Matt Fredrikson**
Carnegie Mellon University
{syeom, mfredrik}@cs.cmu.edu

Abstract

We turn the definition of individual fairness on its head—rather than ascertaining the fairness of a model given a predetermined metric, we find a metric for a given model that satisfies individual fairness. This can facilitate the discussion on the fairness of a model, addressing the issue that it may be difficult to specify a priori a suitable metric. Our contributions are twofold: First, we introduce the definition of a *minimal* metric and characterize the behavior of models in terms of minimal metrics. Second, for more complicated models, we apply the mechanism of randomized smoothing from adversarial robustness to make them individually fair under a given weighted L^p metric. Our experiments show that adapting the minimal metrics of linear models to more complicated neural networks can lead to meaningful and interpretable fairness guarantees at little cost to utility.

1 Introduction

When machine learning models are deployed to make predictions about people, it is important that the model treats individuals fairly. *Individual fairness* [Dwork *et al.*, 2012] captures the notion that similar people should be treated similarly by imposing a continuity requirement on models. However, this raises the difficult societal question of how to define which people are "similar".

We start in Section 4 from the insight that it may be easier to determine whether a given similarity metric is reasonable than it is to construct one from scratch. Thus, rather than imposing individual fairness with a predetermined similarity metric, we find a metric that corresponds to the behavior of a given model, which can then guide the discussion on whether the model is fair. To facilitate this, we introduce the notion of a *minimal* fairness metric, and show that in many cases there exists a unique metric that best characterizes the behavior of a given model for this purpose.

In Section 5, we deal with more complicated models, such as deep neural networks, whose minimal metrics are not easily computable. We show that we can make *any* model provably individually fair by post-processing it with *randomized smoothing* [Cohen *et al.*, 2019] to impose a given weighted L^p metric. As randomized smoothing was originally applied as a defense against adversarial examples, our result brings to light the connection between individual fairness and adversarial robustness. However, our theorems are in a sense stronger because individual fairness is a uniform requirement that applies to all points in the input space, whereas the certified threshold of Cohen *et al.* is a function of the input point. Our Laplace and Gaussian smoothing mechanisms are versatile in that they can make a model provably individually fair under any given weighted L^p metric, and we show the minimality of this metric for the smoothed model to argue that we do not add more noise than is necessary.

Finally, our experiments combine the two main elements of our paper—we smooth neural networks to be individually fair under a metric that is proportional to the minimal metrics of linear models trained on the same datasets. Our results on four real datasets show that the neural networks smoothed with Gaussian noise in particular are often approximately as accurate as the original models. Moreover, we can achieve models with similar favorable individual fairness guarantees to those of linear models while still enjoying the increased predictive accuracy enabled by the neural network.

2 Related Work

Dwork *et al.* [2012] introduced the definition of individual fairness, which contrasts with group-based notions of fairness [Hardt *et al.*, 2016; Zafar *et al.*, 2017] that require demographic groups to be treated similarly on average. Motivated in part by group fairness, Zemel *et al.* [2013] learn a representation of the data that excludes information about a protected attribute, such as race or gender, whose use is often legally prohibited. This work has spurred more research on fair representations [Calmon *et al.*, 2017; Madras *et al.*, 2018; Tan *et al.*, 2019], and the resulting representations implicitly define a similarity metric. However, unlike the weighted L^p metrics that we use, these metrics are harder for humans to interpret and are primarily designed to attain group fairness.

Others approximate individual fairness based on a limited number of oracle queries, which represent human judgments, about whether pair of individuals is similar. Gillen *et al.* [2018] attempt to learn a similarity metric that is consistent with the human judgments in the setting of online linear contextual bandits. In a more general setting, Ilvento [2019] derives an approximate metric using comparison queries that

ask which of two individuals a given third individual is more similar to. Finally, Jung *et al.* [2019] apply constrained optimization directly without assuming that the human judgments are consistent with a metric.

By contrast, we post-process a model using randomized smoothing to provably ensure individual fairness. Cohen *et al.* [2019] previously analyzed randomized smoothing in the context of adversarial robustness. In the context of fairness, most post-processing approaches [Hardt *et al.*, 2016; Canetti *et al.*, 2019] do not take individual fairness into account, and although Lohia *et al.* [2019] consider individual fairness, they define two individuals to be similar if and only if they differ only in the pre-specified protected attribute.

3 Background

In this section, we present the definitions and notation that we will use throughout the paper.

Definition 1 (Distance metric)**.** *A nonnegative function $D : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a* distance metric *in $\mathcal{X}$ if it satisfies the following three conditions: nonnegativity, symmetry, and triangle inequality.*

In common mathematical usage, Definition 1 is a pseudometric, and metrics must also satisfy the condition that $D(x_1, x_2) = 0$ if and only if $x_1 = x_2$. However, throughout this paper we will refer to pseudometrics as metrics, following the convention in the field of metric learning.

One commonly used family of metrics is the standard L^p metric, which is defined over $\mathbb{R}^d$. In this paper, we consider a more general family of metrics that allows each coordinate to be weighted differently.

Definition 2 (Weighted L^p metric)**.** *The* weighted L^p metric*, with $p \geq 1$ and weights $w_i \geq 0$, is a distance metric in $\mathbb{R}^d$ that is defined by the equation*

$$D(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sqrt[p]{\textstyle\sum_{i=1}^{d} w_i \cdot |x_{1i} - x_{2i}|^p}, \tag{1}$$

where x_{1i} and x_{2i} are the i-th coordinates of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, respectively.

We place the restriction that $p \geq 1$ because otherwise the function D does not satisfy the triangle inequality. When $w_i = 1$ for all i, we have the *standard L^p metric*.

Throughout this paper, we will use $\mathcal{X}$ and $\mathcal{Y}$ to denote a model's input and output spaces, respectively. Moreover, we will assume a distance metric $D_{\mathcal{Y}} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ that characterizes how close two points in the output space are.

Definition 3 (Individual fairness [Dwork *et al.*, 2012])**.** *A model $h : \mathcal{X} \to \mathcal{Y}$ is* individually fair *under metric $D_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ if, for all $x_1, x_2 \in \mathcal{X}$,*

$$D_{\mathcal{Y}}(h(x_1), h(x_2)) \leq D_{\mathcal{X}}(x_1, x_2). \tag{2}$$

Individual fairness captures the intuition that the model should not behave arbitrarily. In particular, it formalizes the notion that similar individuals should be treated similarly, i.e., given two individuals $x_1, x_2 \in \mathcal{X}$, if the distance $D_{\mathcal{X}}(x_1, x_2)$ between them is small, then the distance $D_{\mathcal{Y}}(h(x_1), h(x_2))$ between the outputs of the model on these individuals should also be small.

