Understanding Fairness Surrogate Functions in Algorithmic Fairness

Wei Yao ⋆ *wei.yao@ruc.edu.cn*
Gaoling School of Artificial Intelligence
Renmin University of China, Beijing
Beijing Key Laboratory of Big Data Management and Analysis Methods, Beijing

Zhanke Zhou ⋆ *cszkzhou@comp.hkbu.edu.hk*
TMLR Group, Department of Computer Science
Hong Kong Baptist University

Zhicong Li *zhicongli@ruc.edu.cn*
Gaoling School of Artificial Intelligence
Renmin University of China, Beijing
Beijing Key Laboratory of Big Data Management and Analysis Methods, Beijing

Bo Han *bhanml@comp.hkbu.edu.hk*
TMLR Group, Department of Computer Science
Hong Kong Baptist University

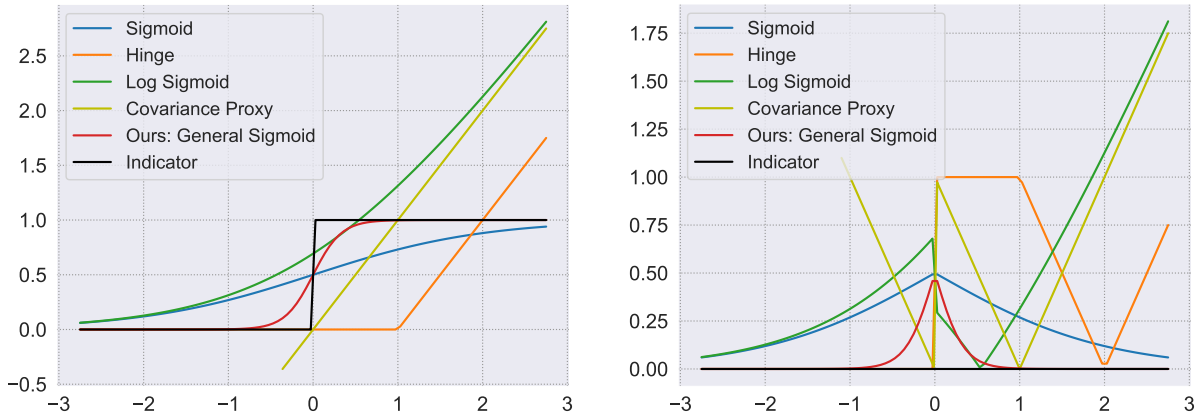
Yong Liu † *liuyonggsai@ruc.edu.cn*
Gaoling School of Artificial Intelligence
Renmin University of China, Beijing
Beijing Key Laboratory of Big Data Management and Analysis Methods, Beijing

Reviewed on OpenReview: *https://openreview.net/forum?id=iBgmoMTlaz*

Abstract

It has been observed that machine learning algorithms exhibit biased predictions against certain population groups. To mitigate such bias while achieving comparable accuracy, a promising approach is to introduce surrogate functions of the concerned fairness definition and solve a constrained optimization problem. However, it is intriguing in previous work that such fairness surrogate functions may yield unfair results and high instability. In this work, in order to deeply understand them, taking a widely used fairness definition—demographic parity as an example, we show that there is a *surrogate-fairness gap* between the fairness definition and the fairness surrogate function. Also, the theoretical analysis and experimental results about the "gap" motivate us that the fairness and stability will be affected by the points far from the decision boundary, which is the *large margin points issue* investigated in this paper. To address it, we propose the general sigmoid surrogate to simultaneously reduce both the surrogate-fairness gap and the variance, and offer a rigorous fairness and stability upper bound. Interestingly, the theory also provides insights into two important issues that deal with the *large margin points* as well as obtaining a more *balanced dataset* are beneficial to fairness and stability. Furthermore, we elaborate a novel and general algorithm called Balanced Surrogate, which iteratively reduces the "gap" to mitigate unfairness. Finally, we provide empirical evidence showing that our methods consistently improve fairness and stability while maintaining accuracy comparable to the baselines in three real-world datasets.

⋆ indicates equal contribution.
† indicates the corresponding author.



(a) Surrogate Functions

(b) Surrogate-Fairness Gap

Figure 1: (a) Some examples of fairness surrogate functions. The closer the surrogate functions are to the indicator function, the better they represent DP. More details are introduced in Section 3.2. (b) The surrogate-fairness gap of different surrogate functions. It measures the difference between surrogate functions and the indicator function. There is a much smaller gap for our general sigmoid surrogate.

1 Introduction

Recently, increasing attention has been paid to the fairness issue in supervised machine learning. That is, although the classifiers seek a higher accuracy, some groups with certain sensitive features (e.g., sex, race, age) may be unfairly treated, which raises ethical problems (Julia Angwin & Kirchner, 2016; Mehrabi et al., 2021; Caton & Haas, 2020). One can be litigated for committing adverse impacts if his/her decision-making process disproportionately treats groups with sensitive attributes (Barocas & Selbst, 2016).

To quantitatively measure the extent of fairness violation, a usual way is adopting the fairness definition, *demographic parity* (DP), which requires the decision makers to accept a roughly equal proportion of each group (Barocas et al., 2019). Existing methods follow the fairness-aware manner of solving a constrained optimization problem, where the learning objective is integrated with the standard loss and a fairness constraint. To incorporate DP into the constraint, fairness surrogate functions are used to replace the indicator function, which is intractable for gradient-based algorithms (Lohaus et al., 2020; Bendekgey & Sudderth, 2021) (refer to Figure 1(a) for some examples). To date, various surrogate functions have been proposed to incorporate fairness definitions into constraints (Wu et al., 2019; Goh et al., 2016; Padh et al., 2021; Zafar et al., 2017a;b;c; Bendekgey & Sudderth, 2021). They are widely applied in various machine learning domains, such as differential privacy (Ding et al., 2020), meta-learning (Zhao et al., 2020), and semi-supervised learning (Zhang et al., 2020). Unfortunately, these surrogate functions encounter two risks. One risk is that if these fairness constraints are used, even when the constraints are perfectly satisfied, there is *no guarantee* whether DP is satisfied (Lohaus et al., 2020). And the fairness surrogate functions may lead to even unfair solutions (Radovanović et al., 2022). Moreover, another risk arises from the high variance issue observed in existing fairness-aware algorithms with surrogate functions (Friedler et al., 2019), rendering them unstable for deployment under fairness requirements (Ganesh et al., 2023).

In this paper, we evaluate these multifarious surrogate functions in algorithmic fairness with both rigorous theorems and extensive experiments. Firstly, we stress the importance of the "**surrogate-fairness gap**", which is *the disparity between the fairness surrogate function and the fairness definition*. It is the decisive factor of whether the fairness surrogate function can lead to fair outcomes and should be minimized. Additionally, we delve into the **variance** of the substitute for DP, highlighting the adverse impact of unbounded surrogates on stability. Drawing upon the inherent property of the surrogate-fairness gap and instability, we conduct an in-depth examination of the **large margin points issue** within the context of unbounded surrogate functions. To reduce the "gap", we propose two solutions to improve the existing surrogate func-

tions: a theoretically motivated fairness surrogate function named *general sigmoid* with upper bounds of the violation of DP, and a novel algorithm called the *balanced surrogate* to iteratively reduce the gap during training.

