Too Relaxed to Be Fair

Michael Lohaus [1 2] **Michaël Perrot** [2 3 4] **Ulrike von Luxburg** [1 2]

Abstract

We address the problem of classification under fairness constraints. Given a notion of fairness, the goal is to learn a classifier that is not discriminatory against a group of individuals. In the literature, this problem is often formulated as a constrained optimization problem and solved using relaxations of the fairness constraints. We show that many existing relaxations are unsatisfactory: even if a model satisfies the relaxed constraint, it can be surprisingly unfair. We propose a principled framework to solve this problem. This new approach uses a strongly convex formulation and comes with theoretical guarantees on the fairness of its solution. In practice, we show that this method gives promising results on real data.

1. Introduction

Informally, a classifier is considered *unfair* when it unjustly promotes a group of individuals while being detrimental to others; it is considered *fair* when it is free of any unjust behavior. However, the details of what is fair and unfair can be vastly different from one application to another. For example, a college might want to admit a diverse student pool with respect to gender or race. This notion of fairness is called *demographic parity*. On the other hand, consider a bank giving out loans. If a group of individuals repays less frequently than others, it is normal that they receive fewer loans. However, it does not mean that all requests should be declined. In particular, any individual that is likely to repay a loan should be given the opportunity to get one, regardless of group membership. This is called *equality of opportunity*.

The problem of learning fair classifiers has mainly been addressed in three ways. First, pre-processing approaches alter the labels of the examples or their representation to increase the intrinsic fairness of a dataset. A classifier learned on this modified data is then more likely to be fair (Feldman et al., 2015; Calmon et al., 2017; Kamiran & Calders, 2012; Dwork et al., 2012; Zemel et al., 2013). Second, post-hoc procedures transform existing accurate but unfair classifiers into fair classifiers (Chzhen et al., 2019; Hardt et al., 2016; Woodworth et al., 2017; Kamiran et al., 2010). Finally, direct methods learn a fair and accurate classifier in a single step (Kamishima et al., 2012; Zafar et al., 2017b;a; Calders & Verwer, 2010; Wu et al., 2019; Donini et al., 2018; Cotter et al., 2019; Agarwal et al., 2018; Goh et al., 2016). In this paper, we focus on the latter kind of approaches.

Motivation: relaxations sometimes fail to produce fair solutions. Recently, several direct methods have been proposed that use relaxed versions of the considered fairness constraint. These approaches work reasonably well for some applications. However, their relaxations are quite coarse and we demonstrate below that they can fail to find fair classifiers. In particular, there is typically no guarantee on the relationship between the relaxed fairness and the true fairness of a solution: a classifier that is perfectly fair in terms of relaxed fairness can be highly unfair in terms of true fairness (see Figure 1 for an illustration). In this paper, we study the limitations of a number of popular approaches (Zafar et al., 2017b;a; Wu et al., 2019; Donini et al., 2018).

Algorithmic contributions. We propose a new principled framework to tackle the problem of fair classification that is particularly relevant for application scenarios where formal fairness guarantees are required. Our approach is based on convex relaxations and comes with theoretical guarantees that ensure that the learned classifier is fair up to sampling errors. Furthermore, we use a learning theory framework for similarity-based classifiers to exhibit sufficient conditions that guarantee the existence of a fair and accurate classifier.

2. Problem Setting

Let $\mathcal{X}$ be a feature space, $\mathcal{Y} = \{-1, 1\}$ a space of binary class labels, and $\mathcal{S} = \{-1, 1\}$ a space of binary sensitive attributes. Assume that there exists a distribution $\mathcal{D}_\mathcal{Z}$ over $\mathcal{Z} = \mathcal{X} \times \mathcal{S} \times \mathcal{Y}$ and that we can draw some examples

[1] University of Tübingen, Germany [2] Max Planck Institute for Intelligent Systems, Tübingen, Germany [3] Univ Lyon, UJM-Saint-Etienne, CNRS, IOGS, LabHC UMR 5516, F-42023, SAINT-ETIENNE, France [4] Most of the work was done when M.P. was affiliated with the Max Planck Institute. Correspondence to: Michael Lohaus <michael.lohaus@uni-tuebingen.de>, Michaël Perrot <michael.perrot@univ-st-etienne.fr>, Ulrike von Luxburg <luxburg@informatik.uni-tuebingen.de>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

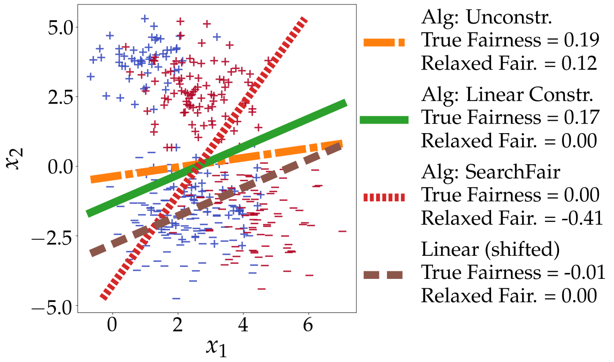


Figure 1. The goal is to separate the positive class (+) from the negative class (−) while remaining fair with respect to two sensitive groups: the blue and the red group. We evaluate the true fairness (DDP) and a linear relaxation of the fairness (Zafar, Section 3.1) of three linear classifiers learned with no fairness constraint (Unconstr., orange), a linear relaxation of the fairness constraint (Linear Constr., green), and our framework (SearchFair, red). We also plot the classifier obtained by translating Linear (Linear (shifted), brown). It has the same relaxed fairness as Linear but a different true fairness: the relaxation is oblivious to the intercept parameter. SearchFair finds the fairest classifier.

$(x, s, y) \sim \mathcal{D}_\mathcal{Z}$. Our goal in fair classification is to obtain a classifier, a mapping $h : \mathcal{X} \to \mathcal{Y}$ defined as $h(x) = \operatorname{sign}(f(x))$ where $f : \mathcal{X} \to \mathbb{R}$ is a real valued function, that is fair with respect to the sensitive attribute while remaining accurate on the class labels. In this paper, we study two notions of fairness: *demographic parity* and *equality of opportunity*.

