INDUSTRIAL AND SYSTEMS ENGINEERING

ISE

Fair Supervised Learning Through Constraints On Smooth Nonconvex Unfairness-Measure Surrogates

Zahra Khatti[1], Daniel P. Robinson[1], and Frank E. Curtis[1]

[1]Department of Industrial and Systems Engineering, Lehigh University, Bethlehem, PA, 18015
USA

ISE Technical Report 25T-007

LEHIGH
UNIVERSITY

Fair Supervised Learning Through Constraints on Smooth Nonconvex Unfairness-Measure Surrogates

Zahra Khatti
Department of Industrial and Systems Engineering
Lehigh University
zak223@lehigh.edu

Daniel P. Robinson
Department of Industrial and Systems Engineering
Lehigh University
daniel.p.robinson@lehigh.edu

Frank E. Curtis
Department of Industrial and Systems Engineering
Lehigh University
frank.e.curtis@lehigh.edu

Abstract

A new strategy for fair supervised machine learning is proposed. The main advantages of the proposed strategy as compared to others in the literature are as follows. (a) We introduce a new smooth nonconvex surrogate to approximate the Heaviside functions involved in discontinuous unfairness measures. The surrogate is based on smoothing methods from the optimization literature, and is new for the fair supervised learning literature. The surrogate is a tight approximation which ensures the trained prediction models are fair, as opposed to other (e.g., convex) surrogates that can fail to lead to a fair prediction model in practice. (b) Rather than rely on regularizers (that lead to optimization problems that are difficult to solve) and corresponding regularization parameters (that can be expensive to tune), we propose a strategy that employs hard constraints so that specific tolerances for unfairness can be enforced without the complications associated with the use of regularization. (c) Our proposed strategy readily allows for constraints on multiple (potentially conflicting) unfairness measures at the same time. Multiple measures can be considered with a regularization approach, but at the cost of having even more difficult optimization problems to solve and further expense for tuning. By contrast, through hard constraints, our strategy leads to optimization models that can be solved tractably with minimal tuning.

1 Introduction

Prediction models that have been created using supervised machine learning techniques are very effective in modern practice. For a few examples relevant to this paper, researchers and practitioners have witnessed impressive performance by such models in areas such as finance [18], criminal justice [7], hiring [34, 14], and healthcare [30]. However, it has also been found that such prediction models can be affected by and/or introduce biases that can lead to unfair decisions. This can have both discriminatory and legal implications [4], such as when certain demographic groups receive favorable or unfavorable outcomes [16] to a disproportionate degree compared to other groups.

Preprint. Under review.

Overall, these biases—either inherited from historical data [22, 26] or resulting from model training procedures [45, 22]—raise significant concerns about the use of supervised machine learning models in real-world applications. For only a few examples: Research has shown that Black individuals are often offered higher interest rates for auto loans [10], small business loans for women can involve higher interest rates than for men [1], and gender biases persist in algorithmic hiring tools [14].

Various strategies have been and continue to be explored that aim to produce prediction models that may be considered fair decision-making tools. Broadly characterized, these can be referred to as pre-processing strategies that involve modifying training data before model training (see, e.g., [23]), in-processing strategies that involve modifying training algorithms in order to produce fair models (see, e.g., [41]), and post-processing strategies that adjust a model after it has been trained in order to have it yield fairer predictions (see, e.g., [33]). Our focus in this work is to devise a new in-process strategy for generating fair prediction models. In particular, we focus our attention on the incorporation of unfairness measures in the optimization problems that arise for model training, since such an approach has been shown to lead to promising improvements [17]. (The literature refers to both fairness measures and unfairness measures. We refer to unfairness measures in this paper since they lead to quantities that we aim to constrain, i.e., to have low levels of unfairness.)

The use of unfairness measures in an optimization problem that is formulated for model training has been explored previously in the literature. However, due to certain computational challenges that arise from the incorporation of such measures, certain compromises are often made that can diminish the effectiveness of such an approach. For one thing, the straightforward formulation of an explicit unfairness measure can lead to a function that is nonconvex and/or nonsmooth, which in turn causes computational challenges in the empirical risk minimization (ERM) problems that are formulated for model training [35]. As a result, convex surrogate functions [41, 15]—such as those based on covariance approximations [19]—are often employed with the aim of making the ERM problem easier to solve. However, a downside of this approach is that while it may ensure that a covariance measure (between predictions and a sensitive feature) is reduced, it does not guarantee that an original unfairness measure of interest is kept within reasonable or required limits. Second, the widespread availability of software packages and typical focus on (stochastic-)gradient-based algorithms often leads researchers and practitioners to incorporate unfairness measures only through regularization terms in the objective function [24, 5]. However, this leads to formulations that are computationally expensive to tune, and may ultimately fail to enforce fairness strictly [9]. As a result, the prediction model may possess residual biases and lead to disparate outcomes [37].

Previously proposed strategies face other challenges as well. For instance, some previously proposed strategies face difficulties when trying to incorporate multiple unfairness measures that may conflict with each other [25, 12]. This adds to the fundamental challenge of aiming to balance accuracy and fairness [29], where—if one is not careful—model performance can suffer when trying to enforce a limit on an unfairness measure [17]. In addition, some previously proposed strategies—say, involving relaxed constraints [19] or adversarial techniques [43]—can struggle when trying to balance computational efficiency with fairness guarantees and model performance [27].

1.1 Contributions

We propose a new in-process strategy for fair supervised learning with the following contributions.

- We propose the use of smooth *nonconvex* and *bounded* empirical models of discontinuous unfairness measures that, through a scalar parameterization, can approximate an unfairness measure of interest to arbitrarily high accuracy. Moreover, we prove that the unfairness measure is small whenever our proposed model value is small (e.g., if used as a constraint in an optimization formulation). This is in sharp contrast to commonly used *convex* or *unbounded* nonconvex surrogates [8, 28, 38, 41, 43], which do not provide such a guarantee, meaning that the unfairness measure can be large even when such surrogates are small. These claims are supported by our numerical experiments with multiple datasets and multiple (convex and nonconvex) surrogate functions.
- Rather than incorporate unfairness measures only through regularization terms (otherwise known as soft constraints) [24, 5, 32], we propose the use of hard constraints on surrogates of unfairness measures [41, 15]. This is the first paper to combine hard constraints with smooth nonconvex surrogates in such a way that desired bounds on unfairness measures are actually realized (to high accuracy) when a prediction model is ultimately trained. Our approach of using hard constraints

significantly reduces training cost, since otherwise a significant amount of computational expense needs to be devoted to tuning regularization parameters. We also show through our experiments that setting a higher regularization in order to ensure further decrease in an unfairness measure can have a significantly adverse affect on prediction accuracy, whereas by formulating and solving a problem with hard constraints the effect on prediction accuracy can be much more mild.

- Formulations that involve only a single unfairness measure might fail to account for different perspectives that may make a prediction model unfair [21, 41, 43]. In contrast, by employing hard constraints on accurate unfairness-measure surrogates, our approach readily allows the incorporation of multiple (potentially conflicting) unfairness measures at the same time [3, 28, 9]. This flexibility enhances the capability of our approach to mitigate multiple types of biases, ensuring broader applicability and robustness [17] for training fair prediction models.

2 Background on Unfairness Measures

The purpose of this section is to provide background on unfairness measures that are relevant for our setting. We begin by referencing a few fundamental fairness criteria that are well known in the supervised learning literature, then present probabilistic statements of fairness concepts that will, in turn, be used to formulate measures of unfairness that we will employ in our subsequent problem formulations and numerical experiments. We conclude this section with some comments about the use of convex and nonconvex surrogate functions for approximating unfairness measures.

2.1 Fundamental Fairness Criteria

Let us consider the setting of supervised learning for classification, where in contrast to the standard setting of having only a combined feature vector, we distinguish between nonsensitive and sensitive features. (One possibility for trying to produce a fair classifier is to remove the sensitive feature. However, as is well known, this runs the risk of producing an unfair classifier since nonsensitive features can be correlated with sensitive ones [20].) For the sake of simplicity of presentation, let us consider the case where there is a single sensitive feature that takes binary values and a single label that also takes binary values. Our proposed approach can readily be extended to cases involving any finite number of sensitive feature values and multiclass settings with more than two labels.

Let (X, S, Y) be a tuple of random variables that is defined with respect to a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where X represents the nonsensitive feature vector, S represents the binary sensitive feature (in $\{0, 1\}$), and Y represents the binary true label (in $\{0, 1\}$). Through a supervised learning procedure, suppose one defines a prediction function parameterized by w so that, with a *trained* w, the predicted label corresponding to a given pair of features is given by $\hat{Y} \equiv \hat{Y}(X, S, w)$. Three fundamental fairness criteria that are recognized in the literature are independence, separation, and sufficiency [3]. The independence criterion requires that the prediction $\hat{Y}$ is independent of the sensitive feature S; the separation criterion requires that the prediction $\hat{Y}$ is independent of the sensitive feature S conditioned on the true label Y; and the sufficiency criterion requires that the true label Y is independent of the sensitive feature S conditioned on the prediction $\hat{Y}$.

For our purposes, we focus primarily on unfairness measures based on the independence criterion, although our proposed strategy can be extended readily to the context of these alternative criteria.

2.2 Probabilistic Statements of Fairness

From the aforementioned fundamental criteria, various specific probabilistic statements of fairness have been derived [17]. Here, we provide two related examples based on the independence criterion.

Demographic Parity (or Statistical Parity). Demographic parity requires that a model's predictions are independent of the sensitive attribute [42]. For our setting, this can be expressed as

$$\mathbb{P}(\hat{Y}(X, S, w) = 1 | S = 1) = \mathbb{P}(\hat{Y}(X, S, w) = 1 | S = 0). \tag{1}$$

Later on, we will discuss the empirical violation of this equation as an unfairness measure.