Our analysis is general for fairness surrogate functions in the case of a common fairness definition, DP. Additionally, we employ a widely recognized covariance proxy (Zafar et al., 2017c) as an illustrative instance of fairness surrogate function. In particular, we first derive the violation of DP for the fairness surrogate functions in Section 4. The violation of DP depends on two factors: the surrogate function itself and the surrogate-fairness gap. The "gap" (shown in Figure 1(b)) directly determines whether a surrogate function is an appropriate substitute for DP. Secondly, we explore the variance of the surrogate function in Section 4.1, emphasizing the detrimental impact of unbounded surrogates on stability. Furthermore, driven by the "gap" and variance, we recognize that large margin points—those data points lying significantly distant from the decision boundary—pose challenges in constraining the fairness and stability for unbounded surrogate functions. This observation is validated through a case study on three real-world datasets in Section 4.2. With theoretical motivations, we introduce the general sigmoid surrogate in Section 5.1 to address large margin points and simultaneously bound the "gap" and variance. We theoretically demonstrate that there is a reliable fairness and stability guarantee for it. Interestingly, the theorems also shed light on the importance of a balanced dataset for both fairness and stability. Furthermore, in Section 5.2, we propose balanced surrogates, a novel and general algorithm that iteratively reduces the "gap" to improve fairness. It is a plug-and-play learning paradigm for the naive fairness-aware training framework using fairness surrogate functions. In the experiments in Section 6 using three real-world datasets, our methods generally enhance fair predictions and stability, while maintaining accuracy comparable to the baselines. Overall, our main contributions are three-fold:

- We demonstrate the importance of *Surrogate-fairness Gap* for fairness surrogate functions and provide an analysis of the *variance*. We emphasize the importance for researchers to consider the impact of the *large margin points issue* on the fairness and stability of unbounded surrogate functions.
- We propose *General Sigmoid Surrogate* and demonstrate that it achieves fairness and stability guarantees. The theoretical results further provide insights to the community that *large margin points issue* needs to be solved and a *balanced dataset* is beneficial to obtain a fairer and more stable classifier.
- We present *Balanced Surrogate*, a novel and general method that iteratively reduces the "gap" to improve the fairness of any fairness surrogate functions.

2 Related Work

Fairness-aware Algorithms. To mitigate bias, there are various kinds of classical fair algorithms, most of which fall into three categories: pre-processing, in-processing, and post-processing. The pre-processing method is to learn a fair representation that tries to remove information correlated to the sensitive feature while preserving other information for training, e.g., (Calders et al., 2009; Kamiran & Calders, 2011; Zemel et al., 2013; Feldman et al., 2015; Calmon et al., 2017). The downstream tasks then use the fair representation instead of the original biased dataset. The post-processing method is to modify the prediction results to satisfy the fairness definition, e.g., (Kamiran et al., 2012; Fish et al., 2016; Hardt et al., 2016). The in-processing method is to remove unfairness during training. Some intuitive and easy-to-use ideas involve applying fairness constraints (Goh et al., 2016; Zafar et al., 2017a;b;c; Bechavod & Ligett, 2017; Wu et al., 2019; Bendekgey & Sudderth, 2021; Padh et al., 2021) and adding a regularization term to penalize unfairness (Kamishima et al., 2012; Berk et al., 2017; Agarwal et al., 2018; Lohaus et al., 2020; Shui et al., 2022b). Refer to Appendix C.1 for other fairness-aware in-processing approaches. Our paper focuses on in-processing methods, with a particular emphasis on fairness surrogate functions, which are widely used in fairness constraints and fairness regularization methods mentioned above.

Fairness Surrogate Functions. Although many existing popular surrogates work well in practice, for example, linear (Donini et al., 2018; Agarwal et al., 2018; Bechavod & Ligett, 2017), ramp (Goh et al.,

2016; Zafar et al., 2017b), convex-concave (Zafar et al., 2017a), hinge (Wu et al., 2019), sigmoid and log-sigmoid (Bendekgey & Sudderth, 2021). They suffer from the same issue: there is not a fairness guarantee for them (Lohaus et al., 2020). And using fairness constraints or regularization can unexpectedly yield unfair solutions (Radovanović et al., 2022). Refer to Appendix C.2 for a meticulous overview of existing works, most of which present counterexamples for analysis. The high variance issue has been observed in existing fairness-aware algorithms (Ganesh et al., 2023), including the instability of the covariance proxy (Friedler et al., 2019). In this paper, in addition to empirically showing counter-examples, we both theoretically and empirically underscore the significance of the surrogate-fairness gap and variance, which are fundamental factors contributing to the two aforementioned problems, respectively. Our general sigmoid surrogate is shown to deal with the large margin points to simultaneously reduce both the surrogate-fairness gap and the variance. There is also fairness and stability upper bound for it, which is crucial in this field (Gallegos et al., 2023; Mehrabi et al., 2021; Caton & Haas, 2020). Additionally, in order to reduce the gap, we also devise a balanced surrogate approach to further improve the fairness and stability of surrogate functions, which may deserve a deeper exploration in this field for future work.

3 Preliminaries

3.1 Fairness-aware Classification

Note the general purpose of fairness-aware classification is to find a classifier with minimal accuracy loss while satisfying certain fairness constraints. For simplicity, we set up the problem as the binary classification task with only a binary-sensitive feature: with the training set $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ consisting of feature vectors $\mathbf{x}_i \in R^d$ and the corresponding class labels $y_i \in \{0, 1\}$, one needs to predict the labels of a test set. Let $d_\theta(\mathbf{x})$ denotes the signed distance between the feature vector $\mathbf{x}$ and the decision boundary parameterized by θ. Given a point $\mathbf{x}_i$ in the test set, a classifier will predict it as positive if $d_\theta(\mathbf{x}_i) > 0$ and zero if $d_\theta(\mathbf{x}_i) \leq 0$. Among the features of $\mathbf{x}$, there is one binary sensitive attribute $z \in \{-1, +1\}$ (e.g., sex, race, age).

As introduced in Section 1, a widely used fairness definition is called the *demographic parity* (DP) (Mehrabi et al., 2021; Caton & Haas, 2020). It states that each protected class should receive the positive outcome at equal rates, i.e.,

$$P(d_\theta(\mathbf{x}) > 0 | z = +1) = P(d_\theta(\mathbf{x}) > 0 | z = -1).$$

And further, the *difference of demographic parity* (DDP) metric (Lohaus et al., 2020) can be used to measure the degree to which demographic parity is violated:

$$DDP = P(d_\theta(\mathbf{x}) > 0 | z = +1) - P(d_\theta(\mathbf{x}) > 0 | z = -1).$$

Then, with this metric, whether a classifier satisfies demographic parity can be determined by the condition $|DDP| \leq \epsilon$, where $\epsilon \geq 0$ is a given threshold.

3.2 Surrogate Functions

We divide the training set into four classes according to the predicted labels and sensitive features:

$$\mathcal{N}_{1a} = \{(\mathbf{x}_i, y_i) \in \mathcal{S} \mid d_\theta(\mathbf{x}_i) > 0, z_i = +1\}, \qquad \mathcal{N}_{1b} = \{(\mathbf{x}_i, y_i) \in \mathcal{S} \mid d_\theta(\mathbf{x}_i) > 0, z_i = -1\},$$
$$\mathcal{N}_{0a} = \{(\mathbf{x}_i, y_i) \in \mathcal{S} \mid d_\theta(\mathbf{x}_i) \leq 0, z_i = +1\}, \qquad \mathcal{N}_{0b} = \{(\mathbf{x}_i, y_i) \in \mathcal{S} \mid d_\theta(\mathbf{x}_i) \leq 0, z_i = -1\},$$

where $N_{1a}, N_{1b}, N_{0a}, N_{0b}$ are sizes of $\mathcal{N}_{1a}, \mathcal{N}_{1b}, \mathcal{N}_{0a}, \mathcal{N}_{0b}$, respectively. To consider DDP as fairness constraints for optimization, the probability in it cannot be computed directly, so frequency is used to estimate them:

$$\begin{aligned} \widehat{DDP}_{\mathcal{S}} &= \frac{N_{1a}}{N_{1a} + N_{0a}} - \frac{N_{1b}}{N_{1b} + N_{0b}} \\ &= \frac{\sum_{(\mathbf{x},y) \in \mathcal{N}_{1a} \cup \mathcal{N}_{0a}} \mathbb{1}_{d_\theta(\mathbf{x}) > 0}}{N_{1a} + N_{0a}} - \frac{\sum_{(\mathbf{x},y) \in \mathcal{N}_{1b} \cup \mathcal{N}_{0b}} \mathbb{1}_{d_\theta(\mathbf{x}) > 0}}{N_{1b} + N_{0b}}, \end{aligned} \tag{1}$$

where $\mathbb{1}_{[\cdot]} : \mathbb{R} \to \{0, 1\}$ is the indicator function that returns 1 if the condition is true and 0 otherwise.