Demographic Parity. A classifier f is fair for demographic parity when its predictions are independent of the sensitive attribute (Calders et al., 2009; Calders & Verwer, 2010). Formally, this can be written as

$$\underset{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}{\mathbb{P}}[f(x)>0|s=1] = \underset{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}{\mathbb{P}}[f(x)>0|s=-1]\,.$$

In practice, enforcing exact demographic parity might be too restrictive. Instead, we consider a fairness score (Wu et al., 2019) called Difference of Demographic Parity (DDP):

$$\mathrm{DDP}(f) = \underset{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}{\mathbb{E}}\left[\mathbb{I}_{f(x)>0}|s=1\right] - \underset{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}{\mathbb{E}}\left[\mathbb{I}_{f(x)>0}|s=-1\right], \tag{1}$$

where $\mathbb{I}_a$ is the indicator function that returns 1 when a is true and 0 otherwise. The DDP is positive when the favoured group is $s = 1$ and negative when it is $s = -1$. Given a threshold $\tau \geq 0$, we say that a classifier f is τ-DDP fair if $|\mathrm{DDP}(f)| \leq \tau$. When $\tau = 0$, exact demographic parity is achieved and we say that the classifier is DDP fair.

Equality of Opportunity. A classifier f is fair for equality of opportunity when its predictions for positively labelled examples are independent of the sensitive attribute (Hardt et al., 2016). Formally, it is

$$\underset{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}{\mathbb{P}}[f(x)>0|y=1,s=1] = \underset{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}{\mathbb{P}}[f(x)>0|y=1,s=-1]\,.$$

Again, instead of only considering exact equality of opportunity, we use a fairness score (Donini et al., 2018) called Difference of Equality of Opportunity (DEO):

$$\mathrm{DEO}(f) = \underset{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}{\mathbb{E}}\left[\mathbb{I}_{f(x)>0}|y=1,s=1\right] - \underset{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}{\mathbb{E}}\left[\mathbb{I}_{f(x)>0}|y=1,s=-1\right]. \tag{2}$$

This quantity is positive when the favoured group is $s=1$ and negative when it is $s=-1$. Given a threshold $\tau \geq 0$, we say that a classifier f is τ-DEO fair if $|\mathrm{DEO}(f)| \leq \tau$. When $\tau = 0$, exact equality of opportunity is achieved and we say that the classifier is DEO fair.

It is worth noting that demographic parity and equality of opportunity are quite similar from a mathematical point of view. In the remainder of the paper, we focus our exposition on DDP as results that hold for DDP can often be readily extended to DEO by conditioning on the target label. We only provide details in the supplementary when these extensions are more involved.

Learning a fair classifier. Given a function class $\mathcal{F}$, a τ-DDP fair and accurate classifier f^* is given as the solution of the following problem:

$$f^* = \underset{\substack{f\in\mathcal{F}\\|\mathrm{DDP}(f)|\leq\tau}}{\arg\min}\ L(f)\,,$$

where $L(f) = \mathbb{E}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}\left[\ell(f(x), y)\right]$ is the true risk of f for a convex loss function $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. In practice, we only have access to a set $\widehat{\mathcal{D}}_\mathcal{Z} = \{(x_i, s_i, y_i)\}_{i=1}^n$ of n examples drawn from $\mathcal{D}_\mathcal{Z}$. Hence, we consider the empirical version of this problem:

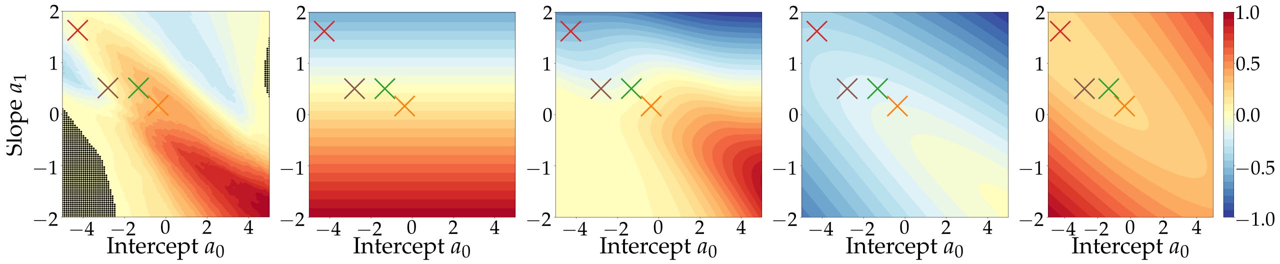
$$f^\beta = \underset{\substack{f\in\mathcal{F}\\|\mathrm{DDP}(f)|\leq\tau}}{\arg\min}\ \widehat{L}(f) + \beta\Omega(f)\,, \tag{3}$$

where $\Omega(f)$ is a convex regularization term used to prevent over-fitting, β is a trade-off parameter, and $\widehat{L}(f) = \frac{1}{n}\sum_{(x,s,y)\in\widehat{\mathcal{D}}_\mathcal{Z}} \ell(f(x), y)$ is the empirical risk. The main difficulty involved in learning a fair classifier is to ensure that $|\mathrm{DDP}(f)| \leq \tau$.

3. When Fairness Relaxations Fail

To obtain a τ-DDP fair classifier, most approaches consider the fully empirical version of Optimization Problem 3:

$$\min_{f\in\mathcal{F}} \quad \hat{L}(f) + \beta\Omega(f)$$
$$\text{subject to } |\widehat{\mathrm{DDP}}(f)| \leq \tau, \tag{4}$$



(a) DDP. (b) Linear. (c) Convex-concave. (d) Wu - Lower. (e) Wu - Upper.