4 Minimal Distance Metric

One criticism of individual fairness is that it is difficult to apply in practice because it requires one to specify the metric $D_{\mathcal{X}}$ [Chouldechova and Roth, 2018]. The choice of a metric in $\mathcal{X}$ dictates which individuals should be considered similar, which is highly context-dependent and often controversial. Thus, we take a slightly different approach—rather than specifying a metric $D_{\mathcal{X}}$ and asking whether a model is individually fair under that metric, we find one metric under which the model is individually fair. Then, we can reason about whether the metric is appropriate for the task at hand.

However, there could be multiple metrics for which a model is individually fair. In fact, if $D_{\mathcal{X}}(x_1, x_2) \geq D'_{\mathcal{X}}(x_1, x_2)$ for all $x_1, x_2 \in \mathcal{X}$, then any model that is individually fair under $D'_{\mathcal{X}}$ is also fair under $D_{\mathcal{X}}$, as the metrics are simply upper bounds on the extent to which a model's outputs can vary. On the other hand, our goal is to characterize the behavior of a model, for which we need a *tight* upper bound. This notion of tightness is captured by the minimality of a distance metric, defined in Definition 4.

Definition 4 (Minimal distance metric)**.** *Let $\mathcal{M}$ be a set of distance metrics in $\mathcal{X}$. A metric $D_{\mathcal{X}} \in \mathcal{M}$ is* minimal *in $\mathcal{M}$ with respect to model $h : \mathcal{X} \to \mathcal{Y}$ if* (1) *h is individually fair under $D_{\mathcal{X}}$, and* (2) *there does not exist a different $D'_{\mathcal{X}} \in \mathcal{M}$ such that h is individually fair under $D'_{\mathcal{X}}$ and $D_{\mathcal{X}}(x_1, x_2) \geq D'_{\mathcal{X}}(x_1, x_2)$ for all $x_1, x_2 \in \mathcal{X}$.*

To see how one may reason about the minimal distance metric, consider a hiring model with a binary output that informs whether a given applicant should be hired. A natural $D_{\mathcal{Y}}$ in this setting is the 0-1 loss $D_{\mathcal{Y}}(y_1, y_2) = \mathbb{1}[y_1 \neq y_2]$. Then, if the hiring model satisfies individual fairness under a metric $D_{\mathcal{X}}$ such that $D_{\mathcal{X}}(x_1, x_2) = 0$ whenever x_1 and x_2 differ only in race, we can reason that it does not directly use race to discriminate.

We now present Theorem 1, which identifies the unique minimal metric among the set of all metrics.

Theorem 1. *Let $h : \mathcal{X} \to \mathcal{Y}$ be a model, and let $\mathcal{M}_{\text{all}}$ be the set of all metrics that satisfy the conditions in Definition 1. Then, the metric $D_{\mathcal{X}}$, defined as*

$$D_{\mathcal{X}}(x_1, x_2) = D_{\mathcal{Y}}(h(x_1), h(x_2)) \tag{3}$$

for all $x_1, x_2 \in \mathcal{X}$, is the unique minimal metric in $\mathcal{M}_{\text{all}}$ with respect to h.

Proof. We first prove that $D_{\mathcal{X}}$ is a minimal metric, and later we will prove that no other metric is minimal. Since we assume $D_{\mathcal{Y}}$ to be a metric, it easily follows that $D_{\mathcal{X}}$ is also a metric under Definition 1. Moreover, the equality in Eq. (2) always holds by our definition of $D_{\mathcal{X}}$, so h is individually fair under $D_{\mathcal{X}}$. Thus, it remains to show that there does not exist a different $D'_{\mathcal{X}} \in \mathcal{M}_{\text{all}}$ such that h is individually fair under $D'_{\mathcal{X}}$ and $D_{\mathcal{X}}(x_1, x_2) \geq D'_{\mathcal{X}}(x_1, x_2)$ for all $x_1, x_2 \in \mathcal{X}$.

Suppose such $D'_{\mathcal{X}}$ exists. Since $D'_{\mathcal{X}} \neq D_{\mathcal{X}}$, there must exist some $x_1, x_2 \in \mathcal{X}$ such that $D_{\mathcal{X}}(x_1, x_2) > D'_{\mathcal{X}}(x_1, x_2)$. This, combined with Eq. (3), contradicts our assumption that h is individually fair under $D'_{\mathcal{X}}$.

Now we prove that $D_{\mathcal{X}}$ is the unique minimal metric, arguing that $D'_{\mathcal{X}}$ cannot be minimal if $D'_{\mathcal{X}} \neq D_{\mathcal{X}}$. If there

exist $x_1, x_2 \in \mathcal{X}$ such that $D_{\mathcal{X}}(x_1, x_2) > D'_{\mathcal{X}}(x_1, x_2)$, then h is not individually fair under $D'_{\mathcal{X}}$. Thus, we must have $D'_{\mathcal{X}}(x_1, x_2) \geq D_{\mathcal{X}}(x_1, x_2)$ for all $x_1, x_2 \in \mathcal{X}$, but then h is individually fair under $D_{\mathcal{X}}$, so $D'_{\mathcal{X}}$ cannot be minimal. □

Theorem 1 shows that the minimal metric $D_{\mathcal{X}}$ in $\mathcal{M}_{\text{all}}$ is defined directly in terms of the model in question. Ideally, we want the minimal metric to be simpler than the model so that it can help us interpret and reason about the fairness of the model. Thus, in the rest of this paper we only consider weighted L^p metrics, which comprise a broad and interpretable family of metrics defined over $\mathbb{R}^d$.

With this set of metrics, we can no longer prove a theorem as general as Theorem 1, so we now prove a result for linear regression models. In this setting, we have $\mathcal{Y} = \mathbb{R}$, and the distance metric is simply the absolute value $D_{\mathcal{Y}}(y_1, y_2) = |y_1 - y_2|$. Theorem 2 identifies the weighted L^p metric that is uniquely minimal for a given linear regression model.