Disparate Impact. Disparate impact relates to the requirement that selection rates are similar regardless of the value of the sensitive feature. The aim to avoid disparate impact can be seen, e.g., in the "four-fifths rule" in US employment law [4, 16, 31]. Stated mathematically in the setting of binary classification, the requirement to avoid disparate impact between two groups can be stated as

$$\begin{aligned} &\delta\mathbb{P}(\hat{Y}(X,S,w)=1|S=1) \leq \mathbb{P}(\hat{Y}(X,S,w)=1|S=0) \\ \text{and } &\delta\mathbb{P}(\hat{Y}(X,S,w)=1|S=0) \leq \mathbb{P}(\hat{Y}(X,S,w)=1|S=1), \end{aligned} \tag{2}$$

where $\delta \in [0,1]$ is a threshold parameter (e.g., $\delta = 0.8$ in the context of the four-fifths rule). Later on, we will employ empirical approximations of the terms in these inequalities in order to formulate constraints (on smooth nonconvex surrogates) to be imposed during model training.

2.3 Empirical Surrogates for Probabilistic Statements of Fairness

The probabilistic statements of fairness from the previous subsection can be incorporated into optimization problem formulations designed for prediction model training if and only if they are replaced by empirical approximations. Suppose that w represents trainable parameters, such as of a deep neural network, and that a set of feature-label tuples $\{(x_i,s_i,y_i)\}_{i\in[N]}$ is available for prediction model training, where $[N] := \{1,\dots,N\}$. For each $i \in [N]$, let the output of the neural network be given by $\mathcal{N}(x_i,s_i,w)$ and suppose that, for a given threshold $\tau \in \mathbb{R}$ and with $\mathbb{1}$ denoting the indicator function, the predicted label for (x_i,s_i) is given by

$$\hat{y}(x_i,s_i,w) = \mathbb{1}\{\mathcal{N}(x_i,s_i,w) > \tau\}.$$

For each value $s \in \{0,1\}$, the probability of a positive prediction over group s can be estimated as

$$\mathbb{P}(\hat{Y}(X,S,w)=1|S=s) \approx \frac{\sum_{\{i\in[N]:s_i=s\}} \hat{y}(x_i,s_i,w)}{N_s}, \text{ where } N_s := \sum_{i\in[N]} \mathbb{1}\{s_i = s\}.$$

However, since the indicator function is discontinuous, it cannot be used directly to formulate an optimization problem that is to be solved using an algorithm for continuous optimization [37]. Thus, as is common in the literature, let us suppose that for all $i \in [N]$ one approximates

$$\hat{y}(x_i,s_i,w) \approx \phi(t(x_i,s_i,w)), \text{ where } t(x_i,s_i,w) := \mathcal{N}(x_i,s_i,w) - \tau. \tag{3}$$

Here, $t(x_i,s_i,w)$ represents the distance-to-threshold function for data point $i \in [N]$ and $\phi : \mathbb{R} \to \mathbb{R}$ aims to approximate the step function from 0 to 1 (at the origin). Various such approximations have been explored in the literature for approximating fairness measures and other purposes as well. To list a few, we mention the linear function $\phi(t) = t$ [41], the max function $\phi(t) = \max\{t,0\}$ [39], the sigmoid function $\phi(t) = \sigma(t) := (1+e^{-t})^{-1}$ [6], and the hyperbolic tangent function $\phi(t) = \tanh(t)$. For our purposes in the next section, we focus on the sigmoid function and another approximation—based on a smoothing function that is common in the optimization literature—for use in the optimization problem formulations that we propose for prediction model training.

A question that arises in the use of a surrogate to approximate a step function is how large discrepancies can be once the model is trained. The following theorem, derived from Yao et al. [40], establishes a revealing bound on such discrepancies *when the codomain of ϕ is the unit interval and the shifted function $\phi - \frac{1}{2}$ is symmetric about the origin*; a proof is given in Appendix A to account for minor changes in the properties of ϕ as compared to [40]. It is important to note that while the corresponding theorem from [40] refers to bounds on absolute values of differences of empirical estimates of probabilities, the numerical studies in that paper involve the use of regularization for model training, meaning that such tight constraints were not necessarily satisfied in practice. This further motivates a unique feature of our work in this paper, which shows that by enforcing hard constraints we can meet the useful bounds stated in this theorem.

Theorem 1. *Suppose $\phi : \mathbb{R} \to [0,1]$, the function $\phi - \frac{1}{2}$ is symmetric about the origin, and for some $\gamma \in (0,\frac{1}{2})$ one has $\phi(t(x_i,s_i,w)) \in [0,\gamma] \cup [1-\gamma,1]$ for all $i \in [N]$. Suppose also that, for some $\epsilon \in (0,\infty)$, the surrogate of the empirical estimate of the violation of demographic parity has*

$$|c_{dp}(w)| \leq \epsilon, \text{ where } c_{dp}(w) := \frac{\sum_{\{i\in[N]:s_i=1\}} \phi(t(x_i,s_i,w))}{N_1} - \frac{\sum_{\{i\in[N]:s_i=0\}} \phi(t(x_i,s_i,w))}{N_0}. \tag{4}$$

Then the actual empirical estimate of the violation of demographic parity has

$$|\bar{c}_{dp}(w)| \leq \epsilon + \gamma, \;\; \textit{where} \;\; \bar{c}_{dp}(w) := \frac{\sum_{\{i\in[N]:s_i=1\}} \hat{y}(x_i, s_i, w)}{N_1} - \frac{\sum_{\{i\in[N]:s_i=0\}} \hat{y}(x_i, s_i, w)}{N_0}.$$

Let us close this section by mentioning that the common covariance-based surrogate for independence [41] can be understood in terms of a specific surrogate of the form mentioned above. A benefit of this surrogate is that it is convex. However, a downside is that it does not readily offer a means to leverage the useful result in Theorem 1. Specifically, the surrogate is given by

$$\begin{aligned} \text{cov}(\hat{Y}(X,S,w), S) &= \mathbb{E}[(\hat{Y}(X,S,w) - \mathbb{E}[\hat{Y}(X,S,w)])(S - \mathbb{E}[S])] \\ &\approx \frac{1}{N} \sum_{i\in[N]} (s_i - \bar{s}) \cdot \mathcal{N}(x_i, s_i, w) =: c_{\text{cov}}(w), \end{aligned}$$

where $\bar{s} = \frac{1}{N}\sum_{i\in[N]} s_i$. The following theorem (see [6, 40]) shows that this surrogate is proportional to c_{dp} from (4) with the specific choice of $\phi(t) = t$ (linear function), namely,

$$\begin{aligned} &\mathbb{E}[\hat{Y}(X,S,w)|S=1] - \mathbb{E}[\hat{Y}(X,S,w)|S=0] \\ &\approx \frac{\sum_{\{i\in[N]:s_i=1\}} t(x_i, s_i, w)}{N_1} - \frac{\sum_{\{i\in[N]:s_i=0\}} t(x_i, s_i, w)}{N_0} =: c_{\text{dp}}^{\phi(t)=t}(w). \end{aligned}$$

Theorem 2. *The functions c_{cov} and $c_{dp}^{\phi(t)=t}$ satisfy $c_{cov}(w) = \frac{N_0 \cdot N_1}{N^2} \cdot c_{dp}^{\phi(t)=t}(w)$ for all w.*

The proportionality shown in Theorem 2 suggests that one might be able to enforce a bound on the violation of demographic parity by enforcing a bound of the form $|c_{\text{cov}}(w)| \leq \epsilon$ for some $\epsilon \in (0, \infty)$. However, the useful bound in Theorem 1 is not readily applicable to this setting since the codomain of the linear function ϕ defined by $\phi(t) = t$ is unbounded; thus, it is essentially impossible to ensure that the surrogate yields, for some $\gamma \in (0, \frac{1}{2})$ and all $i \in [N]$, the inclusion $\phi(t(x_i, s_i, w)) \in [0, \gamma] \cup [1-\gamma, 1]$ for all points in the training set. Therefore, $|c_{\text{cov}}(w)| \leq \epsilon$ does not necessarily enforce a useful bound on the actual empirical estimate of violation of demographic parity, which motivates our choice of studying more accurate, nonconvex and bounded surrogates.

3 Methodology

Motivated by our prior discussions, in this section we present our proposed methodology that combines smooth, nonconvex, and accurate surrogate functions for unfairness measures with hard constraints that are to be imposed during model training. We emphasize that this combination leads to a *tractable* training paradigm where the optimization problem can be solved using stochastic Newton/SQP techniques from the recent literature [13]. The benefits of our proposed methodology are realized in practice in the experimental results that we present in the following section.

3.1 Smooth, Nonconvex, and Accurate Surrogates

At the heart of the definition of a surrogate function for our setting is an approximation of the step function from 0 to 1 (at the origin), also known as the Heaviside function. In particular, due to Theorem 1, we are interested in bounded approximations, say, ones with a codomain of $[0, 1]$. A common choice with these properties is the sigmoid function, so that is one of the choices that we consider for our methodology and experiments. However, we also consider a second approximation that employs ideas from smoothing techniques from the mathematical optimization literature; see, e.g., [11]. In our experiments, we often find better results for this second approximation—which we refer to as the smoothed-step function—as compared to the sigmoid function.