Figure 2. Consider linear classifiers for the dataset in Figure 1. The decision boundaries are of the form $x_2 = a_1x_1 + a_0$ where a_1 controls the slope and a_0 the intercept. For given intercepts and slopes, we plot normalized values of (a) the DDP score (yellow is fair), (b) the linear relaxation (Section 3.1), (c) the convex-concave relaxation (Section 3.2), (d) the concave Wu lower bound, and (d) the convex Wu upper bound (Section 3.2). The black dotted area in (a) corresponds to trivial constant classifiers—the predicted class is the same for all points. The colored crosses correspond to the classifiers in Figure 1. A good relaxation should capture the true DDP reasonably well, in particular the yellow regions should match. However, none of the considered relaxations manage to achieve this.

where the empirical version of DDP is:

$$\widehat{\mathrm{DDP}}(f) = \frac{1}{n}\sum_{\substack{(x,s,y)\in\widehat{\mathcal{D}}_{\mathcal{Z}}\\ s=1}} \mathbb{I}_{f(x)>0} - \frac{1}{n}\sum_{\substack{(x,s,y)\in\widehat{\mathcal{D}}_{\mathcal{Z}}\\ s=-1}} \mathbb{I}_{f(x)>0}.$$

The main issue with this optimization problem is the non-convexity of the constraints that makes it hard to find the optimal solution. A standard approach is then to first rewrite the DDP in an equivalent, but easier to handle form[1] and then replace the indicator functions with a relaxation. Zafar et al. (2017b) and Donini et al. (2018) use a linear relaxation to obtain a fully convex constraint. Zafar et al. (2017a) use a convex relaxation that leads to a convex-concave constraint. Wu et al. (2019) combine a convex relaxation with a concave one to obtain a fully convex problem. Below, we show that these approaches only loosely approximate the true constraint and might lead to suboptimal solutions (see Figure 2). Furthermore, when theoretical guarantees accompany the method, they are either insufficient to ensure that the learned classifier is fair (Wu et al., 2019) or they make assumptions that are hard to satisfy in practice (Donini et al., 2018).

3.1. Linear Relaxations

We first study methods that use a linear relaxation of the indicator function to obtain a convex constraint in Optimization Problem 4. First, Zafar et al. (2017b) rewrite the DDP:

$$\mathrm{DDP}(f) = \mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}} \frac{1}{p_1(1-p_1)}\left(\frac{s+1}{2} - p_1\right)\mathbb{I}_{f(x)>0},$$

where $p_1 = \mathbb{P}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}(s=1)$ is the proportion of individuals in group $s=1$. Then, they consider a linear approximation of $\mathbb{I}_{f(x)>0}$ and obtain the constraint:

$$\left|\frac{1}{n}\sum_{(x,s,y)\in\widehat{\mathcal{D}}_{\mathcal{Z}}} \frac{1}{\hat{p}_1(1-\hat{p}_1)}\left(\frac{s+1}{2} - \hat{p}_1\right) f(x)\right| \leq \tau,$$

where $\hat{p}_1$ is an empirical estimate of p_1. In their original formulation, Zafar et al. (2017b) get rid of the factor $\frac{1}{\hat{p}_1(1-\hat{p}_1)}$ by replacing the right hand side of the constraint with $c = \hat{p}_1(1-\hat{p}_1)\tau$.

Similarly, Donini et al. (2018) rewrite the DDP:

$$\mathrm{DDP}(f) = \mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}} \frac{s}{p_s}\mathbb{I}_{f(x)>0},$$

where $p_s = \mathbb{P}_{(x',s',y')\sim\mathcal{D}_{\mathcal{Z}}}(s'=s)$ is the proportion of individuals in group s. Then, using the same linear relaxation as Zafar et al. (2017b) with $\hat{p}_s$, an empirical estimate of p_s, they obtain the constraint[2]

$$\left|\frac{1}{n}\sum_{(x,s,y)\in\widehat{\mathcal{D}}_{\mathcal{Z}}} \frac{s}{\hat{p}_s} f(x)\right| \leq \tau.$$

Both constraints are mathematically close and only differ in terms of the multiplicative factor in front of $f(x)$ in the inner sum. Thus, they can be rewritten as

$$\left|\mathrm{LR}_{\widehat{\mathrm{DDP}}}(f)\right| = \left|\frac{1}{n}\sum_{(x,s,y)\in\widehat{\mathcal{D}}_{\mathcal{Z}}} C\left(s,\widehat{\mathcal{D}}_{\mathcal{Z}}\right) f(x)\right| \leq \tau.$$

[1] In the supplementary, we provide the derivations for all the alternate formulations of DDP presented in this paper.

[2] Donini et al. (2018) originally consider τ-DEO fairness rather than DDP. In the constraint, instead of drawing the examples from $\mathcal{D}_{\mathcal{Z}}$, they use the conditional distribution $\mathcal{D}_{\mathcal{Z}|y=1}$. However, this does not change the intrinsic nature of the constraint, and the issues raised here remain valid.

where $C\left(s, \widehat{\mathcal{D}}_{\mathcal{Z}}\right)$ can be chosen to obtain any of the two constraints. In the following, we use this general formulation to show that both formulations have shortcomings that can lead to undesired behaviors.

Linear relaxations are too loose. In Figures 2(a) and 2(b) we illustrate the behaviors of $\widehat{\text{DDP}}(f)$ and $\text{LR}_{\widehat{\text{DDP}}}(f)$. In the figures, we consider linear classifiers of the form $f(x) = -x_2 + a_1 x_1 + a_0$ where a_1 controls the slope of the classifier and a_0 the intercept. The underlying data is the same as in Figure 1. It shows that the linear relaxation of DDP can behave completely differently compared to the true DDP. It is particularly striking to notice that the intercept does not have any influence on the constraint. This behavior can be formally verified. Let f be a predictor of the form $f(x) = g(x) + b$ where b is the intercept. Then, $\text{LR}_{\widehat{\text{DDP}}}(f)$ is independent of changes in b since $\frac{1}{n}\sum_{(x,s,y)\in\widehat{\mathcal{D}}_{\mathcal{Z}}} C\left(s, \widehat{\mathcal{D}}_{\mathcal{Z}}\right) = 0$ for both constraints presented above. The proofs are given in the supplementary.