Theorem 2. *Let $h : \mathbb{R}^d \to \mathbb{R}$ be a linear regression model with coefficients $\beta_1, \ldots, \beta_d$, and let $\mathcal{M}_L$ be the set of all weighted L^p metrics. Then, the L^1 metric with weights $w_i = |\beta_i|$ is the unique minimal metric in $\mathcal{M}_L$ with respect to h.*

Proof. $D_{\mathcal{X}}$ is clearly in $\mathcal{M}_L$ by definition. To see that h is individually fair under $D_{\mathcal{X}}$, note that for all $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^d$

$$\begin{aligned} D_{\mathcal{Y}}(h(\boldsymbol{x}_1), h(\boldsymbol{x}_2)) &= |\textstyle\sum_{i=1}^{d} \beta_i (x_{1i} - x_{2i})| \\ &\leq \textstyle\sum_{i=1}^{d} |\beta_i (x_{1i} - x_{2i})| = D_{\mathcal{X}}(\boldsymbol{x}_1, \boldsymbol{x}_2). \end{aligned} \tag{4}$$

The rest of the proof closely mirrors the argument given in the proof of Theorem 1, so we only mention how the proofs differ. As in the proof of Theorem 1, we assume that there exists $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^d$ such that $D_{\mathcal{X}}(\boldsymbol{x}_1, \boldsymbol{x}_2) > D'_{\mathcal{X}}(\boldsymbol{x}_1, \boldsymbol{x}_2)$. Our goal is to show that $D'_{\mathcal{X}}$ is not individually fair, and for this proof we have the additional condition that $D'_{\mathcal{X}} \in \mathcal{M}_L$. However, it is not necessarily true that $D'_{\mathcal{X}}(\boldsymbol{x}_1, \boldsymbol{x}_2) < D_{\mathcal{Y}}(h(\boldsymbol{x}_1), h(\boldsymbol{x}_2))$, so we instead construct $\boldsymbol{x}'_2$ such that $D'_{\mathcal{X}}(\boldsymbol{x}_1, \boldsymbol{x}'_2) < D_{\mathcal{Y}}(h(\boldsymbol{x}_1), h(\boldsymbol{x}'_2))$.

Let $x'_{2i} = x_{1i} - \text{sgn}(\beta_i)|x_{1i} - x_{2i}|$. With Eq. (1), we can verify that $D_{\mathcal{X}}(\boldsymbol{x}_1, \boldsymbol{x}_2) = D_{\mathcal{X}}(\boldsymbol{x}_1, \boldsymbol{x}'_2)$ for any $D_{\mathcal{X}} \in \mathcal{M}_L$. Moreover, $\beta_i(x_{1i} - x'_{2i}) \geq 0$ for all i, so the equality in Eq. (4) holds if we replace $\boldsymbol{x}_2$ by $\boldsymbol{x}'_2$. Combining all of these relations, we arrive at the desired result:

$$\begin{aligned} D'_{\mathcal{X}}(\boldsymbol{x}_1, \boldsymbol{x}'_2) &= D'_{\mathcal{X}}(\boldsymbol{x}_1, \boldsymbol{x}_2) < D_{\mathcal{X}}(\boldsymbol{x}_1, \boldsymbol{x}_2) \\ &= D_{\mathcal{X}}(\boldsymbol{x}_1, \boldsymbol{x}'_2) = D_{\mathcal{Y}}(h(\boldsymbol{x}_1), h(\boldsymbol{x}'_2)). \quad \square \end{aligned}$$

5 Randomized Smoothing

For settings without a simple linear relation between the inputs and the outputs, neural networks often replace linear models. However, neural networks are often susceptible to adversarial examples [Szegedy *et al.*, 2014; Goodfellow *et al.*, 2015], which are inputs to the model that are created by applying a small perturbation to an original input with the goal of causing a very large change in the model's output. The frequent success of these attacks show that a small change in $\mathcal{X}$ can cause a large change in $\mathcal{Y}$, which is contrary to individual fairness.

Previously, Cohen *et al.* [2019] introduced randomized smoothing, a post-processing method that ensures that the post-processed model is robust against perturbations of size, measured with the standard L^2 norm, up to a threshold that depends on the input point. In this section, for any given metric, we apply a modified version of randomized smoothing and prove that the resulting model is individually fair under that metric. We note that this result does not immediately follow from prior results—individual fairness imposes the same constraint on every point in the input space, whereas the certified threshold of Cohen *et al.* is a function of the input point.

In the rest of this paper, we assume that $\mathcal{Y}$ is categorical, following the setting of Cohen *et al.* [2019]. In this section, we present and prove two methods for deriving an individually fair model from an arbitrary function $f : \mathbb{R}^d \to \mathcal{Y}$. Like the models considered by Dwork *et al.* [2012], our fair model $h_{f,g}$ maps $\mathbb{R}^d$ to $\Delta(\mathcal{Y})$, which is the set of probability distributions over $\mathcal{Y}$. It is important to note that $h_{f,g}$ is deterministic and that we treat its output simply as an array of probabilities. To avoid confusion with the randomness that we introduce in Section 6, we will write $h_{f,g}(\boldsymbol{x})[y]$ to denote the probability $\Pr[h_{f,g}(\boldsymbol{x}) = y]$.

Definition 5 (Randomized smoothing)**.** *Let $f : \mathbb{R}^d \to \mathcal{Y}$ be an arbitrary model, and let $g : \mathbb{R}^d \to \mathbb{R}$ be a probability distribution*[1]*. Then, the* smoothed model $h_{f,g} : \mathbb{R}^d \to \Delta(\mathcal{Y})$ *is defined by*

$$h_{f,g}(\boldsymbol{x})[y] = \int_{\mathbb{R}^d} \mathbb{1}[f(\boldsymbol{x} + \boldsymbol{t}) = y] \cdot g(\boldsymbol{t})\, d\boldsymbol{t} \tag{5}$$

for all $y \in \mathcal{Y}$, and g is called the smoothing distribution*.*

Intuitively, f is the original model, and the value of the smoothed model $h_{f,g}$ at $\boldsymbol{x}$ is found by querying f on points around $\boldsymbol{x}$. We choose the points around $\boldsymbol{x}$ according to the distribution g, and the output $h_{f,g}(\boldsymbol{x})$ of the smoothed model is a probability distribution of the values of f at the queried points. To reason about the individual fairness of $h_{f,g}$, we use the total variation distance (Eq. (6)) to define the distance $D_{\Delta(\mathcal{Y})}$ between probability distributions.

$$D_{\Delta(\mathcal{Y})}(Y_1, Y_2) = \tfrac{1}{2} \textstyle\sum_{y \in \mathcal{Y}} |Y_1[y] - Y_2[y]|. \tag{6}$$

5.1 Laplace Smoothing Distribution

One main difference between this setting and that in Section 4 is that we have a choice of the smoothing distribution g. Thus, instead of simply finding a metric under which the model is individually fair, we adapt the smoothing distribution to a given metric. Theorem 3 shows that, for any weighted L^p metric $D_{\mathcal{X}}$, there exists a smoothing distribution g that guarantees that $h_{f,g}$ is individually fair under $D_{\mathcal{X}}$ for all f.

Theorem 3 (Laplace smoothing)**.** *Let $\mathcal{M}_L$ be the set of all weighted L^p metrics. For any $D_{\mathcal{X}} \in \mathcal{M}_L$, let $g(\boldsymbol{t}) = \exp(-2D_{\mathcal{X}}(\boldsymbol{0}, \boldsymbol{t}))/Z$, where Z is the normalization factor $\int_{\mathbb{R}^d} \exp(-2D_{\mathcal{X}}(\boldsymbol{0}, \boldsymbol{t}))\, d\boldsymbol{t}$. Then, $h_{f,g}$ is individually fair under $D_{\mathcal{X}}$ for all f.*

[1] We abuse notation and use g to denote both the distribution and its probability density function.