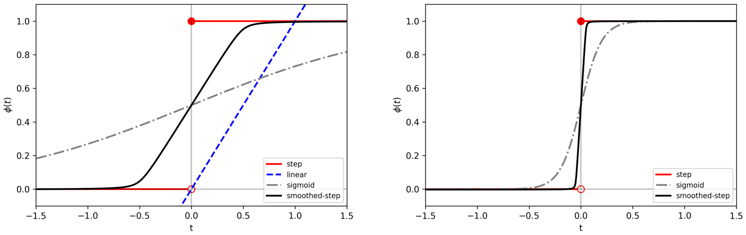
A piecewise-affine approximation of the step function that, as required by Theorem 1, has a graph that passes through $(0, 0.5)$ is given by $\phi : \mathbb{R} \to [0, 1]$ with $\phi(t) = \min\{\max\{0, t + 0.5\}, 1\}$ for all $t \in \mathbb{R}$. This can be viewed as a composite involving two max functions; indeed, one finds

$$\phi(t) = \underline{\phi}(\overline{\phi}(t)), \;\; \text{where} \;\; \underline{\phi}(\bar{t}) = \min\{\bar{t}, 1\} = 1 - \max\{1 - \bar{t}, 0\} \;\; \text{and} \;\; \overline{\phi}(t) = \max\{0, t + 0.5\}.$$

Now employing the smooth approximation given by $\max\{t, 0\} \approx \frac{1}{2}(t + \sqrt{t^2 + \mu})$, where $\mu \in (0, \infty)$ is a user-defined smoothing parameter, one obtains the smooth, nonconvex approximation

$$\phi_\mu(t(x_i, s_i, w)) := 1 - \tfrac{1}{2}\left(1 - \overline{\phi}_\mu(t(x_i, s_i, w)) + \sqrt{(1 - \overline{\phi}_\mu(t(x_i, s_i, w)))^2 + \mu}\right)$$
$$\text{where } \overline{\phi}_\mu(t(x_i, s_i, w)) := \tfrac{1}{2}\left(t(x_i, s_i, w) + \tfrac{1}{2} + \sqrt{(t(x_i, s_i, w) + \tfrac{1}{2})^2 + \mu}\right).$$

Figure 1 shows the step function (Heaviside function) along with linear, sigmoid, and smoothed-step approximations. We also want to emphasize that, to obtain improved results in practice, these functions should be scaled in order to encourage values that are closer to 0 or 1 [40]; in particular, in place of $\sigma(t)$ and $\phi_\mu(t)$ one should consider $\sigma(\alpha t)$ and $\phi_\mu(\alpha t)$ for $\alpha \in (0, \infty)$, likely greater than 1. Such scalings make the derivative of the function larger near the origin. Overall, such scaling is beneficial in terms of Theorem 1, but one might wonder if it will make the training problem more difficult to solve. We did not find this to be the case with our proposed algorithm, discussed next. Indeed, overall, our experiments demonstrate improved results with larger scaling values.



(a) Surrogate functions

(b) Scaled surrogate functions

Figure 1: On the left, graphs of the step/Heaviside function ($\mathbb{1}\{t \geq 0\}$), linear function ($\phi(t) = t$), sigmoid function ($\phi(t) = \sigma(t)$), and smoothed-step function ($\phi_\mu(t)$ defined in this section). On the right, graphs of a scaled sigmoid function ($\sigma(10t)$) and scaled smoothed-step function ($\phi_\mu(10t)$) to illustrate that scaling can make the functions more closely approximate the step function.

3.2 Complete Training Problem with Unfairness Constraints

Given a set of feature-label tuples $\{(x_i, s_i, y_i)\}_{i\in[N]}$, prediction function $\mathcal{N}$, loss function ℓ, regularization function r, regularization parameter $\lambda \in (0, \infty)$, (hard) constraint function c, and pair of bounds (l, u), we propose that a fair prediction model can be trained in a supervised learning context by solving the following *tractable* optimization problem with unfairness constraints:

$$\min_{w} \sum_{i\in[N]} \ell(\mathcal{N}(x_i, s_i, w), y_i) + \tfrac{1}{\lambda} r(w) \ \text{ subject to } \ l \leq c(w) \leq u. \tag{5}$$

For example, $\mathcal{N}$ may be a (deep) neural network with w as the trainable parameters, ℓ can be any typical loss function for classification, r may be a sparsity-promoting regularizer or one aimed to penalize a prescribed surrogate for an unfairness measure, and c may be a vector-valued function corresponding to a surrogate for an unfairness measure (or surrogates for multiple unfairness measures). In our experiments in Section 4, we consider for illustrative purposes a regularization function based on the desire to impose demographic parity, namely, $r_{\text{dp}}(w) = |q(w)|^2$ where

$$q(w) = \frac{\sum_{\{i\in[N]:s_i=1\}} \phi(t(x_i, s_i, w))}{N_1} - \frac{\sum_{\{i\in[N]:s_i=0\}} \phi(t(x_i, s_i, w))}{N_0} \tag{6}$$

and ϕ is a prescribed approximation of the step/Heaviside function and t is defined as in (3). Given a trained prediction function determined by w, one can evaluate its performance through $\bar{r}_{\text{dp}}(w)$, which is evaluated in the same way as $r_{\text{dp}}(w)$ except with $\phi(t(\cdot))$ replaced by $\hat{y}(\cdot)$. As for the constraint function c, we focus primarily on one to constrain disparate impact. Specifically, given

$\delta \in [0,1]$, we consider $l = (-\infty, -\infty)$, $u = (0,0)$, and $c = (c_{\text{di},1}, c_{\text{di},2})$ where

$$\begin{aligned} c_{\text{di},1}(w) &= \delta \left(\frac{\sum_{\{i \in [N]: s_i = 0\}} \phi(t(x_i, s_i, w))}{N_0} \right) - \frac{\sum_{\{i \in [N]: s_i = 1\}} \phi(t(x_i, s_i, w))}{N_1} \\ \text{and} \;\; c_{\text{di},2}(w) &= \delta \left(\frac{\sum_{\{i \in [N]: s_i = 1\}} \phi(t(x_i, s_i, w))}{N_1} \right) - \frac{\sum_{\{i \in [N]: s_i = 0\}} \phi(t(x_i, s_i, w))}{N_0}. \end{aligned} \tag{7}$$

Given a trained prediction function determined by w, one can evaluate any potential violation of these constraints through $c_{\text{di}}(w) = \max\{c_{\text{di},1}(w), c_{\text{di},2}(w)\}$, where in particular a violation occurs if and only if this value is greater than 0. One can also consider $\bar{c}_{\text{di}}(w) = \max\{\bar{c}_{\text{di},1}(w), \bar{c}_{\text{di},2}(w)\}$, where these functions are defined similarly to $c_{\text{di},1}$ and $c_{\text{di},2}$, except with $\phi(t(\cdot))$ replaced by $\hat{y}(\cdot)$.

For the sake of tractability, the constraints can be formulated with only a subset of a full training dataset and one can employ stochastic objective gradients. One needs to be careful with a regularizer defined through (6) in order to ensure that the stochastic gradient estimate is unbiased. The following informal theorem shows that unbiased stochastic estimates of the gradient of q in (6) can be obtained as long as the mini-batches involve the same number of data points with $s_i = 0$ and $s_i = 1$ in all iterations, which is straightforward to enforce. A formal theorem and proof are in Appendix B.

Theorem 3. *(Informal) Suppose that N_0 points with $s_i = 0$ are chosen uniformly at random and N_1 points with $s_i = 1$ are chosen uniformly at random and that these points are used to compute a stochastic estimate of $\nabla q(w_k)$, where q is defined in* (6). *Then, the estimate is unbiased.*

4 Experiments

Our experiments focus on two benchmark datasets: *Adult* [2] and *Law* School [36]. Detailed statistics for these datasets are provided in Appendix C. Adult contains 32,561 data points that are split at a ratio of 80:20 into training data (26,048 points) and testing data (6,513 points). Each data point has 90 features (after one-hot encoding) with gender as the sensitive attribute. Law contains 20,798 data points that are split at a ratio of 80:20 into training data (16,638 points) and testing data (4,160 points). Each data point has 11 features with race as the sensitive attribute.

Our proposed approach is agnostic to the prediction model. For our experiments for Adult, the prediction model was a feed-forward neural network with two hidden layers (with 128 and 64 nodes, respectively), Leaky ReLU activation at the hidden layers, and sigmoid activation at the output layer. For Law, the prediction model was a linear neural network with sigmoid activation. In all of the experimental results displayed in this section, the models were trained using a sequential quadratic optimization (SQP) method, Binary Cross-Entropy loss (BCELoss), and 500 epochs with an initial learning rate of 0.5 that was adjusted dynamically during training; see Appendix D for further details about the training algorithm. The results in this section use full-batch gradients, but Appendix E contains results with stochastic gradients as well, using the result of Theorem 3.

All results presented here are with respect to the training data. Appendix E shows that prediction accuracies, unfairness measures, etc. are similar for all experiments when one considers testing data.

4.1 Effect of Surrogate Function on Actual Limit on Unfairness that is Achieved

Our first set of experiments demonstrates that having a loose approximation of an unfairness measure can cause a trained prediction model to be less fair than desired. Toward this end, we first trained prediction models using unscaled surrogate functions. Constraints were imposed as in (7) for various values of δ. For each δ, we considered the trained model and determined the maximum value of δ, call it $\hat{\delta}$, such that $c_{\text{di}}(w) \leq 0$ (a similar calculation is also performed based on $\bar{c}_{\text{di}}(w) \leq 0$); the results are plotted in Figure 2. In particular, the graphs labeled $\phi(t)$ and $\hat{y}$ are the plots of the $(\delta, \hat{\delta})$ pairs that correspond to $c_{\text{di}}(w) \leq 0$ and $\bar{c}_{\text{di}}(w) \leq 0$, respectively. These results show that the desired limit on disparate impact was almost achieved when the smoothed-step surrogate was employed, but that the desired limit was not nearly achieved when the sigmoid surrogate was employed. By contrast, when the surrogate functions are scaled (with $\alpha = 50$) as described at the end of §3.1, one obtains the results shown in Figure 3, where for both the smoothed-step and sigmoid functions one finds that the imposed δ and the realized $\hat{\delta}$ values are tightly matched. Note that these experiments

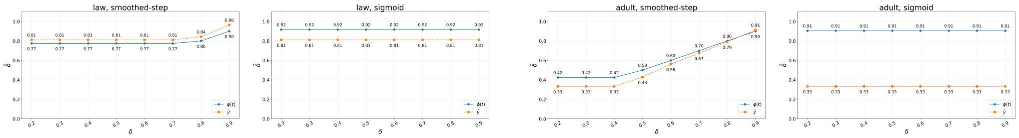


Figure 2: Levels of disparate impact actually achieved ($\hat{\delta}$) when prediction models are trained with constraints as in (7) for varying values of δ *without* scaling of the surrogate functions. The surrogate approximations not being tight causes large gaps between the levels of disparate impact desired and the levels actually achieved. The graphs indicated by $\phi(t)$ show the values such that $c_{\text{di}}(w) \leq 0$, whereas the graphs indicated by $\hat{y}$ show the values such that $\bar{c}_{\text{di}}(w) \leq 0$.