In application, $\widehat{DDP}_{\mathcal{S}}$ usually serves as a substitute for DDP to judge the fairness of a classifier. However, due to $\mathbb{1}_{d_\theta(\mathbf{x})>0}$, it is intractable to directly incorporate $\widehat{DDP}_{\mathcal{S}}$ into constraints for gradient-based algorithms. So smooth **surrogate function** $\phi : \mathbb{R} \to \mathbb{R}$ is used to replace $\mathbb{1}_{d_\theta(\mathbf{x})>0}$ with $\phi(d_\theta(\mathbf{x}))$:

$$\widetilde{DDP}_{\mathcal{S}}(\phi) = \frac{\sum_{(\mathbf{x},y)\in\mathcal{N}_{1a}\cup\mathcal{N}_{0a}} \phi(d_\theta(\mathbf{x}))}{N_{1a}+N_{0a}} - \frac{\sum_{(\mathbf{x},y)\in\mathcal{N}_{1b}\cup\mathcal{N}_{0b}} \phi(d_\theta(\mathbf{x}))}{N_{1b}+N_{0b}}. \tag{2}$$

Then $\widetilde{DDP}_{\mathcal{S}}(\phi)$ can be incorporated into constraints as $\left|\widetilde{DDP}_{\mathcal{S}}(\phi)\right| \leq \epsilon$, where ϵ is the threshold. In this way, the fairness-aware classification problem becomes a feasible constrained optimization problem.

One popular fairness surrogate function is *covariance proxy* (CP), which is introduced by (Zafar et al., 2017c). Empirical study shows that CP can reflect the difference of demographic parity and can be incorporated as constraints for fairness-aware classification problem (Ding et al., 2020; Zhao et al., 2020; Zhang et al., 2020). We defined a general version of it as

$$\widehat{Cov}_{\mathcal{S}}(\phi) = \frac{1}{N}\sum_{i=1}^{N}(z_i - \overline{z})\phi(d_\theta(\mathbf{x}_i)), \tag{3}$$

where $\overline{z}$ is the mean of z over the training set. If $\phi(x) = x$, the equation (3) recovers the original definition of CP in (Zafar et al., 2017c). Also, $\widetilde{DDP}_{\mathcal{S}}(\phi) \propto \widehat{Cov}_{\mathcal{S}}(\phi)$ and the proof can be found in Appendix A.1. It means that *the original CP is equivalent to the linear surrogate function* $\phi(x) = x$, which is also explained in previous work (Lohaus et al., 2020; Bendekgey & Sudderth, 2021). We provide theoretical results for $\widetilde{DDP}_{\mathcal{S}}(\phi)$ in the main paper, and extend them to $\widehat{Cov}_{\mathcal{S}}(\phi)$ in Appendix A.2 with the same conclusions.

4 The Surrogate-fairness Gap

We first emphasize the importance of surrogate-fairness gap, and then point out two issues: instability (Section 4.1) and large margin points (Section 4.2), which may influence the surrogate-fairness gap.

Firstly, we connect $\widehat{DDP}_{\mathcal{S}}$ and $\widetilde{DDP}_{\mathcal{S}}(\phi)$ together, and build the surrogate-fairness gap between them.

Proposition 1. *Define the magnitude of the signed distance by $D_\theta(\mathbf{x})$, i.e., $D_\theta(\mathbf{x}) = |d_\theta(\mathbf{x})|$. It satisfies:*

$$\underbrace{\widehat{DDP}_{\mathcal{S}} - \widetilde{DDP}_{\mathcal{S}}(\phi)}_{\textit{surrogate-fairness gap}} = \frac{\sum_{(\mathbf{x},y)\in\mathcal{N}_{1a}\cup\mathcal{N}_{0a}} \left[\mathbb{1}_{d_\theta(\mathbf{x})>0} - \phi(d_\theta(\mathbf{x}))\right]}{N_{1a}+N_{0a}} - \frac{\sum_{(\mathbf{x},y)\in\mathcal{N}_{1b}\cup\mathcal{N}_{0b}} \left[\mathbb{1}_{d_\theta(\mathbf{x})>0} - \phi(d_\theta(\mathbf{x}))\right]}{N_{1b}+N_{0b}}. \tag{4}$$

There is a surrogate-fairness gap between $\widehat{DDP}_{\mathcal{S}}$ and $\widetilde{DDP}_{\mathcal{S}}(\phi)$. For $\widetilde{DDP}_{\mathcal{S}}(\phi)$, it can serve as a fairness constraint or regularization in the algorithm. The algorithm will automatically find a solution that penalizes large $\left|\widetilde{DDP}_{\mathcal{S}}(\phi)\right|$. Unfortunately, it is different for the "gap", which comes from the inherent difference between the indicator function and the fairness surrogate function. For the ideal case $\phi(x) = \mathbb{1}_{x>0}$, which means that $\phi(d_\theta(\mathbf{x})) = \mathbb{1}_{d_\theta(\mathbf{x})>0}$, the gap is zero and reducing the constraint or regularization term $\left|\widetilde{DDP}_{\mathcal{S}}(\phi)\right|$ is equivalent to reducing $\left|\widehat{DDP}_{\mathcal{S}}\right|$. However, for any surrogate function ϕ, the gap will be inevitably introduced unless $\phi(x) = \mathbb{1}_{x>0}$. When the surrogate-fairness gap is small enough, there is fairness guarantee for the classifier. In practice, Figure 1(b) shows the surrogate-fairness gap of different fairness surrogate functions, including CP (which is equivalent to linear surrogate function) (Zafar et al., 2017c), hinge (Wu et al., 2019), log-sigmoid as well as sigmoid (Bendekgey & Sudderth, 2021), and our general sigmoid surrogate. It suggests that unbounded surrogate functions tend to exhibit a larger surrogate-fairness gap. The bounded surrogate functions, such as sigmoid and general sigmoid, both exhibit a bounded "gap".

4.1 Instability

The variance of $\widetilde{DDP}_{\mathcal{S}}(\phi)$ is $Var\left(\widetilde{DDP}_{\mathcal{S}}(\phi)\right) = \mathbb{E}\left(\widetilde{DDP}_{\mathcal{S}}(\phi)\right)^2 - \left[\mathbb{E}\left(\widetilde{DDP}_{\mathcal{S}}(\phi)\right)\right]^2$. If we choose bounded surrogate $\phi(x) \in [0,1]$, then $\widetilde{DDP}_{\mathcal{S}}(\phi) \in [-1,1]$, which means that $Var\left(\widetilde{DDP}_{\mathcal{S}}(\phi)\right) \in [0,1]$. Therefore, there is an stability guarantee for $\widetilde{DDP}_{\mathcal{S}}(\phi)$ if $\phi(x) \in [0,1]$. However, if we choose unbounded surrogate function (such as $\phi(x) = x \in [-\infty, +\infty]$ for the original CP), the resulting values of $\phi(x)$ are not constrained within the range $[0,1]$. Therefore, we cannot conclude that $\widetilde{DDP}_{\mathcal{S}}(\phi) \in [-1,1]$. Consequently, we also cannot conclude that $Var\left(\widetilde{DDP}_{\mathcal{S}}(\phi)\right) \in [0,1]$. As a result, there is no longer a stability guarantee for $\widetilde{DDP}_{\mathcal{S}}(\phi)$.

To summarize the aforementioned problems, incorporating $\widetilde{DDP}_{\mathcal{S}}(\phi)$ into fairness regularization and constraints to indirectly minimize $\widehat{DDP}_{\mathcal{S}}$ may encounter difficulties for two reasons. Firstly, due to the existence of the surrogate-fairness gap, minimizing $\widetilde{DDP}_{\mathcal{S}}(\phi)$ is not equivalent to minimizing $\widehat{DDP}_{\mathcal{S}}$. Secondly, if unbounded surrogate functions are employed, the uncontrollable variance of $\widetilde{DDP}_{\mathcal{S}}(\phi)$ makes it even more challenging to be an appropriate estimator of $\widehat{DDP}_{\mathcal{S}}$.