Proof. We will show that $D_{\Delta(\mathcal{Y})}(h_{f,g}(\boldsymbol{x}), h_{f,g}(\boldsymbol{x}+\boldsymbol{\epsilon})) \leq D_{\mathcal{X}}(\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{\epsilon})$ for all $\boldsymbol{x}, \boldsymbol{\epsilon} \in \mathbb{R}^d$.

First, since $D_{\mathcal{X}}(\boldsymbol{t}, \boldsymbol{t}-\boldsymbol{\epsilon}) = D_{\mathcal{X}}(\boldsymbol{0}, \boldsymbol{\epsilon})$ for all weighted L^p metric $D_{\mathcal{X}}$, we have

$$D_{\mathcal{X}}(\boldsymbol{0}, \boldsymbol{t}) - D_{\mathcal{X}}(\boldsymbol{0}, \boldsymbol{\epsilon}) \leq D_{\mathcal{X}}(\boldsymbol{0}, \boldsymbol{t}-\boldsymbol{\epsilon}) \leq D_{\mathcal{X}}(\boldsymbol{0}, \boldsymbol{t}) + D_{\mathcal{X}}(\boldsymbol{0}, \boldsymbol{\epsilon}) \tag{7}$$

by the triangle inequality. We can apply the first inequality in Eq. (7) to bound the probability $g(\boldsymbol{t}-\boldsymbol{\epsilon})$ in terms of $g(\boldsymbol{t})$.

$$\begin{aligned} g(\boldsymbol{t}-\boldsymbol{\epsilon}) &= \exp(-2D_{\mathcal{X}}(\boldsymbol{0}, \boldsymbol{t}-\boldsymbol{\epsilon}))/Z \\ &\geq \exp(-2[D_{\mathcal{X}}(\boldsymbol{0}, \boldsymbol{t}) + D_{\mathcal{X}}(\boldsymbol{0}, \boldsymbol{\epsilon})])/Z \\ &= g(\boldsymbol{t}) / \exp(-2D_{\mathcal{X}}(\boldsymbol{0}, \boldsymbol{\epsilon})) \end{aligned}$$

Then, for all $y \in \mathcal{Y}$ we have

$$\begin{aligned} &h_{f,g}(\boldsymbol{x}+\boldsymbol{\epsilon})[y] \\ &= \textstyle\int_{\mathbb{R}^d} \mathbb{1}[f(\boldsymbol{x}+\boldsymbol{t}) = y] \cdot g(\boldsymbol{t}-\boldsymbol{\epsilon})\, d\boldsymbol{t} \\ &\geq \textstyle\int_{\mathbb{R}^d} \mathbb{1}[f(\boldsymbol{x}+\boldsymbol{t}) = y] \cdot g(\boldsymbol{t}) / \exp(-2D_{\mathcal{X}}(\boldsymbol{0}, \boldsymbol{\epsilon}))\, d\boldsymbol{t} \\ &= h_{f,g}(\boldsymbol{x})[y] / \exp(-2D_{\mathcal{X}}(\boldsymbol{0}, \boldsymbol{\epsilon})). \end{aligned} \tag{8}$$

Similarly, we can apply the second inequality in Eq. (7) to derive the upper bound

$$h_{f,g}(\boldsymbol{x}+\boldsymbol{\epsilon})[y] \leq h_{f,g}(\boldsymbol{x})[y] \cdot \exp(-2D_{\mathcal{X}}(\boldsymbol{0}, \boldsymbol{\epsilon})). \tag{9}$$

We can now apply a previous result by Kairouz *et al.* [2016, Theorem 6] to determine the maximum distance between $h_{f,g}(\boldsymbol{x})$ and $h_{f,g}(\boldsymbol{x}+\boldsymbol{\epsilon})$ that is attainable with the above constraints. For brevity, let c denote $\exp(-2D_{\mathcal{X}}(\boldsymbol{0}, \boldsymbol{\epsilon}))$. In the context of ε-local differential privacy, Kairouz *et al.* showed that the maximum possible total variation distance is $(e^{\varepsilon}-1)/(e^{\varepsilon}+1)$. Replacing e^{ε} with c, we see that the distance between $h_{f,g}(\boldsymbol{x})$ and $h_{f,g}(\boldsymbol{x}+\boldsymbol{\epsilon})$ is at most $(c-1)/(c+1)$.

Finally, it remains to be proven that this quantity is not more than $D_{\mathcal{X}}(\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{\epsilon})$, which is equivalent to $D_{\mathcal{X}}(\boldsymbol{0}, \boldsymbol{\epsilon})$ for weighted L^p metrics. Since this distance can be written as $(\ln c)/2$, it suffices to show that $(c-1)/(c+1) \leq (\ln c)/2$ for all $c \geq 1$. This inequality follows from the fact that equality holds at $c = 1$ and that the derivative of the right-hand side is never less than that of the left-hand side for $c \geq 1$. □

Although Theorem 3 identifies a smoothing distribution that ensures the individual *fairness* of the resulting model under $D_{\mathcal{X}}$, we also want the smoothed model to retain the *utility* of the original model f. In the extreme case where g is the uniform distribution over $\mathbb{R}^d$, the resulting model will be a constant function and therefore satisfy individual fairness under any metric, but it will not be very useful for classification tasks. More generally, smoothed models that are individually fair under smaller distance metrics tend to not preserve as much locally relevant information about f. Thus, we argue that a smoothing distribution does not unnecessarily lower the model's utility by showing that $D_{\mathcal{X}}$ is *minimal*. The definition of minimality that we use here differs from Definition 4 in that the smoothed model must be individually fair for all f.

Definition 6 (Minimal distance metric, smoothing)**.** *Let $\mathcal{M} \subseteq \mathcal{M}_L$ be a set of distance metrics in $\mathbb{R}^d$. A metric $D_{\mathcal{X}} \in \mathcal{M}$ is* minimal *in $\mathcal{M}$ with respect to a smoothing distribution g if* (1) *$h_{f,g}$ is individually fair under $D_{\mathcal{X}}$ for all f, and* (2) *there does not exist a different $D'_{\mathcal{X}} \in \mathcal{M}$ such that $h_{f,g}$ is individually fair under $D'_{\mathcal{X}}$ for all f and $D_{\mathcal{X}}(\boldsymbol{x}_1, \boldsymbol{x}_2) \geq D'_{\mathcal{X}}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ for all $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^d$.*

For general weighted L^p metrics, the inequalities in Eq. (7) are strict for most $\boldsymbol{t}$, so the bounds in Eqs. (8) and (9) are not tight, and $D_{\mathcal{X}}$ is not guaranteed to be minimal. On the other hand, if $D_{\mathcal{X}}$ is a weighted L^1 metric, Theorem 4 shows that it is minimal with respect to its Laplace smoothing distribution.