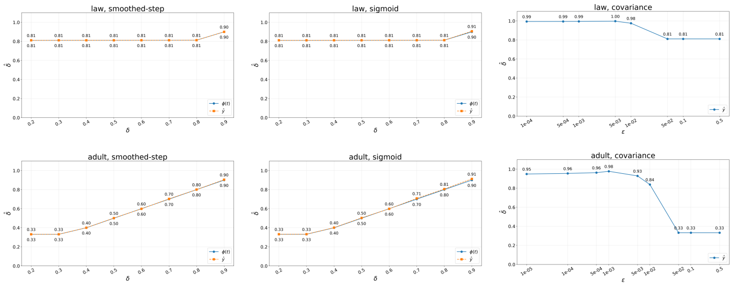


Figure 3: Levels of disparate impact actually achieved ($\hat{\delta}$) when prediction models are trained with constraints as in (7) for varying values of δ *with* scaling of the surrogate functions. These results should be contrasted with those in Figure 2. In particular, it should be observed that scaling the surrogate functions leads to much tighter correspondence between δ and $\hat{\delta}$ when the constraints are tight. Also, the plots on the right are the levels of disparate impact actually achieved when a constraint on a covariance surrogate (less than or equal to ϵ) is imposed.

show that the constraints on disparate impact become *active* for Law and Adult around $\delta \in [0.7, 0.8]$ and $\delta \in [0.3, 0.4]$, respectively. In Figure 3, we also show that one can impose a constraint on disparate impact through a constraint on a covariance surrogate. However, in such an approach, it is nontrivial to determine the specific limit on the covariance (ϵ) that corresponds to a specific desired limit on disparate impact. *This demonstrates that our use of accurate, nonconvex surrogates can make it easier to impose a specific desired limit on an unfairness measure such as disparate impact.*

Figure 4 confirms our prior claim that, through the use of hard constraints on nonconvex surrogates of unfairness measures, one is able to train prediction models that also offer high-accuracy predictions. The results in this figure correspond to those in Figure 3, i.e., with the models trained through scaled surrogate functions. The prediction accuracies remain high, even as the value of δ or ϵ is adjusted to the point where the imposed hard constraints become active.

4.2 Comparison of Constraint-Only-Based vs. Regularization-Only-Based Approaches

Our second set of experiments demonstrates that enforcing hard constraints on unfairness-measure surrogates yields better prediction models and offers more precise control over unfairness measure values. We consider two approaches: A constraint-only-based approach that imposes hard constraints as defined by (7) with no regularization (i.e., $\lambda = \infty$ in (5)) versus a regularization-only-based approach that employs $\delta = 0$ and varying values of the regularization parameter λ in (5).

As depicted in Figure 5, the regularization-only-based approach has notable limitations. First, the unfairness measure does not always improve monotonically with decreasing λ (see the results for the Law data set). This unexpected behavior is a consequence of formulating optimization problems that are increasingly difficult to solve as λ decreases. Second, the relationship between λ and the resulting unfairness measure is nonlinear, thus requiring expensive tuning to reach a desired level of unfairness. Overall, it is clear that the complicated relationship between the regularization parameter and unfairness measure makes it difficult to effectively use a regularization-only-based approach.

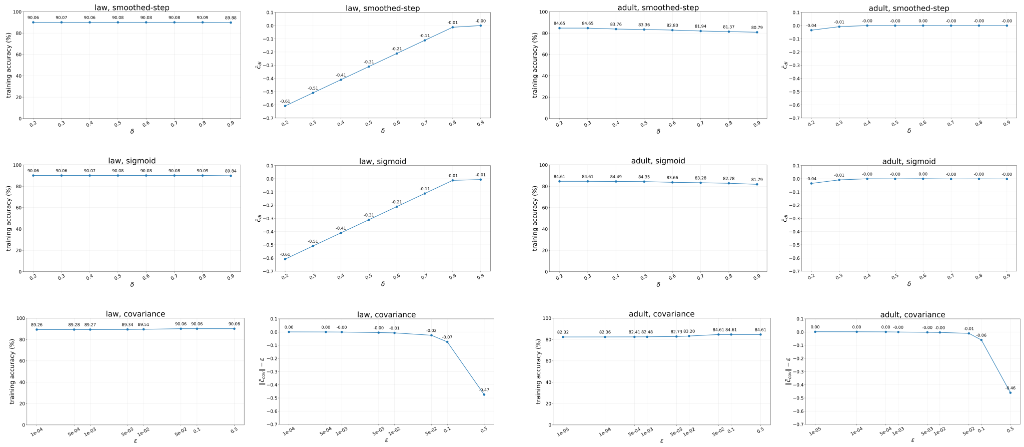


Figure 4: Training accuracy and constraint violation measures for smoothed-step, sigmoid, and covariance models on the Law and Adult datasets. The plots show training accuracy alongside constraint violation metrics for different surrogate functions. The results indicate that fairness constraints can be satisfied without compromising prediction accuracy.

In contrast, the constraint-only-based approach ensures monotonicity in the unfairness measure and allows a desired bound on unfairness (defined through group-specific probability ratios) to be enforced *explicitly*. Together, these properties allow the constraint-only-based approach to avoid requiring any significant hyperparameter tuning. It should also be noted that the constraint-only-based approach maintains high accuracy across many values of the unfairness thresholds, which is yet another advantage of it over the regularization-only-based approach.

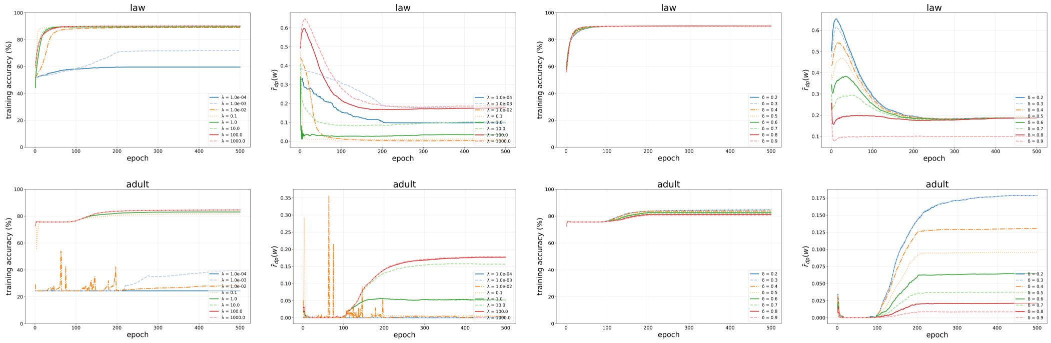


Figure 5: Comparison between the constraint-only (two right columns) and regularization-only (two left columns) approaches for enforcing fairness using the smoothed-step surrogate on Law and Adult datasets. The constraint-only-based approach consistently meets the targeted threshold (δ) with minimal impact on accuracy, while the regularization-only-based method exhibits unpredictable outcomes and greater accuracy loss, underscoring an advantage of using explicit constraints.

4.3 Simultaneous Implementation of Multiple Unfairness Measures

The proposed framework can easily accommodate multiple unfairness constraints at the same time. To demonstrate this capability, we tested our framework when constraints on both disparate impact and equal impact were enforced simultaneously, each with varying thresholds. As shown in Table 1, combining multiple unfairness constraints often leads to a reduction in accuracy that exceeds the reduction observed when only one of the constraints is enforced, especially as the thresholds become more demanding. Since our constraint-only-based approach consistently satisfies both unfairness criteria when enforced as constraints, it highlights the practical value of our approach in real-world applications when multiple unfairness measures must be taken into account.

Table 1: Accuracies (overall, male, and female) and constraint violations when disparate impact (DI) and/or equal impact (EI) constraints are enforced with $\delta = 0.8$ for the Law and Adult datasets. Enforcing both constraints reduces overall accuracy, particularly for Adult, but the constraint-only method consistently satisfies both unfairness constraints, as indicated by non-positive constraint model violations (c_{di} and c_{ei}) and non-positive constraint violation measures ($\bar{c}_{di}$ and $\bar{c}_{ei}$).

Dataset	**Scenario**	**Overall Acc**	**Male Acc**	**Female Acc**	c_{di}	c_{ei}	$\bar{c}_{di}$	$\bar{c}_{ei}$
Law ($\delta = 0.8$)	Only DI	90.095	92.3973	77.9039	-0.0124	0.0602	-0.0127	-0.1082
	Only EI	89.8606	92.4116	76.3526	-0.1186	-0.0	-0.1187	-0.1704
	Both DI-EI	90.089	92.3973	77.8661	-0.0124	-0.1076	-0.0131	-0.1082
Adult ($\delta = 0.8$)	Only DI	81.3729	76.3707	91.5002	0.0	0.1088	-0.0	0.1102
	Only EI	84.6437	81.022	91.9763	0.1255	-0.0371	0.1251	-0.0383
	Both DI-EI	75.7294	69.259	88.8295	0.0001	0.0	-0.0001	0.0008

5 Discussion and Conclusion

We present a hard-constraint-based approach to fair supervised learning that accurately yields prediction models that meet prescribed limits on unfairness. Additionally, our method readily allows limits on multiple unfairness measures simultaneously. A limitation of our approach is that there are slightly higher computational costs to train a model with hard constraints, although these additional costs are outweighed by the savings that our approach offers in terms of parameter tuning costs.

References

[1] Alberto F Alesina, Francesca Lotti, and Paolo Emilio Mistrulli. Do women pay more for credit? evidence from italy. *Journal of the European Economic Association*, 11(suppl_1):45–66, 2013.

[2] Arthur Asuncion, David Newman, et al. Uci machine learning repository, 2007.

[3] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning: Limitations and opportunities. *NIPS Tutorial*, 2017.