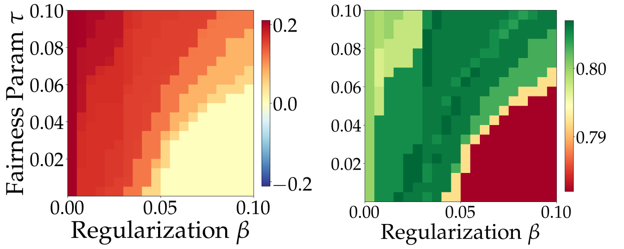
Theoretical guarantees for linear relaxations are not satisfactory. Donini et al. (2018) study a sufficient condition under which the linear fairness relaxation $\text{LR}_{\widehat{\text{DDP}}}(f)$ of a function f is close to its true fairness, that is it holds that $\left|\widehat{\text{DDP}}(f)\right| \leq \left|\text{LR}_{\widehat{\text{DDP}}}(f)\right| + \hat{\Delta}$. The condition that needs to be satisfied by f is

$$\frac{1}{2}\sum_{s'\in\{-1,1\}}\left|\frac{1}{2}\sum_{\substack{(x,s,y)\in\widehat{\mathcal{D}}_{\mathcal{Z}}\\ s=s'}}(\text{sign}(f(x)) - f(x))\right| \leq \hat{\Delta}.$$

Unfortunately, the left hand side of this condition is nonconvex and thus, it is difficult to use in practice. In particular, when they learn a classifier with their linear relaxation, Donini et al. (2018) do not ensure that it also has a small $\hat{\Delta}$. They only verify this condition when the learning process is over, that is when a classifier f has already been produced. However, at this time, it is also possible to compute $\widehat{\text{DDP}}(f)$ directly, so the bound is not needed anymore.

If one could show that for a given function class $\mathcal{F}$, there exists a small $\hat{\Delta}$ such that the condition holds for all $f \in \mathcal{F}$, then any classifier learned from this function class would be guaranteed to be fair when $\left|\text{LR}_{\widehat{\text{DDP}}}(f)\right|$ is small. However, it is not clear whether such function class exists. Nevertheless, this argument hints that for linear relaxations of the fairness constraint, the complexity of the function class largely controls the DDP that can be achieved.

Linear relaxations should not be combined with complex classifiers. We demonstrate that, if the class of classifiers $\mathcal{F}$ is complex, then the linear relaxation constraint has almost no influence on the outcome of the optimization problem. In Figure 3, we compare the performance, in terms of empirical DDP and accuracy, of several models learned by Optimization Problem 4 equipped with the linear relaxation for different parameters β (for regularization) and τ (for fairness). Intuitively, one would expect that varying τ leads to changes in the fairness level while varying β leads to changes in accuracy. However, this is not the case: τ only has an effect on the result when β is sufficiently large. It means that the fairness of the model is mainly controlled by the regularization parameter rather than the fairness one.



(a) DDP.

(b) Accuracy.

Figure 3. We consider a similarity-based classifier (Section 5) with rbf kernel and 1000 train and test points from the Adult dataset. Using a varying regularization parameter β and fairness parameter τ, we train several classifiers using the linear fairness relaxation (Section 3.1). We plot the empirical test DDP of the learned models in Figure 3(a) (red and blue are bad, yellow is good) and their accuracy in Figure 3(b) (red is bad, green is good). We can see that, if β is small (complex model), the fairness relaxation parameter τ has no influence on the DDP score. For higher values of β (simpler models), decreasing τ improves the DDP. Best viewed in color.

This would not be an issue if the fairness of complex classifiers was small. Unfortunately, high-complexity models have a high capacity to alter their decision boundaries. It means that to achieve both high accuracy and high fairness at the same time, they tend to alter their prediction margin for a few selected examples. While this might not affect the accuracy by a lot, the linear relaxation is sensible to this kind of changes and thus can be largely improved—which is what the optimization aims for. However, altering labels of individual points does not have a big influence on the true DDP: it remains high. This effect is reduced when one learns models of low capacity, which have less freedom to deliberately change labels of individual points. Overall, linear relaxations are mainly relevant for simple classifiers and tend to have little effect on complex ones. We outline this undesirable behavior in the experiments.

3.2. Other Relaxations

In the previous section we demonstrated that linear relaxations are not sufficient to ensure fairness of the learned classifier. We now focus on two approaches that use nonlinear relaxations of the indicator function to stay close to the original fairness definition.

Convex-concave relaxation. In a second paper, Zafar et al. (2017a) use the same fairness formulation as Zafar et al. (2017b), but, instead of a linear relaxation of the indicator function, they use a non-linear relaxation.[3] Hence, given $\hat{p}_1$ defined as in Section 3.1, they obtain the constraint:

$$\left|\mathrm{CCR}_{\widehat{\mathrm{DDP}}}(f)\right| = \left|\frac{1}{n}\sum_{(x,s,y)\in\widehat{\mathcal{D}}_{\mathcal{Z}}} \frac{\left(\frac{s+1}{2}-\hat{p}_1\right)}{\hat{p}_1(1-\hat{p}_1)} \min\left(0, f(x)\right)\right| \leq \tau.$$

In Figure 2(c) we give an illustration of $\mathrm{CCR}_{\widehat{\mathrm{DDP}}}(f)$. It more closely approximates the original $\widehat{\mathrm{DDP}}(f)$ than the linear relaxation. Nevertheless, it remains quite far from the original definition—in particular for classifiers that are not constant. Moreover, using such a convex relaxation leads to a convex-concave problem that turns out to be difficult to optimize without guarantees on the global optimality.