Theorem 4. *Let $\mathcal{M}_1$ be the set of all weighted L^1 metrics. For any $D_{\mathcal{X}} \in \mathcal{M}_1$, let $g(\boldsymbol{t}) = \exp(-2D_{\mathcal{X}}(\boldsymbol{0}, \boldsymbol{t}))/Z$, where Z is the normalization factor $\int_{\mathbb{R}^d} \exp(-2D_{\mathcal{X}}(\boldsymbol{0}, \boldsymbol{t}))\, d\boldsymbol{t}$. Then, $D_{\mathcal{X}}$ is uniquely minimal in $\mathcal{M}_1$ with respect to g.*

Proof. We have already proven in Theorem 3 that $h_{f,g}$ is individually fair under $D_{\mathcal{X}}$ for all f. It remains to show that there does not exist a different $D'_{\mathcal{X}} \in \mathcal{M}_1$ such that $h_{f,g}$ is individually fair under $D'_{\mathcal{X}}$ for all f and $D_{\mathcal{X}}(\boldsymbol{x}_1, \boldsymbol{x}_2) \geq D'_{\mathcal{X}}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ for all $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^d$.

Let $w_1, \ldots, w_d$ and $w'_1, \ldots, w'_d$ be the weights of $D_{\mathcal{X}}$ and $D'_{\mathcal{X}}$, respectively. If $D_{\mathcal{X}}(\boldsymbol{x}_1, \boldsymbol{x}_2) \geq D'_{\mathcal{X}}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ for all $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^d$, we must have $w_i \geq w'_i$ for all i. Moreover, since $D_{\mathcal{X}} \neq D'_{\mathcal{X}}$, there exists i such that $w_i > w'_i$. We now construct f such that $h_{f,g}$ is not individually fair under $D'_{\mathcal{X}}$.

Let $f : \mathbb{R}^d \to \{0, 1\}$ be a function such that $f(\boldsymbol{x}) = \mathbb{1}[x_i \geq 0]$. We will show that there exists $\epsilon > 0$ such that $D_{\Delta(\mathcal{Y})}(h_{f,g}(\boldsymbol{0}), h_{f,g}(\epsilon \boldsymbol{e}_i)) > D'_{\mathcal{X}}(\boldsymbol{0}, \epsilon \boldsymbol{e}_i)$, where $\boldsymbol{e}_i$ is the basis vector that is one in the i-th coordinate and zero in all others. Applying Eq. (5) and simplifying, we get $h_{f,g}(\boldsymbol{0})[0] = 1/2$ and $h_{f,g}(\epsilon \boldsymbol{e}_i)[0] = \exp(-2w_i\epsilon)/2$. Therefore, the distance $D_{\Delta(\mathcal{Y})}$ is $(1 - \exp(-2w_i\epsilon))/2$. Moreover, we have $D'_{\mathcal{X}} = w'_i\epsilon$. The ratio $D_{\Delta(\mathcal{Y})}/D'_{\mathcal{X}}$ approaches $w_i/w'_i > 1$ as $\epsilon \to 0$, so when ϵ is sufficiently small, we have $D_{\Delta(\mathcal{Y})} > D'_{\mathcal{X}}$.

Uniqueness follows from the argument given in the last paragraph of the proof of Theorem 1. □

5.2 Gaussian Smoothing Distribution

As we show in Section 7, in practice Laplace smoothing distributions do not preserve well the utility of f due to their relatively high densities at the tails. Thus, we present Gaussian smoothing as an alternative, which Theorem 5 shows is individually fair under any weighted L^2 metric. Since $D_2(\boldsymbol{x}_1, \boldsymbol{x}_2) \leq d^{\max(0, 1/2-1/p)} D_p(\boldsymbol{x}_1, \boldsymbol{x}_2)$ for any weighted L^2 and L^p metrics D_2 and D_p with the same weights, we can then scale the weights accordingly to make $h_{f,g}$ fair under any given L^p metric. For simplicity, we only consider the setting of binary classification, i.e., $\mathcal{Y} = \{0, 1\}$.

Theorem 5 (Gaussian smoothing)**.** *Let $D_{\mathcal{X}}$ be a weighted L^2 metric with weights $w_1, \ldots, w_d$, and let $\boldsymbol{\Sigma}$ be a diagonal matrix with $\Sigma_{ii} = (2\pi w_i)^{-1}$. If g is Gaussian with mean $\boldsymbol{0}$ and variance $\boldsymbol{\Sigma}$, $h_{f,g}$ is individually fair under $D_{\mathcal{X}}$ for all $f : \mathbb{R}^d \to \{0, 1\}$.*

To prove this theorem, we will apply the Neyman–Pearson lemma [Neyman and Pearson, 1933], as formulated by Cohen *et al.* [2019, Lemma 3].

Lemma 6 (Neyman–Pearson)**.** *Let X_1 and X_2 be random variables in $\mathbb{R}^d$ with densities μ_{X_1} and μ_{X_2}, and let*

$f, f^\star : \mathbb{R}^d \to \{0,1\}$ such that $f^\star(\boldsymbol{t}) = 1$ if and only if $\mu_{X_2}(\boldsymbol{t})/\mu_{X_1}(\boldsymbol{t}) \geq k$ for some threshold $k > 0$. Then,

$$\Pr[f(X_1) = 1] = \Pr[f^\star(X_1) = 1] \\ \textit{implies } \Pr[f(X_2) = 1] \leq \Pr[f^\star(X_2) = 1].$$

Proof of Theorem 5. We proceed by showing that $D_{\Delta(\mathcal{Y})}(h_{f,g}(\boldsymbol{x}), h_{f,g}(\boldsymbol{x}+\boldsymbol{\epsilon})) \leq D_{\mathcal{X}}(\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{\epsilon})$ for all $\boldsymbol{x}, \boldsymbol{\epsilon} \in \mathbb{R}^d$ and $f : \mathbb{R}^d \to \{0,1\}$. For any given f, we will first find $f^\star$ such that

$$D_{\Delta(\mathcal{Y})}(h_{f,g}(\boldsymbol{x}), h_{f,g}(\boldsymbol{x}+\boldsymbol{\epsilon})) \\ \leq D_{\Delta(\mathcal{Y})}(h_{f^\star,g}(\boldsymbol{x}), h_{f^\star,g}(\boldsymbol{x}+\boldsymbol{\epsilon})). \tag{10}$$

We will then show that $h_{f^\star,g}$ is individually fair under $D_{\mathcal{X}}$, which together with Eq. (10) implies that $h_{f,g}$ is also individually fair under $D_{\mathcal{X}}$.