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

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

[6] Henry C Bendekgey and Erik Sudderth. Scalable and stable surrogates for flexible classifiers with fairness constraints. *Advances in Neural Information Processing Systems*, 34:30023–30036, 2021.

[7] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021.

[8] Felix Biggs, Tobias Friedrich, and Martin Schirneck. Too relaxed to be fair. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119, pages 960–970. PMLR, 2020.

[9] L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 319–328, 2019.

[10] Kerwin Kofi Charles, Erik Hurst, and Melvin Stephens Jr. Rates for vehicle loans: race and loan source. *American Economic Review*, 98(2):315–320, 2008.

[11] Xiaojun Chen. Smoothing methods for nonsmooth, nonconvex minimization. *Mathematical Programming*, 134(1):71–99, August 2012.

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

[13] Frank E Curtis, Suyun Liu, and Daniel P Robinson. Fair machine learning through constrained stochastic optimization and an ϵ-constraint method. *Optimization Letters*, pages 1–17, 2023.

[14] Jeffrey Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. In *Ethics of data and analytics*, pages 296–299. Auerbach Publications, 2022.

[15] Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. *Advances in neural information processing systems*, 31, 2018.

[16] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268, 2015.

[17] 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. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 329–338, 2019.

[18] Andreas Fuster, Paul Goldstein, Tarun Goldstein, and William Miner. Predictably unequal? the effects of machine learning on credit markets. *The Journal of Finance*, 77(1):5–47, 2022.

[19] Gabriel Goh, Andrew Cotter, Maya Gupta, and Michael P Friedlander. Satisfying real-world goals with dataset constraints. *Advances in Neural Information Processing Systems*, 29, 2016.

[20] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS symposium on machine learning and the law*, volume 1, page 11. Barcelona, Spain, 2016.

[21] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29, 2016.

[22] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–16, 2019.

[23] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

[24] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 35–50, 2012.

[25] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *Innovations in Theoretical Computer Science Conference*, 2017.

[26] Kristian Lum and William Isaac. To predict and serve? the disparate impact of historical data in predictive policing. *Significance*, 13(5):14–19, 2016.

[27] Natalia Martinez, Martin Bertran, and Guillermo Sapiro. Minimax pareto fairness: A multi objective perspective. *International Conference on Machine Learning*, pages 6755–6764, 2020.

[28] Jeremie Mary, Clement Calauzenes, and Noureddine El Karoui. Fairness-aware learning for continuous attributes and treatments. *International Conference on Machine Learning*, pages 4382–4391, 2019.

[29] Aditya Krishna Menon and Robert C Williamson. The cost of fairness in binary classification. *Proceedings of the Conference on Fairness, Accountability and Transparency*, pages 107–118, 2018.

[30] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[31] Code of Federal Regulations. §1607.4, Section D. `https://www.ecfr.gov/current/title-29/subtitle-B/chapter-XIV/part-1607/subject-group-ECFRdb347e844acdea6/section-1607.4`, 2025. [Online; accessed 13-May-2025].

[32] Manish Padala and Sankarshan Gupta. Fnnc: Achieving fairness through neural networks. *International Joint Conference on Artificial Intelligence*, pages 2277–2283, 2020.

[33] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. *Advances in neural information processing systems*, 30, 2017.

[34] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. Mitigating bias in algorithmic hiring: Evaluating claims and practices. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 469–481, 2020.

[35] Vladimir Vapnik. *The nature of statistical learning theory*. Springer, 1999.

[36] Linda F Wightman. Lsac national longitudinal bar passage study. lsac research report series. 1998.

[37] Blake Woodworth, Suriya Gunasekar, Mikhail I Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. *Conference on Learning Theory*, pages 1920–1953, 2017.

[38] Yongkai Wu, Lu Zhang, and Xintao Wu. Fairness-aware classification: Criterion, convexity, and bounds. *arXiv preprint arXiv:1809.04737*, 2018.

[39] Yongkai Wu, Lu Zhang, and Xintao Wu. On convexity and bounds of fairness-aware classification. In *The World Wide Web Conference*, pages 3356–3362, 2019.

[40] Wei Yao, Zhanke Zhou, Zhicong Li, Bo Han, and Yong Liu. Understanding fairness surrogate functions in algorithmic fairness. *arXiv preprint arXiv:2310.11211*, 2023.

[41] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. *Artificial Intelligence and Statistics*, pages 962–970, 2017.

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

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

[44] Hongchao Zhang and William W Hager. A nonmonotone line search technique and its application to unconstrained optimization. *SIAM journal on Optimization*, 14(4):1043–1056, 2004.

[45] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, 2017.

A Proof of Theorem 1

For the purposes of the proof, let us simplify notation and define $\hat{y}_i(w) := \hat{y}(x_i, s_i, w)$ along with $t_i(w) := t(x_i, s_i, w)$ for all $i \in [N]$. Let us also define the sample sets and their sizes given by

$$\begin{aligned}
\mathcal{N}_0^- &:= \{i \in [N] : s_i = 0 \wedge \hat{y}_i(w) = 0\},\ N_0^- := |\mathcal{N}_0^-|,\\
\mathcal{N}_0^+ &:= \{i \in [N] : s_i = 0 \wedge \hat{y}_i(w) = 1\},\ N_0^+ := |\mathcal{N}_0^+|,\\
\mathcal{N}_1^- &:= \{i \in [N] : s_i = 1 \wedge \hat{y}_i(w) = 0\},\ N_1^- := |\mathcal{N}_1^-|,\\
\text{and } \mathcal{N}_1^+ &:= \{i \in [N] : s_i = 1 \wedge \hat{y}_i(w) = 1\},\ N_1^+ := |\mathcal{N}_1^+|.
\end{aligned}$$

Proof. Consider the shifted function $\varphi = 2(\phi - \frac{1}{2}) = 2\phi - 1$, which under the conditions of the theorem is symmetric about the origin, meaning $\varphi(t) = -\varphi(-t)$ for all $t \in \mathbb{R}$. One finds that

$$\begin{aligned}
2c_{\text{dp}}(w) &= \frac{\sum_{\{i\in[N]:s_i=1\}} 2\phi(t_i(w))}{N_1} - \frac{\sum_{\{i\in[N]:s_i=0\}} 2\phi(t_i(w))}{N_0} \\
&= \frac{\sum_{\{i\in[N]:s_i=1\}} (\varphi(t_i(w)) + 1)}{N_1} - \frac{\sum_{\{i\in[N]:s_i=0\}} (\varphi(t_i(w)) + 1)}{N_0} \\
&= \frac{\sum_{\{i\in[N]:s_i=1\}} \varphi(t_i(w))}{N_1} - \frac{\sum_{\{i\in[N]:s_i=0\}} \varphi(t_i(w))}{N_0} \\
&= \frac{\sum_{i\in\mathcal{N}_1^+} \varphi(|t_i(w)|) - \sum_{i\in\mathcal{N}_1^-} \varphi(|t_i(w)|)}{N_1^+ + N_1^-} - \frac{\sum_{i\in\mathcal{N}_0^+} \varphi(|t_i(w)|) - \sum_{i\in\mathcal{N}_0^-} \varphi(|t_i(w)|)}{N_0^+ + N_0^-}.
\end{aligned}$$

Now the condition that $|c_{\text{dp}}(w)| \le \epsilon$ implies that

$$\begin{aligned}
\Bigg| \frac{N_1^+ - N_1^-}{N_1^+ + N_1^-} - \frac{N_0^+ - N_0^-}{N_0^+ + N_0^-} &+ \frac{N_1^- - N_1^+ + \sum_{i\in\mathcal{N}_1^+} \varphi(|t_i(w)|) - \sum_{i\in\mathcal{N}_1^-} \varphi(|t_i(w)|)}{N_1^+ + N_1^-} \\
&+ \frac{N_0^+ - N_0^- - \sum_{i\in\mathcal{N}_0^+} \varphi(|t_i(w)|) + \sum_{i\in\mathcal{N}_0^-} \varphi(|t_i(w)|)}{N_0^+ + N_0^-} \Bigg| \le 2\epsilon.
\end{aligned}$$

This inequality has the form $|A + B| \le 2\epsilon$, where by the triangle inequality one has that $|A + B| \ge |A| - |B|$, which in turn means that $|A| \le |A + B| + |B| \le 2\epsilon + |B|$. Consequently, from above,

$$\begin{aligned}
&\left| \frac{N_1^+ - N_1^-}{N_1^+ + N_1^-} - \frac{N_0^+ - N_0^-}{N_0^+ + N_0^-} \right| \\
&\le \Bigg| \frac{N_1^- - N_1^+ + \sum_{i\in\mathcal{N}_1^+} \varphi(|t_i(w)|) - \sum_{i\in\mathcal{N}_1^-} \varphi(|t_i(w)|)}{N_1^+ + N_1^-} \\
&\quad + \frac{N_0^+ - N_0^- - \sum_{i\in\mathcal{N}_0^+} \varphi(|t_i(w)|) + \sum_{i\in\mathcal{N}_0^-} \varphi(|t_i(w)|)}{N_0^+ + N_0^-} \Bigg| + 2\epsilon \\
&= \Bigg| \frac{\sum_{i\in\mathcal{N}_1^+} (\varphi(|t_i(w)|) - 1) - \sum_{i\in\mathcal{N}_1^-} (\varphi(|t_i(w)|) - 1)}{N_1^+ + N_1^-} \\
&\quad + \frac{-\sum_{i\in\mathcal{N}_0^+} (\varphi(|t_i(w)|) - 1) + \sum_{i\in\mathcal{N}_0^-} (\varphi(|t_i(w)|) - 1)}{N_0^+ + N_0^-} \Bigg| + 2\epsilon.
\end{aligned}$$