4.2 The Large Margin Points

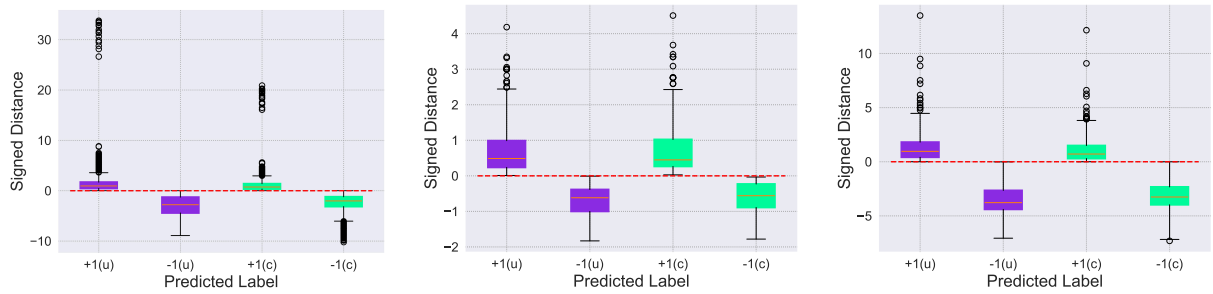
In this section, we emphasize the trouble of large margin points, which may influence the surrogate-fairness gap. In this paper, the points with too large $D_\theta(\mathbf{x})$ are called *large margin points* and others are normal points. Unfortunately, we regret to assert that these large margin points may simultaneously worsen the first two issues mentioned above. To illustrate, we take CP (linear surrogate $\phi(x) = x$) as an example. For three famous real-world data sets, Adult (Kohavi, 1996), COMPAS (Julia Angwin & Kirchner, 2016) and Bank Marketing (S. Moro & Rita, 2014), we provide the boxplot of $d_\theta(\mathbf{x})$ in the test set in Figure 2. The experimental details are in Appendix B.2. There are three main observations in Figure 2: (i) Most of the points are near the decision boundary. (ii) Over 5% points are large margin points for Adult and COMPAS. (iii) Almost all the large margin points are predicted as positive class.

Firstly, for the surrogate-fairness gap problem, the gap in Equation (1) may be amplified in the presence of such large margin points. For example, Figure 2 shows that most of the large margin points are predicted positive, so there is not a tight bound for $\sum_{(\mathbf{x},y)\in\mathcal{N}_{1a}} d_\theta(\mathbf{x})$ and $\sum_{(\mathbf{x},y)\in\mathcal{N}_{1b}} d_\theta(\mathbf{x})$. Also, Figure 2 suggests that the points with negative prediction exhibit a relatively smaller $|d_\theta(\mathbf{x})|$ comparing to those points with positive prediction. Thus, $\sum_{(\mathbf{x},y)\in\mathcal{N}_{0a}} d_\theta(\mathbf{x})$ and $\sum_{(\mathbf{x},y)\in\mathcal{N}_{0b}} d_\theta(\mathbf{x})$ are bounded (for instance, In Figure 2(b), we have $\left|\sum_{(\mathbf{x},y)\in\mathcal{N}_{0a}} d_\theta(\mathbf{x})\right| \leq 2N_{0a}$). Therefore, there is still not a tight bound for $\sum_{(\mathbf{x},y)\in\mathcal{N}_{1a}\cup\mathcal{N}_{0a}} d_\theta(\mathbf{x})$ and $\sum_{(\mathbf{x},y)\in\mathcal{N}_{1b}\cup\mathcal{N}_{0b}} d_\theta(\mathbf{x})$, which may lead to a large surrogate-fairness gap in Equation (1). Finally, if the surrogate-fairness gap becomes large, constraining the fairness surrogate function is inconsistent with the specific fairness definition, which may lead to unfair result.

Secondly, regarding the instability issue, while the majority of points are close to the decision boundary, a small number of large margin points contribute to the increased variance of $d_\theta(\mathbf{x})$, thereby influencing both $\widetilde{DDP}_{\mathcal{S}}(\phi)$ and $Var\left(\widetilde{DDP}_{\mathcal{S}}(\phi)\right)$. The presence of large margin points, along with the use of an unbounded surrogate function, surpasses the constraint on $Var\left(\widetilde{DDP}_{\mathcal{S}}(\phi)\right)$ and may result in unstable fairness guidance for the classifier. These analytical insights above will be further validated through our experiments.

5 Our Approach

We devise the general sigmoid surrogate function with fairness and stability guarantees in Section 5.1. The theory suggests that addressing the large margin points issue and obtaining a more balanced dataset contribute to a fairer classifier. Then we present our balanced surrogates in Section 5.2, which is a novel iterative approach to reduce the "gap" and thus improving fairness.



(a) Adult. (b) COMPAS. (c) Bank.

Figure 2: The boxplot for the unconstrained logistic classifier (u) and logistic classifier with fairness constraints using linear surrogate (c) in three datasets. $+1$ and -1 represent the predicted label. The red dashed line means $d_\theta(\mathbf{x}) = 0$. The orange line in the box is the median. The circles outside the box are large margin points. The rates of large margin points are 7.61%, 5.12% and 0.82% respectively.

5.1 General Sigmoid Surrogates

We generalize sigmoid function as

$$G(x) = \sigma(wx), \tag{5}$$

where $\sigma(x)$ is the sigmoid function and $w > 0$ is the parameter. The general sigmoid surrogate is flexible because of the adjustable w. Moreover, of paramount importance, it achieves a much lower surrogate-fairness gap, making it more consistent with DP, which is shown in Figure 1(b). Additionally, it enjoys stability guarantees as its values fall within the range $[0, 1]$, ensuring that $Var\left(\widetilde{DDP}_{\mathcal{S}}(G)\right) \in [0, 1]$.

5.1.1 Fairness Guarantees

The following Theorem 1 provides the upper bound of $\left|\widehat{DDP}_{\mathcal{S}}\right|$ when $G(D_\theta(\mathbf{x}))$ is close to 1 for all the points under the $\widetilde{DDP}_{\mathcal{S}}(\phi)$ fairness constraint.

Theorem 1. *We assume that $G(D_\theta(\mathbf{x})) \in [1-\gamma, 1]$, where $\gamma > 0$. $\forall \epsilon > 0$, if $\left|\widetilde{DDP}_{\mathcal{S}}(G)\right| \leq \epsilon$, then it holds:*

$$\left|\widehat{DDP}_{\mathcal{S}}\right| \leq \frac{1}{2}\epsilon + \gamma. \tag{6}$$

The proof can be found in Appendix A.3. The first term is similar to that in (4). Now with the assumption $G(D_\theta(\mathbf{x})) \in [1-\gamma, 1]$, the gap here can be limited to a small range of variation. If the general sigmoid surrogates limit $G(D_\theta(\mathbf{x}))$ to around 1 so that γ is small enough, then there are fairness guarantees for the classifier. In contrast, the gap for CP is influenced by the magnitude of $D_\theta(\mathbf{x})$ for every large margin point and thus hard to be bounded.

Remark. *In Appendix D.2, considering CP, we provide extensions of Theorem 1 to other five fairness definitions: three kinds of disparate mistreatment (Theorem 5-10 in Appendix D.2) and balance for positive (negative) class (Theorem 11-12 in Appendix D.4). In particular, CP is generalized to disparate mistreatment (Zafar et al., 2017a), and we theoretically devise the form of CP to better meet with disparate mistreatment, which is empirically validated in (Zafar et al., 2019).*

However, in some cases, we do not need to guarantee that the assumption $G(D_\theta(\mathbf{x})) \in [1-\gamma, 1]$ holds for all the points. So the theorem below relaxes the assumption by giving the upper bound of $\left|\widehat{DDP}_{\mathcal{S}}\right|$ when most of the points satisfy the assumption in Theorem 1.