Fix $\boldsymbol{x}, \boldsymbol{\epsilon} \in \mathbb{R}^d$, and assume without loss of generality that $h_{f,g}(\boldsymbol{x})[1] \leq h_{f,g}(\boldsymbol{x}+\boldsymbol{\epsilon})[1]$. Then, we have

$$D_{\Delta(\mathcal{Y})}(h_{f,g}(\boldsymbol{x}), h_{f,g}(\boldsymbol{x}+\boldsymbol{\epsilon})) = h_{f,g}(\boldsymbol{x}+\boldsymbol{\epsilon})[1] - h_{f,g}(\boldsymbol{x})[1]. \tag{11}$$

We apply Lemma 6 by choosing X_1 and X_2 such that $g(\boldsymbol{t}) = \mu_{X_1}(\boldsymbol{x}+\boldsymbol{t}) = \mu_{X_2}(\boldsymbol{x}+\boldsymbol{\epsilon}+\boldsymbol{t})$. By Eq. (5), we have $\Pr[f(X_1) = 1] = h_{f,g}(\boldsymbol{x})[1]$ and $\Pr[f(X_2) = 1] = h_{f,g}(\boldsymbol{x}+\boldsymbol{\epsilon})[1]$, and similar relations hold between $f^\star$ and $h_{f^\star,g}$. Therefore, if there exists $f^\star$ that satisfies the condition in Lemma 6 such that $h_{f,g}(\boldsymbol{x})[1] = h_{f^\star,g}(\boldsymbol{x})[1]$, then $h_{f,g}(\boldsymbol{x}+\boldsymbol{\epsilon})[1] \leq h_{f^\star,g}(\boldsymbol{x}+\boldsymbol{\epsilon})[1]$. Combining these two (in)equalities, we get

$$h_{f,g}(\boldsymbol{x}+\boldsymbol{\epsilon})[1] - h_{f,g}(\boldsymbol{x})[1] \leq h_{f^\star,g}(\boldsymbol{x}+\boldsymbol{\epsilon})[1] - h_{f^\star,g}(\boldsymbol{x})[1],$$

and Eq. (10) follows from Eq. (11) and its $h_{f^\star,g}$ counterpart.

We now show that it is possible to find $f^\star$ such that $h_{f^\star,g}(\boldsymbol{x})[1] = h_{f,g}(\boldsymbol{x})[1]$. By construction, we have that $f^\star(\boldsymbol{t}) = 1$ if and only if

$$\frac{\mu_{X_2}(\boldsymbol{t})}{\mu_{X_1}(\boldsymbol{t})} = \frac{g(\boldsymbol{t}-\boldsymbol{x}-\boldsymbol{\epsilon})}{g(\boldsymbol{t}-\boldsymbol{x})} \geq k$$

for some $k > 0$. Substituting in the Gaussian density function and solving for $\boldsymbol{t}$, we see that this inequality holds whenever $\boldsymbol{\epsilon}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{t} \geq \kappa$, where κ is a constant with respect to $\boldsymbol{t}$. When evaluating $h_{f^\star,g}(\boldsymbol{x})$ as per Eq. (5), $\boldsymbol{t}$ is distributed normally, and therefore $\boldsymbol{\epsilon}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{t}$ is also (univariate) Gaussian. Thus, with the appropriate value of κ we can obtain the desired $f^\star$.

Finally, it remains to show that $h_{f^\star,g}$ is individually fair under $D_{\mathcal{X}}$. Let $\tau = \boldsymbol{\epsilon}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{t}$, and let γ be the density function of τ. With some computation, we see that $f^\star(\boldsymbol{t})$ and $f^\star(\boldsymbol{t}+\boldsymbol{\epsilon})$ differ if and only if $\kappa \leq \tau < \kappa + \boldsymbol{\epsilon}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\epsilon}$. Moreover, since $\boldsymbol{t}$ has variance $\boldsymbol{\Sigma}$, the variance of $\tau = \boldsymbol{\epsilon}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{t}$ is $\boldsymbol{\epsilon}^T\boldsymbol{\Sigma}^{-1}\mathrm{Var}(\boldsymbol{t})(\boldsymbol{\epsilon}^T\boldsymbol{\Sigma}^{-1})^T = \boldsymbol{\epsilon}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\epsilon}$, and thus the maximum value of γ is $(2\pi\boldsymbol{\epsilon}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\epsilon})^{-1/2}$. We apply these two facts to arrive at the desired result:

$$\begin{aligned} &D_{\Delta(\mathcal{Y})}(h_{f^\star,g}(\boldsymbol{x}), h_{f^\star,g}(\boldsymbol{x}+\boldsymbol{\epsilon})) \\ &= h_{f^\star,g}(\boldsymbol{x}+\boldsymbol{\epsilon})[1] - h_{f^\star,g}(\boldsymbol{x})[1] \\ &= \textstyle\int_{\mathbb{R}^d} (f^\star(\boldsymbol{x}+\boldsymbol{\epsilon}+\boldsymbol{t}) - f^\star(\boldsymbol{x}+\boldsymbol{t})) \cdot g(\boldsymbol{t})\, d\boldsymbol{t} \\ &= \textstyle\int_{\mathbb{R}^d} (f^\star(\boldsymbol{t}+\boldsymbol{\epsilon}) - f^\star(\boldsymbol{t})) \cdot g(\boldsymbol{t}-\boldsymbol{x})\, d\boldsymbol{t} \\ &= \textstyle\int_{\kappa}^{\kappa+\boldsymbol{\epsilon}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\epsilon}} \gamma(\tau - \boldsymbol{\epsilon}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x})\, d\tau \\ &\leq \boldsymbol{\epsilon}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\epsilon} \cdot (2\pi\boldsymbol{\epsilon}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\epsilon})^{-1/2} \\ &= \sqrt{\textstyle\sum_{i=1}^d \epsilon_i^2/(2\pi\Sigma_{ii})} \\ &= \sqrt{\textstyle\sum_{i=1}^d \epsilon_i^2 \cdot w_i} = D_{\mathcal{X}}(\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{\epsilon}). \end{aligned}$$ □

We end this section with Theorem 7, which states that an L^2 metric $D_{\mathcal{X}}$ is minimal with respect to its Gaussian smoothing distribution. We omit the proof since it is very similar to that of Theorem 4.

Theorem 7. *Let $\mathcal{M}_2$ be the set of all weighted L^2 metrics. For any $D_{\mathcal{X}} \in \mathcal{M}_2$, let $w_1, \ldots, w_d$ be the weights, and let $\boldsymbol{\Sigma}$ be a diagonal matrix with $\sigma_{ii} = (2\pi w_i)^{-1}$. If g is a Gaussian with mean $\mathbf{0}$ and variance $\boldsymbol{\Sigma}$, $D_{\mathcal{X}}$ is uniquely minimal in $\mathcal{M}_2$ with respect to g.*

6 Practical Implementation

In practice, it is infeasible to compute $h_{f,g}(\boldsymbol{x})$ because of the integral in Eq. (5). Therefore, to apply randomized smoothing in practice, we approximate the integral by sampling n points independently from the smoothing distribution g, evaluating the model with this noise added to $\boldsymbol{x}$, and returning the observed probability of predicting each class on the sampled points. However, the resulting model may not be individually fair due to the finite sample size. Thus, we define and prove (ϵ, δ)-individual fairness, which requires that the model be close to individually fair with high probability.