Now defining

$$S_1^+ = \sum_{i\in\mathcal{N}_1^+} (\varphi(|t_i(w)|) - 1) \tag{8}$$

and similarly for S_1^-, S_0^+, and S_0^-, one finds from above that

$$\begin{aligned}
&\left| \frac{S_1^+ - S_1^-}{N_1^+ + N_1^-} - \frac{S_0^+ - S_0^-}{N_0^+ + N_0^-} \right| + 2\epsilon \\
&\ge \left| \frac{N_1^+ - N_1^-}{N_1^+ + N_1^-} - \frac{N_0^+ - N_0^-}{N_0^+ + N_0^-} \right| \\
&\ge \left| \frac{N_1^+}{N_1^+ + N_1^-} - \left(1 - \frac{N_1^+}{N_1^+ + N_1^-}\right) - \frac{N_0^+}{N_0^+ + N_0^-} + \left(1 - \frac{N_0^+}{N_0^+ + N_0^-}\right) \right| \\
&= 2 \left| \frac{N_1^+}{N_1^+ + N_1^-} - \frac{N_0^+}{N_0^+ + N_0^-} \right|,
\end{aligned}$$

which in turn implies that

$$|\bar{c}_{\text{dp}}(w)| = \left| \frac{N_1^+}{N_1^+ + N_1^-} - \frac{N_0^+}{N_0^+ + N_0^-} \right| \le \frac{1}{2} \left(\left| \frac{S_1^+ - S_1^-}{N_1^+ + N_1^-} - \frac{S_0^+ - S_0^-}{N_0^+ + N_0^-} \right| + 2\epsilon \right).$$

Under the conditions of the theorem, one has for all $i \in [N]$ that $\phi(t_i(w)) \in [0, \gamma] \cup [1-\gamma, 1]$, which in turn means that $\varphi(|t_i(w)|) \in [1-\gamma, 1] \subseteq [1-\gamma, 1+\gamma]$, so $S_1^+ \in [-\gamma N_1^+, \gamma N_1^+]$ and similarly for S_1^-, S_0^+, and S_0^-. Thus, $\frac{S_1^+ - S_1^-}{N_1^+ + N_1^-} \in [-\gamma, \gamma]$ and $\frac{S_0^+ - S_0^-}{N_0^+ + N_0^-} \in [-\gamma, \gamma]$, which in turn shows

$$\left| \frac{S_1^+ - S_1^-}{N_1^+ + N_1^-} - \frac{S_0^+ - S_0^-}{N_0^+ + N_0^-} \right| \in [0, 2\gamma].$$

With the prior conclusion, one obtains that $|\bar{c}_{\text{dp}}(w)| \leq \gamma + \epsilon$, as desired. □

B Formal Statement and Proof of Theorem 3

Let the features be defined by a pair of random variables (X, S), which are in turn defined by a probability measure $(\Omega, \mathcal{F}, \mathbb{P})$. Here, S represents the sensitive feature, and $S(\omega) \in \{0, 1\}$ for all $\omega \in \Omega$. Subsequently, let $\{(X_i, S_i)\}_{i=1}^N$ be a set of N random-variable pairs, where for all $i \in [N]$ the pair (X_i, S_i) has the same distribution as (X, S). The set of possible outcomes of $\{(X_i, S_i)\}_{i=1}^N$ is $\Omega \times \cdots \times \Omega = \prod_{i=1}^N \Omega$, and the corresponding σ-algebra and probability measure for $\{(X_i, S_i)\}_{i=1}^N$, call it $\mathbb{P}_N$, can be derived from $(\mathcal{F}, \mathbb{P})$.

In general, a realization of $\{(X_i, S_i)\}_{i=1}^N$, call it $\{(x_i, s_i)\}_{i=1}^N$, can have $0 \leq \sum_{i=1}^n s_i \leq N$. (For example, if S indicates female or male, then a random sample of N data points could have any numbers of females or males.) It would be problematic to work with all such realizations when aiming to ensure an unbiased estimate of an unfairness measure; e.g., if a sample contains no males or no females, then the unfairness-measure estimate is not well defined. Instead, we want to work with the conditional distribution of $\{(X_i, S_i)\}_{i=1}^N$ subject to the condition that any realization has $N_0 \in \{1, \ldots, N\}$ values with $s_i = 0$ and $N_1 = N - N_0$ values with $s_i = 1$. Let this conditional distribution have measure $\mathbb{P}_{N_0,N_1}$. Let $\mathbb{E}_{N_0,N_1}$ denote expectation taken with respect to $\mathbb{P}_{N_0,N_1}$.

At the same time, we can define the conditional distribution of X subject to $S = 0$ and the conditional distribution of X subject to $S = 1$. Let the corresponding probability measures be $\mathbb{P}_0$ and $\mathbb{P}_1$, respectively. Let $\mathbb{E}_0$ and $\mathbb{E}_1$ denote expectation taken with respect to $\mathbb{P}_0$ and $\mathbb{P}_1$, respectively.

Given w, the values in which we are interested are the real number

$$c(w) = \mathbb{E}_0[\phi(t(X, 0, w))] - \mathbb{E}_1[\phi(t(X, 1, w))]$$

and the random variable

$$C(w) = \frac{1}{N_0} \sum_{\{i:S_i=0\}} \phi(t(X_i, 0, w)) - \frac{1}{N_1} \sum_{\{i:S_i=1\}} \phi(t(X_i, 1, w)).$$

Theorem 3. (Formal) *For any w, the random variable $C(w)$ is an unbiased estimator of $c(w)$. That is, $\mathbb{E}_{N_0,N_1}[C(w)] = c(w)$ for all w.*

Proof. We start by taking the expectation of $C(w)$

$$\mathbb{E}_{N_0,N_1}[C(w)] = \mathbb{E}_{N_0,N_1}\left[\frac{1}{N_0} \sum_{\{i:S_i=0\}} \phi(t(X_i, 0, w)) - \frac{1}{N_1} \sum_{\{i:S_i=1\}} \phi(t(X_i, 1, w))\right]$$

Using the linearity of expectation

$$\mathbb{E}_{N_0,N_1}[C(w)] = \mathbb{E}_{N_0,N_1}\left[\frac{1}{N_0} \sum_{\{i:S_i=0\}} \phi(t(X_i, 0, w))\right] - \mathbb{E}_{N_0,N_1}\left[\frac{1}{N_1} \sum_{\{i:S_i=1\}} \phi(t(X_i, 1, w))\right]$$

Using the property that the expectation of a sum is the sum of expectations

$$\mathbb{E}_{N_0,N_1}[C(w)] = \frac{1}{N_0} \sum_{\{i:S_i=0\}} \mathbb{E}_0\left[\phi(t(X, 0, w))\right] - \frac{1}{N_1} \sum_{\{i:S_i=1\}} \mathbb{E}_1\left[\phi(t(X, 1, w))\right]$$

Since the samples are i.i.d., the expectation of each term in the sums is the same

$$\mathbb{E}_{N_0,N_1}[C(w)] = \mathbb{E}_0\left[\phi(t(X, 0, w))\right] - \mathbb{E}_1\left[\phi(t(X, 1, w))\right]$$

By comparing the above result with the definition of $c(w)$, we see that

$$\mathbb{E}_{N_0,N_1}[C(w)] = c(w),$$

as claimed. □

C Overview of Datasets

An overview of the datasets employed in our numerical experiments is provided in Table 2. Along with basic information including the number of data points N, numbers of data points with sensitive feature value equal to 1 or 0, etc., the table includes some statistics that are relevant for measuring how innately fair or unfair is the dataset. These are motivated in the following paragraphs. In each of the descriptions below, the measure $\mathbb{P}$ corresponds to the empirical distribution of the dataset.

Demographic Parity Violation with Rote Learning. If a prediction model were to predict the true labels with 100% accuracy, i.e., $\hat{Y} = Y$ over the entire dataset, which corresponds to rote learning of the dataset, then the violation of demographic parity is given by

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

This measure quantifies the absolute difference in the likelihood of a positive outcome $Y = 1$ between one of the groups ($S = 1$) and the other group ($S = 0$) based on the sensitive feature.

Equal Opportunity Violation with Rote Learning. If a prediction model were to predict the true labels with 100% accuracy, then the violation of equal opportunity is given by

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

This measure assesses the disparity in true positive rates between the protected and unprotected groups, focusing on individuals who actually received a positive outcome ($Y = 1$).

Disparate Impact with Rote Learning If a prediction model were to predict the true labels with 100% accuracy, then the disparate impact—namely, the largest value of δ such that both of the inequalities in Equation (7) of the main paper (Section 3.2) would be satisfied—is given by

$$\min\left\{\frac{\mathbb{P}(Y=1|S=0)}{\mathbb{P}(Y=1|S=1)}, \frac{\mathbb{P}(Y=1|S=1)}{\mathbb{P}(Y=1|S=0)}\right\}.$$

This measures the relative disparity in the likelihood of a positive outcome between the protected and unprotected groups, providing insight into potential biases in the model's predictions. For the datasets shown in Table 2, this measure provides insight into why our constraint on disparate impact was not active until $\delta \approx 0.78$ for the Law dataset and was active with $\delta \approx 0.36$ for the Adult dataset.

All-Zero Prediction and All-One Prediction Accuracies A potential concern in the pursuit of a fair prediction model is that model training would result in consistent predictions for all individuals regardless of their feature data. This corresponds to no learning, and simply to predicting the same outcome for all individuals. To determine whether such predictions could be desirable from the viewpoint of achieving near-optimal accuracy, one can compute

$$\frac{\mathbb{P}(Y=0)}{\mathbb{P}(Y=0)+\mathbb{P}(Y=1)} \text{ and } \frac{\mathbb{P}(Y=1)}{\mathbb{P}(Y=0)+\mathbb{P}(Y=1)}$$

For the datasets shown in Table 2, one finds that the optimal accuracy in terms of a learned prediction model should be at least 89.1% for the Law dataset and at least 75.6% for the Adult dataset.

Table 2: Statistics for the Law and Adult datasets.