Theorem 2. *We assume that k points satisfy $G(D_\theta(\mathbf{x})) \in [0, 1-\gamma]$ and others satisfy $G(D_\theta(\mathbf{x})) \in [1-\gamma, 1]$, where $\gamma > 0$. $\forall \epsilon > 0$, if $\left|\widetilde{DDP}_{\mathcal{S}}(G)\right| \leq \epsilon$, then it holds:*

$$\left|\widehat{DDP}_{\mathcal{S}}\right| \leq \frac{1}{2}\epsilon + \gamma + \underbrace{\frac{1}{2}\left(\frac{1}{N_{1a}+N_{0a}} + \frac{1}{N_{1b}+N_{0b}}\right)k}_{relaxation\ factor}. \tag{7}$$

The proof can be found in Appendix A.4. Comparing Theorem 1 with Theorem 2, the relaxation of the assumption produces an extra relaxation factor. According to Theorem 1, if w is large, then γ can be small enough, thus leading to a classifier with fairness guarantee. But an arbitrarily too large w makes training more challenging because of the diminished gradient magnitude for general sigmoid surrogate. Theorem 2 tells us that we need not to assume that all points satisfy $G(D_\theta(\mathbf{x})) \in [1-\gamma, 1]$. Few points violating the assumption (a small k) can also be tolerated. So it supports the idea that we do not have to choose a large w because a relatively small w can also guarantee fairness.

5.1.2 Insights from the Theorems

Large Margin Points Issue. An obvious takeaway of Theorem 2 is that addressing the large margin points issue makes k lower and thus obtaining a tighter bound of $\left|\widehat{DDP}_{\mathcal{S}}\right|$. While Bendekgey & Sudderth (2021) also explores the influence of outliers on the model under fairness constraints, their theoretical analysis in primarily centers on loss degeneracy under fairness constraints, whereas our paper aims to establish upper bounds concerning fairness.

Balanced Dataset. Another interesting take-away of Theorem 1-2 is that we need a *balanced dataset* to obtain a fairer classifier. The approximately same number between two sensitive groups contributes to a tighter $\left|\widehat{DDP}_{\mathcal{S}}\right|$ upper bound. First of all, for Theorem 1, the coefficient of ϵ satisfies $\frac{N^2}{4(N_{1a}+N_{0a})(N_{1b}+N_{0b})} \geq 1$. The equality holds if and only if $N_{1a}+N_{0a} = N_{1b}+N_{0b}$, which means that the two sensitive groups share the equal size. Therefore, a more balanced dataset will obtain a tighter $\left|\widehat{DDP}_{\mathcal{S}}\right|$ upper bound. Similarly, comparing to Theorem 1, we discover that those k points in Theorem 2 relaxed the original bound by a relaxation factor. The coefficient of the relaxation factor $\frac{1}{2}\left(\frac{N}{N_{1a}+N_{0a}} + \frac{N}{N_{1b}+N_{0b}}\right) \geq 2$, and the equality holds if and only if $N_{1a}+N_{0a} = N_{1b}+N_{0b}$. So, if we use a balanced dataset, then $N_{1a}+N_{0a}$ and $N_{1b}+N_{0b}$ are close to each other, thus making the fairness bounded tighter. The imbalance issue also applies to other surrogates and one can balance the dataset in advance to achieve better fairness performance.

Notably, certain theoretical investigations illuminate the positive impact of *a balanced dataset* on fostering fairness in machine learning. For example, the impossibility theorem (Bell et al., 2023) in fairness literature states that, in the context of binary classification, equalizing some specific set of multiple common performance metrics between protected classes is impossible, except in two special cases: a perfect predictor and equal *base rate* (Chouldechova, 2017; Kleinberg et al., 2017; Pleiss et al., 2017). Furthermore, the reduction of variation in group *base rates* has been demonstrated to yield a diminished lower bound for separation gap and independence gap (Liu et al., 2019). Moreover, minimizing the difference in *base rates* results in a decreased lower bound for joint error across both sensitive groups (Zhao & Gordon, 2022). In contrast, our Theorem 2 provides an elucidation from the perspective of *upper bound*. It implies that a more balanced dataset results in a tighter upper bound on the violation of DP.

Remark. *In Appendix A.6, Theorem 4 shows that* $Var\left(\widehat{DDP}_{\mathcal{S}}\right) \leq \frac{1}{4}\left(\frac{1}{N_a}+\frac{1}{N_b}\right)$. *Therefore, it also highlights the advantage of a balanced dataset in reducing variance.*

5.2 Balanced Surrogates

The naive fairness-aware training framework can be formulated as

$$\min \quad L(\theta, \mathbf{x}, y) + \rho \cdot \widetilde{DDP}_{\mathcal{S}}(\phi), \tag{8}$$

where $L(\theta, \mathbf{x}, y) = \frac{1}{N}\sum_{(\mathbf{x},y)\in\mathcal{S}} \ell(\theta, \mathbf{x}, y)$ is the empirical loss over the training set, ℓ is a convex loss function, $\widetilde{DDP}_{\mathcal{S}}(\phi)$ is the fairness regularization, and $\rho > 0$ is the coefficient. Recall in Proposition 1 that whether

$\widetilde{DDP}_{\mathcal{S}}(\phi)$ is an appropriate estimation of $\widehat{DDP}_{\mathcal{S}}$ depends on ϕ. And as stated in Proposition 1, the fairness surrogate function ϕ directly affects the surrogate-fairness gap. Thus, the existence of such "gap" motivates us that it is still necessary to reduce "gap" to improve fairness.

Our balanced surrogates approach mitigates unfairness by treating different sensitive groups differently using a parameter being updated during training. The key idea of the updating procedure is *making the magnitude of "gap" as small as possible*. It is a general plug-and-play learning paradigm for training framework using fairness surrogate functions like (8), which is validated in the experiments.

Specifically, we consider different surrogates for two sensitive groups, i.e.,

$$\phi(d_\theta(\mathbf{x}_i)) = \begin{cases} \phi_1(d_\theta(\mathbf{x}_i)), & z_i = +1. \\ \phi_2(d_\theta(\mathbf{x}_i)), & z_i = -1. \end{cases} \tag{9}$$

With (9), we rewrite $\widetilde{DDP}_{\mathcal{S}}(\phi)$ as:

$$\widetilde{DDP}_{\mathcal{S}}(\phi) = \frac{\sum_{(\mathbf{x},y)\in\mathcal{N}_{1a}\cup\mathcal{N}_{0a}} \phi_1(d_\theta(\mathbf{x}))}{N_{1a}+N_{0a}} - \frac{\sum_{(\mathbf{x},y)\in\mathcal{N}_{1b}\cup\mathcal{N}_{0b}} \phi_2(d_\theta(\mathbf{x}))}{N_{1b}+N_{0b}}, \tag{10}$$

For simplicity, we assume

$$\phi_2(x) = \lambda\phi_1(x), \tag{11}$$

where $\lambda \geq 0$ is the *balance factor* to be updated to reduce the gap. Our objective is

$$\widehat{DDP}_{\mathcal{S}} - \widetilde{DDP}_{\mathcal{S}}(\phi) = 0, \tag{12}$$

which is equivalent to a surrogate-fairness gap of zero. Plug equations (1), (10) and (11) into the equation (12), we then solve for λ:

$$\lambda = \frac{(N_{1b}+N_{0b})\sum_{(\mathbf{x},y)\in\mathcal{N}_{1a}\cup\mathcal{N}_{0a}} \phi_1(d_\theta(\mathbf{x})) - (N_{1a}N_{0b} - N_{0a}N_{1b})}{(N_{1a}+N_{0a})\sum_{(\mathbf{x},y)\in\mathcal{N}_{1b}\cup\mathcal{N}_{0b}} \phi_1(d_\theta(\mathbf{x}))} \tag{13}$$