Lower-upper relaxation with guarantees. To derive their fairness constraint, Wu et al. (2019) propose to first equivalently rewrite the DDP as follows:

$$\mathrm{DDP}(f) = \mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}\left[\frac{\mathbb{I}_{s=1}}{p_1}\mathbb{I}_{f(x)>0} + \frac{\mathbb{I}_{s=-1}}{1-p_1}\mathbb{I}_{f(x)<0} - 1\right]$$

where p_1 is defined as in Section 3.1. Replacing the indicator functions with a convex surrogate other than the linear one would lead to a convex-concave problem due to the absolute value in the constraint. Instead, Wu et al. (2019) propose to use a convex surrogate function κ for the requirement $\mathrm{DDP}(f) < \tau$ and a concave surrogate function δ for $\mathrm{DDP}(f) > -\tau$. The corresponding relaxation is

$$\mathrm{DDP}_{\kappa}(f) = \mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}\left[\frac{\mathbb{I}_{s=1}}{p_1}\kappa(f(x)) + \frac{\mathbb{I}_{s=-1}}{1-p_1}\kappa(-f(x)) - 1\right],$$

and $\mathrm{DDP}_{\delta}(f)$ is defined analogously by simply replacing κ with δ. It leads to the following convex problem:

$$\begin{aligned} \min_{f\in\mathcal{F}} \quad & \hat{L}(f) + \beta\Omega(f) \qquad (5)\\ \text{subject to} \quad & \widehat{\mathrm{DDP}}_{\kappa}(f) \leq \tau_{\kappa} \\ & -\widehat{\mathrm{DDP}}_{\delta}(f) \leq \tau_{\delta}. \end{aligned}$$

Individually, the relaxations are far from the original fairness constraint (as illustrated in Figures 2(e) and 2(d)) but the idea is that combining the upper bound and the lower bound will help to learn a fair classifier. However, one needs to choose τ_{κ} and τ_{δ} appropriately. To address this, Wu et al. (2019) show that choosing

$$\begin{aligned} \tau_{\kappa} &= \psi_{\kappa}\left(\tau_{\text{upper}} - \widehat{\mathrm{DDP}}^{+}\right) + \widehat{\mathrm{DDP}}_{\kappa}^{-} \\ \tau_{\delta} &= \psi_{\delta}\left(\tau_{\text{lower}} + \widehat{\mathrm{DDP}}^{-}\right) + \widehat{\mathrm{DDP}}_{\delta}^{+}, \end{aligned}$$

guarantees that $-\tau_{\text{lower}} \leq \widehat{\mathrm{DDP}}(f) \leq \tau_{\text{upper}}$. Here $\widehat{\mathrm{DDP}}^{+}$ and $\widehat{\mathrm{DDP}}^{-}$ are the worst possible scores of $\widehat{\mathrm{DDP}}(f)$: they are attained by those functions in the given function class that advantage either group $s=-1$ or group $s=1$ the most. The values $\widehat{\mathrm{DDP}}_{\kappa}^{-}$ and $\widehat{\mathrm{DDP}}_{\delta}^{+}$ are defined in the same way for the relaxed scores. The functions ψ_{κ} and ψ_{δ} are invertible functions that depend on the selected surrogate.

While this solution is appealing at a first glance, it turns out that Optimization Problem 5 is often infeasible for meaningful values of τ_{upper} and τ_{lower} as the constraints form disjoint convex sets. To illustrate this, consider $\kappa(x) = \max\{0, 1+x\}$ and $\delta(x) = \min\{1, x\}$ as proposed by Wu et al. (2019) and the dataset used in Figure 1. Then, if $\tau_{\text{upper}} = \tau_{\text{lower}} \leq 1.13$, the problem is infeasible. If $\tau_{\text{lower}} = 0$ and $\tau_{\text{upper}} \leq 1.95$ the problem is also infeasible. Overall, the guarantees are often meaningless: they either make statements about the empty set (no feasible solution) or they are too loose to ensure meaningful levels of fairness.

4. New Approach with Guaranteed Fairness

In the previous section, we have seen that existing approaches use relaxations of the fairness constraint that lead to tractable optimization problems but have little control over the true fairness of the learned model. For this reason, we propose a new framework that solves the problem of finding *provably fair* solutions: given a convex approximation of the fairness constraint, our method is guaranteed to find a classifier with a good level of fairness.

We consider the following optimization problem:

$$f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda) = \mathop{\arg\min}_{f\in\mathcal{F}} \hat{L}(f) + \lambda \mathrm{R}_{\widehat{\mathrm{DDP}}}(f) + \beta\Omega(f), \qquad (6)$$

where $\mathrm{R}_{\widehat{\mathrm{DDP}}}(f)$ is a convex approximation of the signed fairness constraint, that is we do not consider the usual absolute value. In other words, we obtain a trade-off between accuracy and fairness that is controlled by two hyperparameters $\lambda \geq 0$ and $\beta > 0$ and, given β fixed, we can vary λ to move from strongly preferring one group to strongly preferring the other group. Our goal is then to find a parameter setting that is in the neutral regime and does not favor any of the two groups. The main theoretical ingredient for this procedure to succeed is the next theorem, which states that the function $\lambda \mapsto \mathrm{DDP}\left(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda)\right)$ is continuous under reasonable assumptions on the data distribution, the candidate classifiers, and the convex relaxation.

[3] Zafar et al. (2017a) originally consider other notions of fairness than DPP, among them is the τ-DEO fairness (Equation (5) in their paper). Instead of drawing the examples from $\mathcal{D}_{\mathcal{Z}}$, they consider the conditional distribution $\mathcal{D}_{\mathcal{Z}|y=1}$.