Definition 7 ((ϵ, δ)-individual fairness)**.** *A randomized model $h : \mathcal{X} \to \Delta(\mathcal{Y})$ is* (ϵ, δ)-individually fair *under metric $D_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ if, for all $x_1, x_2 \in \mathcal{X}$,*

$$D_{\Delta(\mathcal{Y})}(h(x_1), h(x_2)) \leq D_{\mathcal{X}}(x_1, x_2) + \epsilon \tag{12}$$

with probability at least $1-\delta$. The probability is taken over the randomness of h.

Theorem 8. *Let $h^n_{f,g}$ be a model that approximates $h_{f,g}$ with n samples. If $|\mathcal{Y}| = m$ and $h_{f,g}$ is fair under $D_{\mathcal{X}}$, then $h^n_{f,g}$ is $(\epsilon, 2me^{-4n\epsilon^2/m^2})$-individually fair under $D_{\mathcal{X}}$.*

Proof. Consider any two points $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^d$. Since $h_{f,g}$ is individually fair under $D_{\mathcal{X}}$, we have

$$\begin{aligned} &\tfrac{1}{2}\textstyle\sum_{y\in\mathcal{Y}} |h_{f,g}(\boldsymbol{x}_1)[y] - h_{f,g}(\boldsymbol{x}_2)[y]| \\ &\qquad = D_{\Delta(\mathcal{Y})}(h_{f,g}(\boldsymbol{x}_1), h_{f,g}(\boldsymbol{x}_2)) \leq D_{\mathcal{X}}(\boldsymbol{x}_1, \boldsymbol{x}_2). \end{aligned}$$

We will show that $|\mathrm{diff}[y]| > 2\epsilon/m$ with probability less than $2e^{-4n\epsilon^2/m^2}$, where $\mathrm{diff}[y] = (h^n_{f,g}(\boldsymbol{x}_1)[y] - h^n_{f,g}(\boldsymbol{x}_2)[y]) - (h_{f,g}(\boldsymbol{x}_1)[y] - h_{f,g}(\boldsymbol{x}_2)[y])$. Then, by union bound, with probability at least $1 - 2me^{-4n\epsilon^2/m^2}$ we will have $\mathrm{diff}[y] > 2\epsilon/m$ for all $y \in \mathcal{Y}$, which leads to our desired result.

$$\begin{aligned} &D_{\Delta(\mathcal{Y})}(h^n_{f,g}(\boldsymbol{x}_1), h^n_{f,g}(\boldsymbol{x}_2)) \\ &\leq \tfrac{1}{2}\textstyle\sum_{y\in\mathcal{Y}} |h_{f,g}(\boldsymbol{x}_1)[y] - h_{f,g}(\boldsymbol{x}_2)[y]| + \tfrac{1}{2}\textstyle\sum_{y\in\mathcal{Y}} |\mathrm{diff}[y]| \\ &\leq D_{\mathcal{X}}(x_1, x_2) + \epsilon \end{aligned}$$

Fix $y \in \mathcal{Y}$, and let $X_{ij} = \mathbb{1}[f(\boldsymbol{x}_i + \boldsymbol{t}_{ij}) = y]$, where $\boldsymbol{t}_{ij}$ is the j-th sample drawn from the smoothing distribution g

while evaluating $h^n_{f,g}(\boldsymbol{x}_i)$. Then, $h^n_{f,g}(\boldsymbol{x}_i)[y] = \frac{1}{n}\sum_{j=1}^n X_{ij}$ and $h_{f,g}(\boldsymbol{x}_i)[y] = \frac{1}{n}\mathbb{E}[\sum_{j=1}^n X_{ij}]$, so $\text{diff} = \frac{1}{n}\sum_{j=1}^n (X_{1j} - X_{2j} - \mathbb{E}[X_{1j} - X_{2j}])$. The theorem follows from Hoeffding's inequality.

$$\begin{aligned}&\Pr[|\text{diff}[y]| > 2\epsilon/m] \\ &= \Pr[|\tfrac{1}{2n}\textstyle\sum_{j=1}^n (X_{1j} - X_{2j} - \mathbb{E}[X_{1j} - X_{2j}])| > \epsilon/m] \\ &< 2e^{-4n\epsilon^2/m^2}\end{aligned}$$ □

In the extended version of this paper [Yeom and Fredrikson, 2020], we provide pseudocode of an implementation of randomized smoothing.

7 Experiments

Theorems 3, 5, and 8 show that smoothed models created using randomized smoothing are individually fair, but we have no similar results about their utility except heuristic arguments from minimality. In this section we measure the utility of smoothed models $h_{f,g}$ on four real-world datasets (detailed below), using the smoothing distributions described in Theorems 3 and 5.

The weights of $D_{\mathcal{X}}$ were chosen to be proportional to those of a logistic regression model trained on the same dataset, so the features that receive little weight in the linear model, which are thus less likely to be predictive, have little effect on the output of the smoothed model. The linear model weights were multiplied by a constant between 0.5 and 5 (depending on the dataset) to make the mean weight of $D_{\mathcal{X}}$ equal 1. For each dataset, we trained a neural network f with two dense hidden layers of 128 ReLU neurons each, as well as a logistic regression model to use for deriving the targeted metric. When training neural networks, we augmented training data with noise drawn from the smoothing distribution, as prior work [Cohen *et al.*, 2019] shows that this improves the utility of the smoothed model. For smoothing, we sampled $n = 10^5$ points, which by Theorem 8 corresponds to a guarantee of $\delta = 1.8 \times 10^{-4}$ at $\epsilon = 10^{-2}$.

Adult. Our model uses the five numerical features from the UCI Adult dataset [Dua and Karra Taniskidou, 2017] to predict whether a person earns more than $50,000 per year.

COMPAS. We use the dataset compiled by ProPublica [Angwin *et al.*, 2016] to analyze the COMPAS recidivism prediction model [Equivant, 2019]. Our model uses eight features (15 when one-hot encoded) to predict whether a person will recidivate within the next two years.

SSL. The Strategic Subject List dataset [City of Chicago, 2017] contains scores given by Chicago Police Department's model to rate a person's risk of being involved in a shooting incident, either as a perpetrator or a victim. Our model uses the same eight numerical features used by Chicago's model to predict a person's SSL risk score.