In this way, we can first specify ϕ_1 in (9), initiate λ as λ_0, and train an unconstrained classifier with the produced θ_0 as the start point of the iteration. Then we iteratively solve (8) and compute (13) until convergence. In order to avoid oscillation and accelerate the convergence, we use exponential smoothing for λ:

$$\lambda_t = \begin{cases} \lambda_0, & t = 0. \\ \alpha\lambda'_t + (1-\alpha)\lambda_{t-1}, & t = 1, 2, \cdots, N. \end{cases} \tag{14}$$

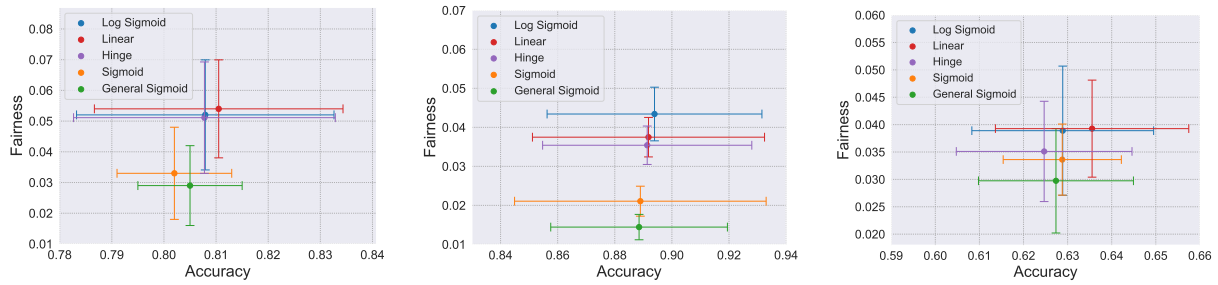
Where λ'_t comes from (13) after t iterations, λ_t is the result of λ'_t after exponential smoothing and $0 \leq \alpha \leq 1$ is the smoothing factor. Notice that $\lambda \leq 0$ is meaningless, so when this happens, we abandon this algorithm and set λ_t to 1, which recovers (8). When the difference of λ_t between two successive iterations is less than a termination threshold η, the algorithm is over. If we choose the smoothing factor α and the termination threshold η properly, then the loop will terminate after a few runs. The algorithmic representation of the balanced surrogates can be found in Appendix B.1.

6 Experiments

6.1 Experimental Setup

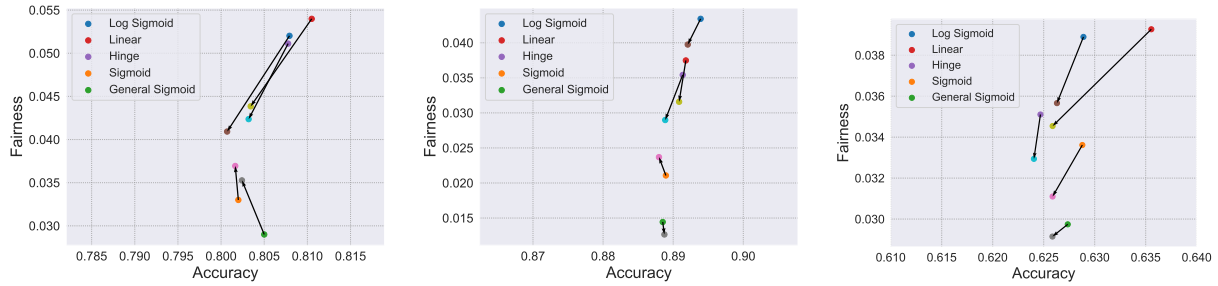
Dataset. We use three real-world datasets: Adult (Kohavi, 1996), Bank Marketing (S. Moro & Rita, 2014) and COMPAS (Julia Angwin & Kirchner, 2016), which are commonly used in fair machine learning (Mehrabi et al., 2021).

- **Adult**. The Adult dataset contains 48842 instances and 14 attributes. The goal is predicting whether the income for a person is more than $50,000 a year. We consider sex as the sensitive feature with values male and female.



(a) Adult. (b) Bank. (c) COMPAS.

Figure 3: Results of different surrogate functions.



(a) Adult. (b) Bank. (c) COMPAS.

Figure 4: Results of applying balanced surrogate to different surrogate functions. The arrow starts with the result of a surrogate and ends with the result of the same surrogate function using balanced surrogate method.

- **Bank**. The Bank Marketing dataset contains 41188 instances and 20 input features. The goal is predicting whether the client will subscribe a term deposit. We follow (Zafar et al., 2017c) and consider age as the binary sensitive attribute, which is discretized into the case whether the client's age is between 25 and 60 years.
- **COMPAS**. The COMPAS dataset was compiled to investigate racial bias in recidivism prediction. The goal is predicting whether a criminal defendant will be a recidivist in two years. We use only the subset of the data with sensitive attribute Caucasian or African-American.

Baseline. In addition to an unconstrained logistic regression classifier (denoted as 'Unconstrained'), we compare our general sigmoid surrogate (denoted as 'General Sigmoid') with other four surrogate functions below, which have also appeared in Figure 1.

- Linear surrogate (equivalent to CP) $\phi(x) = x$ (Zafar et al., 2017c) (denoted as 'Linear').
- Hinge-like surrogate $\phi(x) = \max(x + 1, 0)$ (Wu et al., 2019) (denoted as 'Hinge').
- Sigmoid surrogate $\phi(x) = \sigma(x)$ (Bendekgey & Sudderth, 2021) (denoted as 'Sigmoid').
- Log-sigmoid surrogate $\phi(x) = -\log \sigma(-x)$ (Bendekgey & Sudderth, 2021) (denoted as 'Log-Sigmoid').

6.2 Learning Fair Classifiers

We conduct two main experiments. One is the comparison among general sigmoid surrogate and other surrogate functions on classification tasks. The other is validating the effect of balanced surrogates method

by applying it to different surrogate functions. Following Bendekgey & Sudderth (2021), a linear classifier is used as the base classifier. The dataset is randomly divided into training set (70%), validation set (5%) and test set (25%). The parameter setting is discribed in Appendix B.3. We report two metrics on the test set: $\left|\widehat{DDP}_{\mathcal{S}}\right|$ (lower is better) and accuracy (higher is better), and standard deviation is shown for the metrics.

6.3 Main Results

The results are in Figures 3-4. Refer to Appendix B.5 for specific numerical results of all three datasets.

General Sigmoid Surrogate. In Figure 3, on the one hand, we observe that the general sigmoid surrogate achieves better fairness than unbounded surrogate functions (Log sigmoid, Linear and Hinge), which indicates that the general sigmoid surrogate does effectively shorten the surrogate-fairness gap and therefore improve fairness. On the other hand, comparing to the sigmoid surrogate function, our proposed surrogate not only further reduces the gap, but it is also more flexible due to the parameter w. Moreover, it is intriguing to note that the variance of fairness and accuracy of general sigmoid surrogate is also comparatively smaller than other surrogate functions, demonstrating its superiority in terms of stability. It offers a simple solution to the long-standing high variance issue observed in existing fairness-aware algorithms (Friedler et al., 2019; Ganesh et al., 2023). We provide some theoretical analysis on variance to Appendix A.6. Exploring automated methods to search for a suitable parameter w for improved fairness performance while reducing variance presents an intriguing avenue for future research.