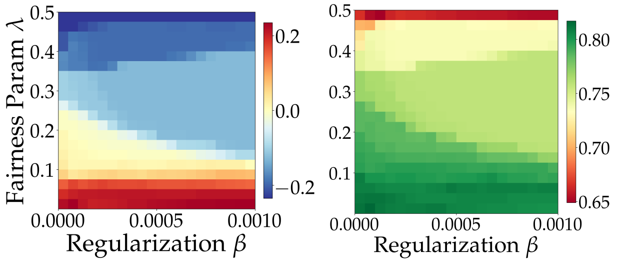
Theorem 1 (Continuity of DDP$\left(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda)\right)$). *Let $\mathcal{F}$ be a function space, we define the set of learnable functions as $\mathcal{F}_{\Lambda} = \left\{f \in \mathcal{F} : \exists \lambda \geq 0, f = f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda)\right\}$. Assume that the following conditions hold:*
(i) *Optimization Problem 6 is m-strongly convex in f,*
(ii) *$\forall f \in \mathcal{F}$, $R_{\widehat{DDP}}(f)$ is bounded in $[-B, B]$,*
(iii) *$\exists \rho$, a metric, such that $(\mathcal{F}_{\Lambda}, \rho)$ is a metric space,*
(iv) *$\forall x \in \mathcal{X}$, $g(f) : f \mapsto f(x)$ is continuous,*
(v) *$\forall f \in \mathcal{F}_{\Lambda}$, f is Lebesgue measurable and the set $\{x : (x, s, y) \in \mathcal{Z}, s = 1, f(x) = 0\}$ is a Lebesgue null set, as well as $\{x : (x, s, y) \in \mathcal{Z}, s = -1, f(x) = 0\}$,*
(vi) *the probability density functions $f_{\mathcal{Z}|s=1}$ and $f_{\mathcal{Z}|s=-1}$ are Lebesgue-measurable.*
Then, the function $\lambda \mapsto$ DDP$\left(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda)\right)$ is continuous.

The proof of this theorem is given in the supplementary. The conditions (i) - (vi) are of a technical nature, but not very restrictive: Condition (i) can be satisfied by using a strongly convex regularization term, for example the squared L_2 norm. Condition (ii) can be satisfied by assuming that $\mathcal{X}$ is bounded. Condition (iii) is, for example, satisfied by any Hilbert Space equipped with the standard dot product. This includes, but is not restricted to, the set of linear classifiers. Condition (iv) ensures that small changes in the hypothesis, with respect to the metric associated to $\mathcal{F}$, also yield small changes in the predictions. Condition (v) ensures that the number of examples for which the predictions are zero is negligible, for example this happens when the decision boundary is sharp. Condition (vi) is satisfied by many usual distributions, for example the Gaussian distribution.

We demonstrate the continuous behavior of DDP on a real dataset in Figure 4. We plot the DDP score and the accuracy of classifiers learned with Optimization Problem 6 for varying parameters λ and β. Given a fixed β, the results support our theoretical findings: there is a smooth transition between favouring the group $s = 1$ with small λ and favouring the group $s = -1$ with higher λ. In between, there is always a region of perfect fairness. In the next corollary, we formally state the conditions necessary to ensure the existence of such a DDP-fair classifier.

Corollary 1 (Existence of a DDP-fair classifier). *Assume that the conditions of Theorem 1 hold and that the convex approximation $R_{\widehat{DDP}}(f)$ is chosen such that for Optimization Problem (6) there exist*
(i) *λ_{+} such that DDP$\left(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_{+})\right) > 0$,*
(ii) *λ_{-} such that DDP$\left(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_{-})\right) < 0$.*
Then, there exists at least one value λ_0 in the interval $[\min(\lambda_{+}, \lambda_{-}), \max(\lambda_{+}, \lambda_{-})]$ such that DDP$\left(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_0)\right) = 0$.

We prove this corollary in the supplementary. This suggests a very simple framework to learn provably fair models. First, we choose a convex fairness relaxation (e.g. the one proposed by Wu et al. (2019)) and search for two initial hyperparameters λ_{+} and λ_{-} that fulfill the assumptions of Corollary 1 (empirically, $\lambda = 0$ and 1 are good candidates). Then, we use a binary search to find a λ_0 between λ_{+} and λ_{-} such that DDP$\left(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_0)\right) = 0$. We call this procedure *SearchFair* and summarize it in Algorithm 1 in the supplementary. Note that any convex approximation R$_{\widehat{\text{DDP}}}(f)$ can be used as long as the conditions of Corollary 1 are respected. In Appendix A we give more details on how we choose this relaxation. Finally, SearchFair theoretically requires to evaluate the true population fairness DDP$\left(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda)\right)$ on the underlying distribution $\mathcal{D}_{\mathcal{Z}}$. In practice, we follow the example of existing fairness constraints (Woodworth et al., 2017) and simply approximate this quantity by its empirical counterpart $\widehat{\text{DDP}}\left(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda)\right)$.



(a) DDP.

(b) Accuracy.

Figure 4. We consider a similarity-based classifier (Section 5) with rbf kernel and 1000 train and test points from the Adult dataset. Using a varying regularization parameter β and fairness parameter λ, we train several classifiers using Optimization Problem 6 with the same loss, convex relaxation, and regularization as SearchFair in the experiments. We plot the empirical test DDP of the learned models in (a) (red and blue are bad, yellow is good) and their accuracy in (b) (red is bad, green is good). We can see that, given a fixed regularization β, we can move from positive DDP (small λ, in red) to a negative DDP (large λ, in blue) with a region of perfect fairness in between (in yellow).

5. Towards Classifiers that are Fair and Accurate

In the last section, we presented a method that is guaranteed to find a DDP fair classifier. However, there is one important catch: we did not make any statement about the classification accuracy of this solution. Here, we take a step in this direction by proposing some sufficient conditions that ensure the existence of a classifier that is both fair and accurate. To this end, we focus on a particular set of classifiers: the family of similarity-based functions. Given a similarity function $K : \mathcal{X} \times \mathcal{X} \to [-1, 1]$ and a set of points $S = \{(x'_1, s'_1, y'_1), \ldots, (x'_d, s'_d, y'_d)\}$, we define a similarity based classifier as $f(x) = \sum_{i=1}^{d} \alpha_i K(x, x'_i)$. The goal is then to learn the weights α_i.