Seizure. In the UCI Epileptic Seizure dataset [Dua and Karra Taniskidou, 2017; Andrzejak *et al.*, 2001], every row consists of 178 readings from an EEG taken over a second. Our model predicts whether a person is experiencing a seizure during that second.

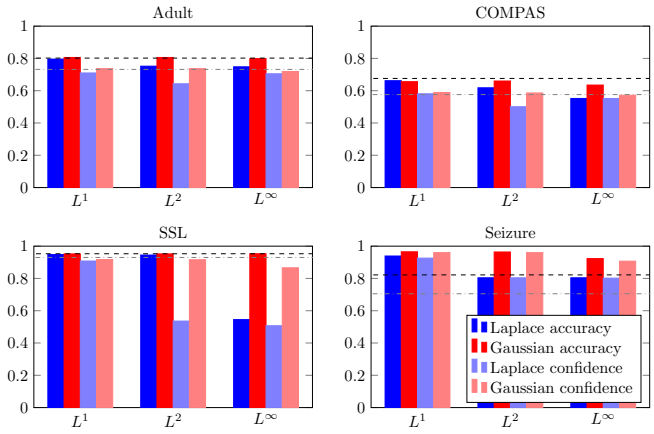

Figure 1: Utility of smoothed models derived from the four datasets described in Section 7. The black dashed line indicates the accuracy of the logistic regression model, and the dash-dotted line its average probit confidence. Because smoothed models output probabilities, accuracy and mean confidence are both reasonable measures of their utility, but only mean confidence preserves the individual fairness of the smoothed model.

7.1 Results

We applied both Laplace and Gaussian smoothing to create models that are individually fair under the weighted L^1, L^2, or L^∞ metrics, with weights derived from those of the corresponding logistic regression model as previously described. Because the outputs of the smoothed models are probabilities, we measured their utilities both in terms of standard accuracy and the mean probit confidence assigned to the correct class, $\mathbb{E}_{(\boldsymbol{x},y)}[h_{f,g}(\boldsymbol{x})[y]]$. Although accuracy is a more common measure of utility, its use of the thresholding operator argmax is incompatible with the individual fairness of $h_{f,g}$.

The results in Fig. 1 show that Gaussian-smoothed models *approximately match or exceed the performance of the logistic model while achieving similar individual fairness guarantees.* We note that for all datasets except for Seizure, the accuracy of the unsmoothed neural network was within 1.5% of the logistic regression model; on Seizure, the neural network achieved 96.6% whereas the logistic model gave 82.2% accuracy. The Gaussian-smoothed Seizure model came very close ($\leq 0.2\%$) to the unsmoothed neural network for L^1 and L^2 metrics, far exceeding the performance of the linear model.

We conclude by noting that, although L^1 metrics are minimal with respect to Laplace smoothing, Gaussian smoooothing outperforms on these metrics in practice. Laplace distributions have higher densities at the tails, resulting in more queries that are very dissimilar to the input point $\boldsymbol{x}$. Thus, in practice it can be preferable to use Gaussian smoothing for every L^p metric, adjusting the weights as shown in Section 5.2 to account for the value of p.

Acknowledgments

This material is based upon work supported by Bosch Corporation, an NVIDIA GPU grant, and the National Science Foundation under Grant No. CNS-1704845. The authors would like to thank Shayak Sen for his helpful feedback.

References

[Andrzejak *et al.*, 2001] Ralph G Andrzejak, Klaus Lehnertz, Florian Mormann, Christoph Rieke, Peter David, and Christian E Elger. Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. *Physical Review E*, 64(6):061907, 2001.

[Angwin *et al.*, 2016] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*, 2016.

[Calmon *et al.*, 2017] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pages 3992–4001, 2017.

[Canetti *et al.*, 2019] Ran Canetti, Aloni Cohen, Nishanth Dikkala, Govind Ramnarayan, Sarah Scheffler, and Adam Smith. From soft classifiers to hard decisions: How fair can we be? In *ACM Conference on Fairness, Accountability, and Transparency*, pages 309–318, 2019.

[Chouldechova and Roth, 2018] Alexandra Chouldechova and Aaron Roth. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*, 2018.

[City of Chicago, 2017] City of Chicago. Strategic Subject List. https://data.cityofchicago.org/Public-Safety/Strategic-Subject-List/4aki-r3np, 2017.

[Cohen *et al.*, 2019] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pages 1310–1320, 2019.

[Dua and Karra Taniskidou, 2017] Dheeru Dua and Efi Karra Taniskidou. UCI machine learning repository. https://archive.ics.uci.edu/ml, 2017.

[Dwork *et al.*, 2012] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science*, pages 214–226, 2012.

[Equivant, 2019] Equivant. Practitioner's guide to COMPAS core. http://www.equivant.com/wp-content/uploads/Practitioners-Guide-to-COMPAS-Core-040419.pdf, 2019.

[Gillen *et al.*, 2018] Stephen Gillen, Christopher Jung, Michael Kearns, and Aaron Roth. Online learning with an unknown fairness metric. In *Advances in Neural Information Processing Systems*, pages 2600–2609, 2018.

[Goodfellow *et al.*, 2015] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[Hardt *et al.*, 2016] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016.

[Ilvento, 2019] Christina Ilvento. Metric learning for individual fairness. *arXiv preprint arXiv:1906.00250*, 2019.

[Jung *et al.*, 2019] Christopher Jung, Michael Kearns, Seth Neel, Aaron Roth, Logan Stapleton, and Zhiwei Steven Wu. Eliciting and enforcing subjective individual fairness. *arXiv preprint arXiv:1905.10660*, 2019.

[Kairouz *et al.*, 2016] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. Extremal mechanisms for local differential privacy. *Journal of Machine Learning Research*, 17(1):492–542, 2016.

[Lohia *et al.*, 2019] Pranay K Lohia, Karthikeyan Natesan Ramamurthy, Manish Bhide, Diptikalyan Saha, Kush R Varshney, and Ruchir Puri. Bias mitigation post-processing for individual and group fairness. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2847–2851, 2019.

[Madras *et al.*, 2018] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3381–3390, 2018.

[Neyman and Pearson, 1933] Jerzy Neyman and Egon Sharpe Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London*, 231(694–706):289–337, 1933.

[Szegedy *et al.*, 2014] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

[Tan *et al.*, 2019] Zilong Tan, Samuel Yeom, Matt Fredrikson, and Ameet Talwalkar. Learning fair representations for kernel models. *arXiv preprint arXiv:1906.11813*, 2019.

[Yeom and Fredrikson, 2020] Samuel Yeom and Matt Fredrikson. Individual fairness revisited: Transferring techniques from adversarial robustness. *arXiv preprint arXiv:2002.07738*, 2020.

[Zafar *et al.*, 2017] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *International Conference on World Wide Web*, pages 1171–1180, 2017.

[Zemel *et al.*, 2013] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013.