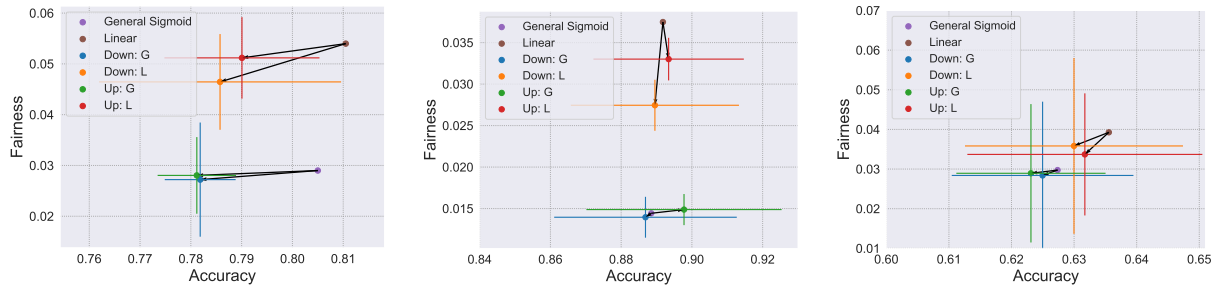
Balanced Surrogates Method. In Figure 4, we observe that balanced surrogates method succeeds in improving fairness of unbounded surrogate functions but sometimes slightly compromising fairness of bounded surrogate functions. For the reason of this phenomenon, fairness-aware algorithms aim to reduce $\left|\widehat{DDP}_{\mathcal{S}}\right|$. Firstly, fairness regularization aims at lowering $\left|\widetilde{DDP}_{\mathcal{S}}(\phi)\right|$. Secondly, the key idea of balanced surrogates is to reduce the magnitude of "gap" $\left|\widehat{DDP}_{\mathcal{S}} - \widetilde{DDP}_{\mathcal{S}}(\phi)\right|$ thus indirectly lower $\left|\widehat{DDP}_{\mathcal{S}}\right|$ (because $\left|\widehat{DDP}_{\mathcal{S}}\right| \leq \left|\widetilde{DDP}_{\mathcal{S}}(\phi)\right| + \left|\widehat{DDP}_{\mathcal{S}} - \widetilde{DDP}_{\mathcal{S}}(\phi)\right|$). However, an infinitesimal $\left|\widehat{DDP}_{\mathcal{S}}\right|$ is not always better. In Appendix A.5, we show in Theorem 3 that there is still a discrepancy between $\widehat{DDP}_{\mathcal{S}}$ and DDP, indicating that a small enough $\left|\widehat{DDP}_{\mathcal{S}}\right|$ is not equivalent to a small $|DDP|$. When the "gap" is large (such as the unbounded surrogate functions), balanced surrogates method can effectively reduce "gap" and achieves a fairer result. But when the "gap" is limited (such as sigmoid and general sigmoid), an infinitesimal magnitude of "gap" may sometimes undermine fairness instead. Interestingly, similar to the stability enhancement observed in general sigmoid surrogate, the numerical results in Appendix B.5 also show that our balanced surrogates method attains smaller variance compared to other surrogate functions. Incorporating mechanisms such as the balanced surrogate method into existing fairness-aware algorithms to enhance both fairness and stability is also a promising direction.

In Appendix B.6, we also compare our methods with other two in-processing methods: reduction (Agarwal et al., 2018) and adaptive sensitive reweighting (Krasanakis et al., 2018). The promising results also demonstrate the superiority of our methods.

6.4 Experimental Verification for the Theoretical Insights

Large Margin Points Issue. The boxplot with our general sigmoid surrogate is shown in Figure 8 in Appendix B.7. Recall that with an unbounded surrogate function ϕ, the large margin points will influence the "gap". Now they have a minor impact on the "gap" mentioned before because they are bounded ($G(D_\theta(\mathbf{x})) \leq 1$). Furthermore, the two edges almost overlap, indicating that the variation of $D_\theta(\mathbf{x})$ is small, contributing to more stable results. Overall, the general sigmoid surrogate successfully deals with the large margin points, mitigating both the surrogate-fairness gap and instability simultaneously.

Balanced Dataset. From the perspective of sensitive attribute, the Adult dataset is an *imbalanced* dataset: there are 32650 male instances and 16192 female instances. The ratio of the two groups is approximately



(a) Adult. (b) Bank. (c) COMPAS.

Figure 5: The result of balanced dataset. "G" and "L" indicates general sigmoid surrogate and linear surrogate, respectively. "Up" and "Down" correspond to upsampling and downsampling, respectively.

2:1. The Bank Marketing and COMPAS datasets also suffer from the imbalance issue. The ratio of the two groups are 39210:1978 (about 20:1) and 3175:2103 (about 3:2), respectively. Such imbalanced datasets lead to a loose bound in Theorem 2. We randomly split the dataset into training set (70%) and test set (30%). According to the number of minority group, we conduct two experiments: Downsampling and Upsampling, which means randomly downsampling (upsampling) the majority (minority) group in the original training set and form the new training set to make two demographic groups more balanced. Refer to Appendix B.4 for the details of our sampling schemes. The test set is partitioned in advance so that it is still imbalanced. We choose an unbounded surrogate function: Linear, and our bounded surrogate: general sigmoid.

The results in Figure 5 show that downsampling the majority group and upsampling the minority group contribute to a balanced dataset and a fairer result. However, in our experiments here, downsampling will lead to reduction of the training set, and upsampling will lead to replication of the training set, which may cause underfitting and overfitting problems, respectively. So the accuracy sometimes decreases. In conclusion, before fairness-aware training, we suggest using fair data augmentation strategies to obtain a balanced dataset, such as Fair Mixup (Chuang & Mroueh, 2021), and algorithms designed to address data imbalance, such as SMOTE (Chawla et al., 2002).

7 Conclusion

In this paper, we research on surrogate functions in algorithmic fairness. We derive the surrogate-fairness gap to indicate the difference between fairness surrogate function and . With boxplots, we find that unbounded surrogates are especially faced with the large margin points issue, which further amplify the "gap" and instability. To address these challenges, we propose general sigmoid surrogate with theoretically validated fairness and stability guarantees to deal with large margin points. The theoretical analysis further provides insights to the community that dealing with the large margin points issue as well as obtaining a more balanced dataset contribute to a fairer and more stable classifier. We further elaborate balanced surrogates method, which is an iterative algorithm to reduce the gap during training. It is also applicable to other fairness surrogate functions. Finally, our experiments using three real-world datasets not only validate the insights of our theorems, but also show that our methods get better fairness and stability performance.

8 Broader Impact and Ethics Statement

This study concentrates on better understanding the fairness surrogate functions in machine learning. Importantly, if someone claims the fairness guarantee of using unbounded fairness surrogate functions, it is worthy of suspicion and further investigation because of the surrogate-fairness gap issue discussed in this paper. Furthermore, the motivation of our general sigmoid surrogate and balanced surrogate methods are both centered on improving the fairness performance.

We acknowledge the sensitive nature of our study and guarantee adherence to all applicable legal and ethical standards. Our research is conducted within a safe and controlled setting to protect real-world systems' security. Only researchers who have received the appropriate clearance can access the most confidential parts of our experiments. Such measures are implemented to preserve the integrity of our research and to reduce any potential risks associated with the experiments.

Acknowledgement

We sincerely appreciate the anonymous reviewers for their helpful suggestions and constructive comments. WY, ZCL and YL were supported by the National Natural Science Foundation of China (NO.62076234); the Beijing Natural Science Foundation (NO.4222029); the Intelligent Social Governance Interdisciplinary Platform, Major Innovation & Planning Interdisciplinary Platform for the "Double First Class" Initiative, Renmin University of China; the Beijing Outstanding Young Scientist Program (NO.BJJWZYJH012019100020098); the Public Computing Cloud, Renmin University of China; the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China (NO.2021030199); the Huawei-Renmin University joint program on Information Retrieval; the Unicom Innovation Ecological Cooperation Plan; the CCF Huawei Populus Grove Fund; and the National Key Research and Development Project (NO.2022YFB2703102). ZKZ and BH were supported by the NSFC General Program No. 62376235, Guangdong Basic and Applied Basic Research Foundation Nos. 2022A1515011652 and 2024A1515012399.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Solon Barocas and Andrew D Selbst. Big data's disparate impact. *California law review*, pp. 671–732, 2016.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. http://www.fairmlbook.org.

Yahav Bechavod and Katrina Ligett. Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044*, 2017.

Andrew Bell, Lucius Bynum, Nazarii Drushchak, Tetiana Zakharchenko, Lucas Rosenblatt, and Julia Stoyanovich. The possibility of fairness: Revisiting the impossibility theorem in practice. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 400–422, 2023.