A theory of learning with such functions has been developed by Balcan et al. (2008). By defining a notion of good similarities, they provide sufficient conditions that ensure the existence of an accurate similarity-based classifier. Here, we build upon this framework and we introduce a notion of good similarities for both accuracy and fairness. Hence, in Definition 1 we give sufficient conditions that ensure the existence of a classifier that is—within a guaranteed margin—fair and accurate at the same time.

Definition 1 (Good Similarities for Fairness). *A similarity function K is $(\varepsilon, \gamma, \tau)$-good for convex, positive, and decreasing loss ℓ and (μ, ν)-fair for demographic parity if there exists a (random) indicator function $R(x, s, y)$ defining a (probabilistic) set of "reasonable points" such that, given that $\forall x \in \mathcal{X}, g(x) = \mathbb{E}_{(x',s',y') \sim \mathcal{D}_{\mathcal{Z}}} [y' K(x, x') \,|R(x', s', y')]$, the following conditions hold:*

(i) $\mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}} \left[\ell\left(\frac{y g(x)}{\gamma} \right) \right] \leq \varepsilon,$

(ii) $\left| \mathop{\mathbb{P}}_{\mathcal{D}_{\mathcal{Z}\,|s=1}} [g(x) \geq \gamma] - \mathop{\mathbb{P}}_{\mathcal{D}_{\mathcal{Z}\,|s=-1}} [g(x) \geq \gamma] \right| \leq \mu,$

(iii) $\mathop{\mathbb{P}}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}} [|g(x)| \geq \gamma] \geq 1 - \nu,$

(iv) $\mathop{\mathbb{P}}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}} [R(x, s, y)] \geq \tau.$

Roughly speaking, a similarity is good for classification if examples of the same class are on average closer to each other than examples of different classes up to a certain margin. Moreover, it is good for fairness if this margin is independent of group membership. Given such a similarity, we can prove the existence of a fair and accurate classifier as is summarized in the next theorem. The proof is given in the supplementary.

Theorem 2 (Existence of a fair and accurate separator). *Let $K \in [-1, 1]$ be a $(\varepsilon, \gamma, \tau)$-good and (μ, ν)-fair metric for a given convex, positive and decreasing loss ℓ with lipschitz constant L. For any $\varepsilon_1 > 0$ and $0 < \delta < \frac{\gamma \varepsilon_1}{2(L + \ell(0))}$, let $S = \{(x'_1, s'_1, y'_1), \ldots, (x'_d, s'_d, y'_d)\}$ be a set of d examples drawn from $\mathcal{D}_{\mathcal{Z}}$ with*

$$d \geq \frac{1}{\tau} \left[\frac{L^2}{\gamma^2 \varepsilon_1^2} + \frac{3}{\delta} + \frac{4L}{\delta \gamma \varepsilon_1} \sqrt{\delta (1 - \tau) \log(2/\delta)} \right].$$

Let $\phi^S : \mathcal{X} \to \mathbb{R}^d$ be a mapping with $\phi^S_i(x) = K(x, x'_i)$, for all $i \in \{1, \ldots, d\}$. Then, with probability at least $1 - \frac{5}{2}\delta$ over the choice of S, the induced distribution over $\phi^S(\mathcal{X}) \times \mathcal{S} \times \mathcal{Y}$ has a linear separator α such that

$$\mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}} \left[\ell\left(\frac{y \left\langle \alpha, \phi^S(x) \right\rangle}{\gamma} \right) \right] \leq \varepsilon + \varepsilon_1,$$

and, with $p_1 = \mathbb{P}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}} [s = 1]$,

$$|DDP(\alpha)| \leq \mu + (\nu + 2\delta) \max\left(\frac{1}{p_1}, \frac{1}{1 - p_1} \right).$$

6. Experiments

In this section, we empirically evaluate SearchFair by comparing it to 5 baselines on 6 real-world datasets. In all the experiments, SearchFair either reliably finds the fairest classifier and is comparable to a very recent non-convex optimization approach.

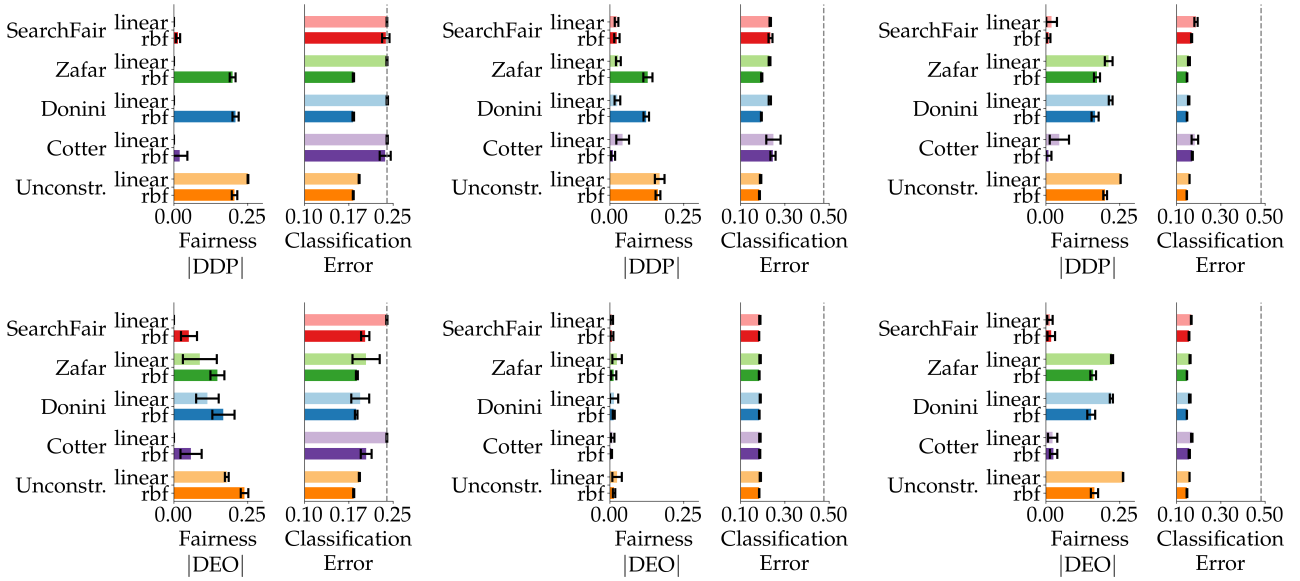
Datasets. We consider 6 different datasets: CelebA (Liu et al., 2015), Adult (Kohavi & Becker, 1996), Dutch (Zliobaite et al., 2011), COMPAS (Larson et al., 2016), Communities and Crime (Redmond & Baveja, 2002), and German Credit (Dua & Graff, 2017). In the supplementary, we give detailed descriptions of these datasets, how we preprocess the data, and the sizes of the train and test splits. Note that we remove the sensitive attribute s from the set of features x so that it is not needed at decision time.