Metric	Law	Adult
N	16,638	26,048
# $S = 1$	13,995	17,436
# $S = 0$	2,643	8,612
# $Y = 1$	14,826	6,354
# $Y = 0$	1,812	19,694
# $Y = 1 \wedge S = 1$	12,915	5,391
# $Y = 1 \wedge S = 0$	1,911	963
# $Y = 0 \wedge S = 1$	1,080	12,045
# $Y = 0 \wedge S = 0$	732	7,649
$\mathbb{P}(Y = 1\|S = 1)$	0.923	0.3092
$\mathbb{P}(Y = 1\|S = 0)$	0.723	0.1118
$\mathbb{P}(S = 1\|Y = 1)$	0.871	0.8487
$\mathbb{P}(S = 0\|Y = 1)$	0.129	0.1513
Demographic Parity Violation with Rote Learning	0.200	0.1974
Equal Opportunity Violation with Rote Learning	0.743	0.6974
Disparate Impact with Rote Learning	0.784	0.3615
All-Zero Prediction Accuracy	0.109	0.7560
All-One Prediction Accuracy	0.891	0.2440

D SQP Algorithm and Learning Rate Adjustment Strategy

For the purposes of describing the SQP algorithm (based on that in [13]) that we employ for model training for our experiments in the paper, let us write the optimization problem of interest in the following form, where we consider the case of having two constraint functions:

$$\min_{w} \sum_{i\in[N]} \ell(\mathcal{N}(x_i, s_i, w), y_i) + \tfrac{1}{\lambda} r(w) \;\text{ subject to }\; \begin{bmatrix} l_1 \\ l_2 \end{bmatrix} \leq \begin{bmatrix} c_1(w) \\ c_2(w) \end{bmatrix} \leq \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}. \tag{9}$$

The algorithm that we state is readily extended to the case of more than 2 constraint functions.

Each iteration of the SQP algorithm involves solving a quadratic optimization (QP) subproblem to compute a step. In each iteration $k \in \mathbb{N}$, in which the current iterate is w_k, this QP has the form

$$\begin{aligned} &\min_{d}\; g(w_k)^T d + \tfrac{1}{2} d^T H_k d \\ &\text{s.t. } J(w_k)d := \begin{bmatrix} -\nabla c_1(w_k)^T \\ \nabla c_1(w_k)^T \\ -\nabla c_2(w_k)^T \\ \nabla c_2(w_k)^T \end{bmatrix} d \leq - \begin{bmatrix} l_1 - c_1(w_k) \\ c_1(w_k) - u_1 \\ l_2 - c_2(w_k) \\ c_2(w_k) - u_2 \end{bmatrix} =: -r(w_k). \end{aligned} \tag{10}$$

where $g(w_k)$ is a gradient or stochastic gradient of the objective function of (9) at w_k and H_k is a positive definite matrix. Since H_k is positive definite and the feasible region of this QP is convex, it follows that the QP has a unique optimal solution. Moreover, as is well known, if it is known which of the inequality constraints are satisfied at equality at the optimal solution of the QP, then the solution can be obtained by solving a symmetric indefinite linear system of the form

$$\begin{bmatrix} H_k & J_{\mathcal{A}}(w_k) \\ J_{\mathcal{A}}(w_k)^T & 0 \end{bmatrix} \begin{bmatrix} d \\ y \end{bmatrix} = - \begin{bmatrix} g \\ r_{\mathcal{A}}(w_k) \end{bmatrix}. \tag{11}$$

Here, $J_{\mathcal{A}}$ and $r_{\mathcal{A}}$ are the rows of J and r (see (10)) corresponding to the indices in $\mathcal{A} \subseteq \{1, 2, 3, 4\}$.

Our SQP algorithm, stated as Algorithm 1 below, specifies a procedure for solving (10) in each iteration by considering, when necessary, different choices for the optimal active set $\mathcal{A}$. We note that for computational efficiency any matrix of the form in (11) only needs to be factored once in order to solve any number of linear systems of the form (11) in the iteration.

Algorithm 1 Stochastic Sequential Quadratic Programming (SQP)

```
Require: initial iterate $w_0$, prescribed learning-rate sequence $\{\beta_k\}$, initial merit parameter $\tau_{-1}$
for $k = 0, 1, 2, ...$ do
    Choose $H_k$ (our implementation uses an Adagrad-type scheme)
    Compute $d_{mm}$ as the unconstrained minimizer of (10), i.e., solve (11) with $\mathcal{A} = \emptyset$
    if $d_{mm}$ is feasible for (10) then
        Set $d_k \leftarrow (d_{mm})$
    else
        Compute $d_{lm}$ as the minimizer of (10) for $\mathcal{A} = \{1\}$
        Compute $d_{um}$ as the minimizer of (10) for $\mathcal{A} = \{2\}$
        Compute $d_{ml}$ as the minimizer of (10) for $\mathcal{A} = \{3\}$
        Compute $d_{mu}$ as the minimizer of (10) for $\mathcal{A} = \{4\}$
        if any of these candidates is feasible for (10) then
            Set $d_k = \operatorname*{argmin}_{d\in\{d_{lm},d_{um},d_{ml},d_{mu}\}} g(w_k)^T d + \frac{1}{2} d^T H_k d$
        else
            Compute $d_{ll}$ as the minimizer of (10) for $\mathcal{A} = \{1, 3\}$
            Compute $d_{lu}$ as the minimizer of (10) for $\mathcal{A} = \{1, 4\}$
            Compute $d_{ul}$ as the minimizer of (10) for $\mathcal{A} = \{2, 3\}$
            Compute $d_{uu}$ as the minimizer of (10) for $\mathcal{A} = \{2, 4\}$
            Set $d_k = \operatorname*{argmin}_{d\in\{d_{ll},d_{lu},d_{lu},d_{uu}\}} g(w_k)^T d + \frac{1}{2} d^T H_k d$
        end if
    end if
    Set $\tau_k$ based on $d_k$; see [13]
    Set $\alpha_k$ based on $\beta_k$ and $\tau_k$; see [13]
    Set $w_{k+1} \leftarrow w_k + \alpha_k d_k$
    Apply Learning Rate Adjustment Strategy (Algorithm 2)
    $k \leftarrow k + 1$
end for
```

Algorithm 2 Learning Rate Adjustment Strategy (based on [44])

```
Require:
$\tau_k$                                                ▷ merit parameter
$\alpha_{\text{min}} \leftarrow 10^{-7}$                ▷ minimum learning rate
$k_{\text{min}} \leftarrow 200$                         ▷ minimum iterations
$\Delta k \leftarrow 5$                                 ▷ adjustment interval
$\gamma \leftarrow 10$                                  ▷ reduction factor
$\eta \leftarrow 0.85$                                  ▷ initial weight for convex combination
$\delta \leftarrow 0$                                   ▷ iterations since last adjustment
$n_c$                                                   ▷ number of constraints
$f(w_k)$                                                ▷ objective value at iteration $k$ (see (9))
$r(w_k)$                                                ▷ constraint value at iteration $k$ (see (10))
$\phi_k \leftarrow \tau f(w_k) + \|r(w_k)\|_1$          ▷ compute merit value
if $k = 0$ then
    Set $\bar{\phi}_k \leftarrow \phi_k$                ▷ initialize moving average
else
    Set $\bar{\phi}_k \leftarrow \eta\phi_k + (1-\eta)\bar{\phi}_{k-1}$    ▷ update moving average
end if
if $k \geq k_{\text{min}}$ and $\alpha_k \geq \alpha_{\text{min}}$ then
    if $\bar{\phi}_k \leq \phi_k$ then
        if $\delta_k \geq \Delta k$ then
            $\alpha_{k+1} \leftarrow \alpha_k / \gamma$
            $\delta_k \leftarrow 0$
        end if
    end if
    $\delta_k \leftarrow \delta_k + 1$
end if
```

E Additional Experimental Results

This section provides supplementary experimental results that complement the findings presented in the experiments section. We first report the compute times for training of each model and dataset at Table 3. Following this, we present additional plots.

Table 3: Compute times for training with the Law and Adult datasets. ”Avg time” is minutes for one 500-epoch run on the laptop CPU (Each number is averaged over all $(\lambda$, constraint threshold$(\delta$ or $\epsilon)$ runs of the given model); ”Max peak” is the largest memory use seen in any run; ”Total CPU” is the overall core-time spent on each dataset.

Dataset	**Model Type**	**Avg Time (min)**	**Max Peak (MB)**	**Total Runs**	**Total CPU (h)**
Law	smoothed-step	11.48	48.50	16	3.06
	sigmoid	10.92	7.10	16	2.91
	covariance	10.85	7.10	16	2.89
Adult	smoothed-step	380.40	52.30	16	101.44
	sigmoid	216.90	10.90	16	57.84
	covariance	149.19	10.90	16	39.78

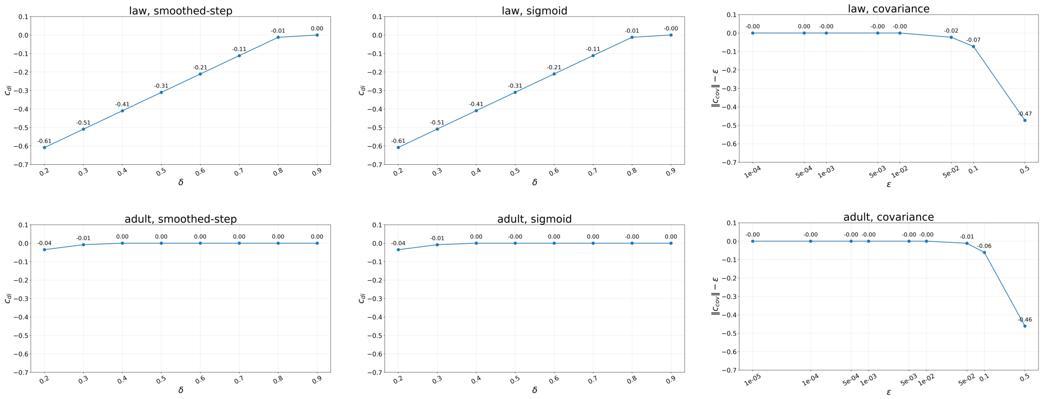


Figure 6: Constraint violation of the model across varying fairness thresholds for smoothed-step, sigmoid, and covariance models on the Law and Adult datasets. These plots show how constraint violations of the model behave as thresholds tighten, with each method activating constraints at different values.