Harry Bendekgey and Erik B. Sudderth. Scalable and stable surrogates for flexible classifiers with fairness constraints. In *Advances in Neural Information Processing Systems*, 2021.

Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. A convex framework for fair regression. In *Fairness, Accountability, and Transparency in Machine Learning (FATML)*, 2017.

Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. Understanding the origins of bias in word embeddings. In *International conference on machine learning*, pp. 803–811, 2019.

Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, 2009.

Flavio P. 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*, 2017.

Simon Caton and Christian Haas. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*, 2020.

Junyi Chai and Xiaoqian Wang. Fairness with adaptive weights. In *International Conference on Machine Learning*, pp. 2853–2866, 2022.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Irene Chen, Fredrik D Johansson, and David Sontag. Why is my classifier discriminatory? *Advances in neural information processing systems*, 31, 2018.

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

Ching-Yao Chuang and Youssef Mroueh. Fair mixup: Fairness via interpolation. In *Proceedings of the International Conference on Machine Learning*, 2021.

André F Cruz, Pedro Saleiro, Catarina Belém, Carlos Soares, and Pedro Bizarro. Promoting fairness through hyperparameter optimization. In *2021 IEEE International Conference on Data Mining (ICDM)*, pp. 1036–1041, 2021.

Jiahao Ding, Xinyue Zhang, Xiaohuan Li, Junyi Wang, Rong Yu, and Miao Pan. Differentially private and fair classification via calibrated functional mechanism. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, 2018.

Sanghamitra Dutta, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, and Kush Varshney. Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing. In *International conference on machine learning*, pp. 2803–2813. PMLR, 2020.

Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015.

Benjamin Fish, Jeremy Kun, and Ádám D. Lelkes. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, 2016.

Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 329–338, 2019.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *arXiv:2309.00770*, 2023.

Prakhar Ganesh, Hongyan Chang, Martin Strobel, and Reza Shokri. On the impact of machine learning randomness on group fairness. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1789–1800, 2023.

Gabriel Goh, Andrew Cotter, Maya Gupta, and Michael P Friedlander. Satisfying real-world goals with dataset constraints. In *Advances in Neural Information Processing Systems*, 2016.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2016.

Max Hort, Zhenpeng Chen, Jie M Zhang, Federica Sarro, and Mark Harman. Bia mitigation for machine learning classifiers: A comprehensive survey. *arXiv preprint arXiv:2207.07068*, 2022.

Surya Mattu Julia Angwin, Jeff Larson and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. propublica, 2016. URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Faisal Kamiran and Toon Calders. Data pre-processing techniques for classification without discrimination. *Knowledge and Information Systems*, 2011.

Faisal Kamiran, Asim Karim, and Xiangliang Zhang. Decision theory for discrimination-aware classification. In *2012 IEEE 12th International Conference on Data Mining*, 2012.

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Proceedings of the 2012th European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part II*, 2012.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274, 2023.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *Innovations in Theoretical Computer Science (ITCS)*, 2017.

Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996.

Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris. Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *Proceedings of the 2018 world wide web conference*, pp. 853–862, 2018.

Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems*, 33:728–740, 2020.

Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. A survey on fairness in large language models. *arXiv preprint arXiv:2308.10149*, 2023.

Zachary Lipton, Julian McAuley, and Alexandra Chouldechova. Does mitigating ml's impact disparity require treatment disparity? *Advances in neural information processing systems*, 31, 2018.

Lydia T Liu, Max Simchowitz, and Moritz Hardt. The implicit fairness criterion of unconstrained learning. In *International Conference on Machine Learning*, pp. 4051–4060, 2019.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*, 2023a.

Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, et al. Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-Radiology*, 2023b.

Michael Lohaus, Michael Perrot, and Ulrike Von Luxburg. Too relaxed to be fair. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pp. 3384–3393, 2018.

Gaurav Maheshwari and Michaël Perrot. Fairgrad: Fairness aware gradient descent. *Transactions on Machine Learning Research*, 2023.

Vien V Mai and Mikael Johansson. Stability and convergence of stochastic gradient clipping: Beyond lipschitz continuity and smoothness. In *International Conference on Machine Learning*, pp. 7325–7335, 2021.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54:1–35, 2021.

Raghav Mehta, Changjian Shui, and Tal Arbel. Evaluating the fairness of deep learning uncertainty estimates in medical image analysis. In *Medical Imaging with Deep Learning*, pp. 1453–1492. PMLR, 2024.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Kirtan Padh, Diego Antognini, Emma Lejal-Glaude, Boi Faltings, and Claudiu Musat. Addressing fairness in classification with a model-agnostic multi-objective algorithm. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, 2021.

Dana Pessach and Erez Shmueli. Algorithmic fairness. *Machine Learning for Data Science Handbook: Data Mining and Knowledge Discovery Handbook*, pp. 867–886, 2023.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems*, pp. 5684–5693, 2017.

Sandro Radovanović, Boris Delibasić, and Milija Suknović. Do we reach desired disparate impact with in-processing fairness techniques? *Procedia Computer Science*, 214:257–264, 2022.

Yuji Roh, Kangwook Lee, Steven Whang, and Changho Suh. Sample selection for fair and robust training. *Advances in Neural Information Processing Systems*, 34:815–827, 2021a.

Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. Fairbatch: Batch selection for model fairness. In *International Conference on Learning Representations*, 2021b.

P. Cortez S. Moro and P. Rita. A data-driven approach to predict the success of bank telemarketing, 2014. URL https://archive.ics.uci.edu/ml/datasets/bank+marketing.

Changjian Shui, Qi Chen, Jiaqi Li, Boyu Wang, and Christian Gagné. Fair representation learning through implicit path alignment. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 20156–20175, 2022a.

Changjian Shui, Gezheng Xu, Qi Chen, Jiaqi Li, Charles X Ling, Tal Arbel, Boyu Wang, and Christian Gagné. On learning fairness and accuracy on multiple subgroups. *Advances in Neural Information Processing Systems*, 35:34121–34135, 2022b.

Changjian Shui, Justin Szeto, Raghav Mehta, Douglas L Arnold, and Tal Arbel. Mitigating calibration bias without fixed attribute grouping for improved fairness in medical imaging analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 189–198. Springer, 2023.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature medicine*, 29(8):1930–1940, 2023.

Saeid Tizpaz-Niari, Ashish Kumar, Gang Tan, and Ashutosh Trivedi. Fairness-aware configuration of machine learning libraries. In *Proceedings of the 44th International Conference on Software Engineering*, pp. 909–920, 2022.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Mingyang Wan, Daochen Zha, Ninghao Liu, and Na Zou. In-processing modeling techniques for machine learning fairness: A survey. *ACM Transactions on Knowledge Discovery from Data*, 17(3):1–27, 2023.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

Yongkai Wu, Lu Zhang, and Xintao Wu. On convexity and bounds of fairness-aware classification. In *The World Wide Web Conference*, 2019.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness beyond disparate treatment & disparate impact. *Proceedings of the 26th International Conference on World Wide Web*, 2017a.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, Krishna P. Gummadi, and Adrian Weller. From parity to preference-based notions of fairness in classification. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017b.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 2017c.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 2019.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning*, 2013.

Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. Data-centric artificial intelligence: A survey. *arXiv preprint arXiv:2303.10158*, 2023.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340, 2018.

Tao Zhang, Tianqing Zhu, Mengde Han, Jing Li, Wanlei Zhou, and Philip S. Yu. Fairness constraints in semi-supervised learning. *arXiv preprint arXiv:2009.06190*, 2020.

Chen Zhao, Changbin Li, Jincheng Li, and Feng Chen. Fair meta-learning for few-shot classification. In *2020 IEEE International Conference on Knowledge Graph (ICKG)*, 2020.

Han Zhao and Geoffrey J Gordon. Inherent tradeoffs in learning fair representations. *The Journal of Machine Learning Research*, 23(1):2527–2552, 2022.