Table 4: Accuracies (overall, male, and female) and constraint violations when disparate impact (DI) and/or equal impact (EI) constraints are enforced with $\delta = 0.9$ for the Law and Adult datasets. Enforcing both constraints reduces overall accuracy, particularly for Adult, but the constraint-only method consistently satisfies both unfairness constraints, as indicated by non-positive constraint model violations (c_{di} and c_{ei}) and non-positive constraint violation measures ($\bar{c}_{di}$ and $\bar{c}_{ei}$).

Dataset	**Scenario**	Overall Acc	Male Acc	Female Acc	c_{di}	c_{ei}	$\bar{c}_{di}$	$\bar{c}_{ei}$
Law ($\delta = 0.9$)	Only DI	89.88	92.38	76.62	0.00	0.07	-0.00	-0.06
	Only EI	89.53	92.40	74.31	-0.07	-0.00	-0.07	-0.09
	Both DI-EI	89.88	92.38	76.62	0.00	-0.06	-0.00	-0.06
Adult ($\delta = 0.9$)	Only DI	80.79	75.50	91.48	0.00	0.18	-0.00	0.18
	Only EI	83.96	80.09	91.78	0.13	-0.01	0.13	-0.01
	Both DI-EI	75.68	69.20	88.79	0.00	-0.00	-0.00	0.00

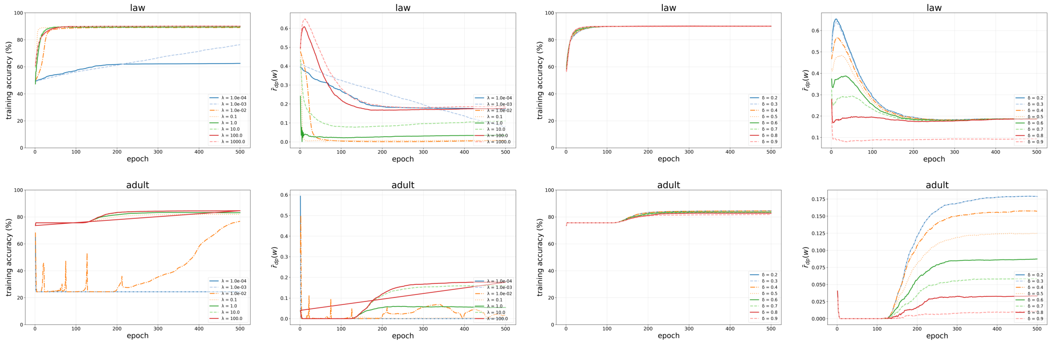


Figure 7: Comparison between constraint-based (right) and regularization-based (left) approaches for enforcing fairness using the sigmoid function on the Law and Adult datasets. This figure compares the effectiveness of constraint-based and regularization-based methods in achieving fairness. The constraint-based approach consistently meets the targeted fairness thresholds (δ or ϵ) with minimal impact on accuracy, while the regularization-based method exhibits unpredictable fairness outcomes and greater accuracy loss, underscoring the advantages of hard constraint enforcement.

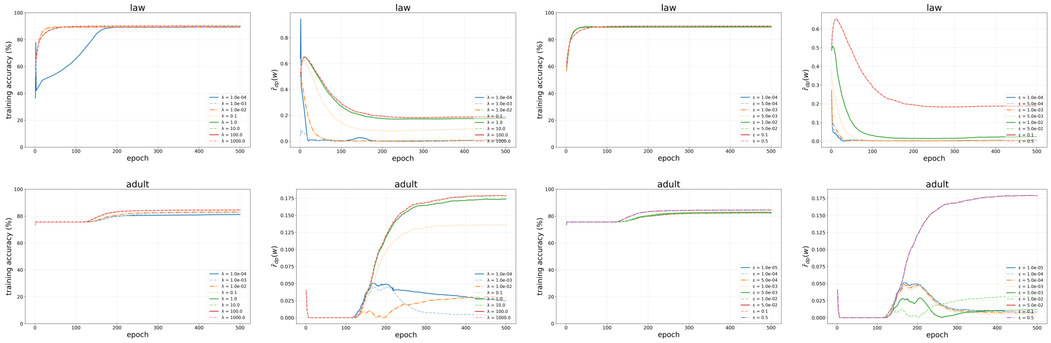


Figure 8: Comparison between constraint-based (right) and regularization-based (left) approaches for enforcing fairness using the covariance function on the Law and Adult datasets. This figure compares the effectiveness of constraint-based and regularization-based methods in achieving fairness. The constraint-based approach consistently meets the targeted fairness thresholds (δ or ϵ) with minimal impact on accuracy, while the regularization-based method exhibits unpredictable fairness outcomes and greater accuracy loss, underscoring the advantages of hard constraint enforcement.

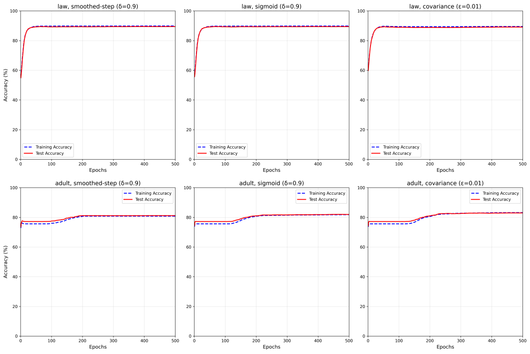


Figure 9: Training and test accuracies for different model formulations: smoothed-step ($\delta = 0.9$), sigmoid ($\delta = 0.9$), and covariance ($\epsilon = 0.01$). All formulations show similar convergence with minimal generalization gap, achieving comparable training and testing accuracy.

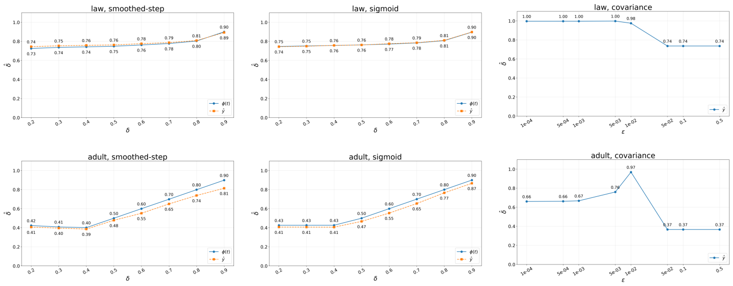


Figure 10: Levels of disparate impact actually achieved ($\hat{\delta}$) in the minibatch training setting when prediction models are trained with constraints as in Equation (7) of the main paper (Section 3.2) for varying values of δ *with* scaling of the surrogate functions. The plots on the right are the levels of disparate impact actually achieved when a constraint on a covariance surrogate (less than or equal to ϵ) is imposed.

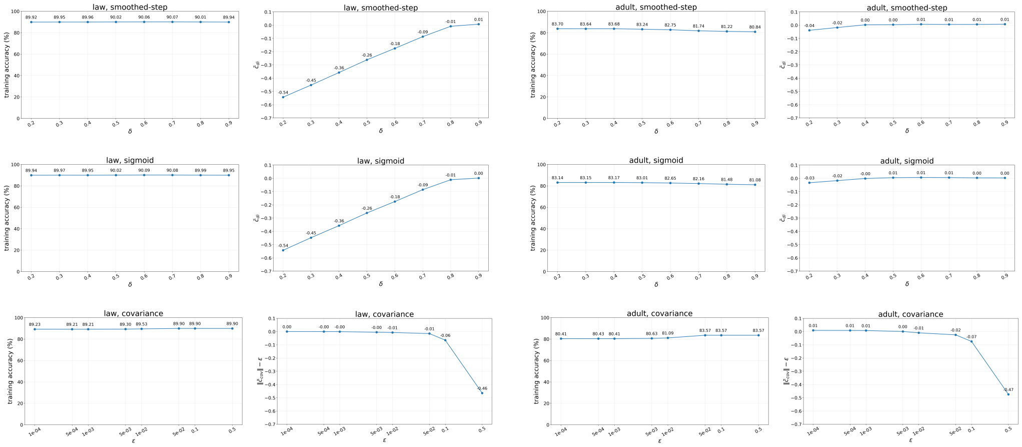


Figure 11: Training accuracy and constraint violation measures in the minibatch training setting for smoothed-step, sigmoid, and covariance models on the Law and Adult datasets. The plots show training accuracy alongside constraint violation metrics for different surrogate functions. The results indicate that fairness constraints can be satisfied without compromising predictoin accuracy.

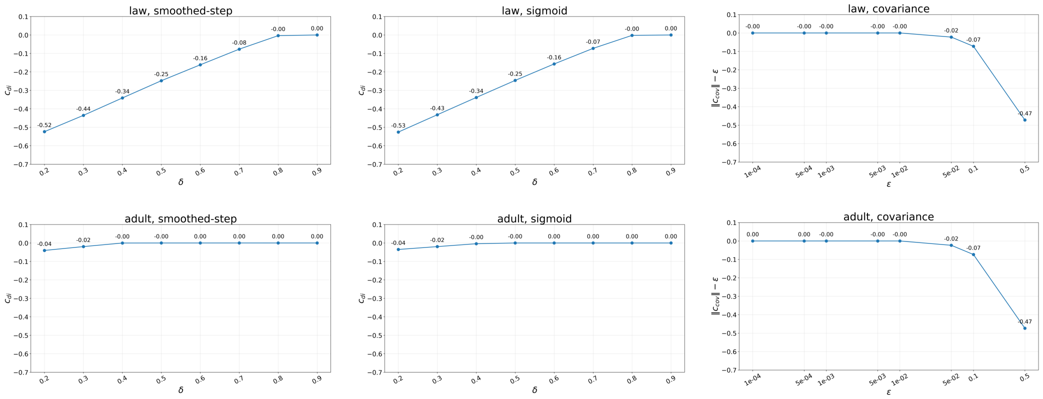


Figure 12: Constraint violation of the model across varying fairness thresholds in the minibatch training setting for smoothed-step, sigmoid, and covariance models on the Law and Adult datasets. These plots show how constraint violations of the model behave as thresholds tighten, with each method activating constraints at different values.