Baselines. We compare SearchFair to 5 baselines. For 3 of them, we use Optimization Problem 4 with hinge loss and a squared ℓ_2 norm as the regularization term. As a function class $\mathcal{F}$, we use similarity-based classifiers presented in Section 5 with either the linear or the rbf kernel and with 70% (at most 1000) of the training examples as reasonable points. As a fairness constraint, we use either the linear relaxation of Zafar et al. (2017b) (Zafar), the linear relaxation of Donini et al. (2018) (Donini), or no constraint at all (Unconst). The fourth baseline is a recent method for non-convex constrained optimization by Cotter et al. (2019) (Cotter). Our last baseline is the constant classifier (Constant) that always predicts the same label but has perfect fairness.

SearchFair.[4] For SearchFair we also use the hinge loss, a squared ℓ_2 norm as the regularization term (it is strongly convex), and similarity-based classifiers. As a convex approximation of the fairness constraint, we use the bounds with hinge loss proposed by Wu et al. (2019) (see Section A in the supplementary for details).

Metrics. Our main goal is to learn fair classifiers. Hence, our main evaluation metrics are the empirical DDP and DEO scores on the test set (lower is better). As a secondary metric (in case of ties in the fairness scores), we consider the classification performance of the models and we report the errors on the test set (lower is better). All the experiments are repeated 10 times and we report the mean and standard deviation for all the metrics.

[4] The code is freely available online: `github.com/mlohaus/SearchFair`.



(a) Adult. (b) Dutch. (c) CelebA.

Figure 5. We report the average and standard deviation of classification error and absolute fairness scores DDP and DEO (closer to 0 is better) over 10 repetitions. The constant classifier is perfectly fair as it always predicts the same label. Its classification error is shown by the grey dashed vertical line. (a) To obtain good fairness on Adult, all DDP fair methods learn the constant classifier. Only SearchFair and Cotter reliably find fair classifiers for both kernels. (b) On Dutch, SearchFair obtains the lowest DDP with a slight loss in accuracy. Cotter performs comparably for both kernels, whereas the other methods only do well with a linear kernel but fail to learn fair classifiers with the rbf kernel. (c) For CelebA, SearchFair and Cotter are the only methods that obtain a low DDP and DEO with only a slight loss in accuracy. The other methods only provide little to no improvement.

Hyperparameters. Zafar, Donini and Cotter use a fairness parameter, that we call τ, to control the fairness level. Since our goal is to learn classifiers that are fair, we set $\tau = 0$ such that perfect fairness is required. For SearchFair, there is no fairness parameter since λ_0 is automatically searched for between a lower bound λ_{min} and an upper bound λ_{max}. We set $\lambda_{\text{min}} = 0$ and $\lambda_{\text{max}} = 1$ as these values usually lead to classifiers with fairness scores of opposite sign (as needed). We use 10 iterations in the binary search.

We use 5-fold cross validation to choose other hyperparameters. For Cotter, only the width of the rbf kernel has to be tuned since we use the framework of the original paper with no regularization term. For all remaining methods we need to choose the regularization parameter β and the width of the rbf kernel. These values are respectively chosen in the sets $\left\{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\right\}$ and $\{10^{-\lceil \log d \rceil -1}, 10^{-\lceil \log d \rceil}, d^{-1}, 10^{-\lceil \log d \rceil +1}, 10^{-\lceil \log d \rceil +2}\}$, with d the number of features. We select the set of parameters that lead to the most accurate classifier on average over the 5 folds. Indeed, the fairness level is automatically taken care of by the methods.

Results. We present the results for 3 out of 6 datasets in Figure 5. The other results are deferred to the supplementary as they follow the same trend. We make two main observations. First, SearchFair always obtains fairness values that are very close to zero. It learns the fairest classifiers out of all the methods and is only matched by Cotter, the non-convex approach. This sometimes comes with a small increase in terms of classification error. For example, in order to achieve perfect DDP fairness on the Adult dataset, SearchFair, and all the other fair methods, yield classifiers close to the trivial constant one. Second, the complexity of the model greatly influences the performances of the linear relaxations. For example, using the complex rbf kernel almost always results in an increase in the fairness score of Zafar and Donini. This is particularly striking for Adult and Dutch where the linear kernel yields reasonable fairness scores. Note that this trend is not always respected. For example, on CelebA, using an rbf kernel improves the fairness score compared to the linear kernel. However, neither of them obtain reasonable fairness levels in the first place.

Discussion on hyperparameter selection. Apart from the hyperparameter selection method used in our experiments, one can think of other cross validation procedures. For example, Donini et al. (2018) proposed NVP, a cross validation method where one selects the set of hyperparameters that gives the fairest classifier while obtaining an average accuracy above a given threshold. Similarly, one could select the set of hyperparameters that yields the most accurate classifier under a given fairness threshold. In the supplementary, we report results that empirically show that these more complex procedures tend to improve the fairness of the baselines (SearchFair remains competitive on all the datasets). Unfortunately, they also blur the dividing line between hyperparameters that control the fairness of the model and the ones that control its complexity. In other words, it becomes unclear whether fairness is achieved thanks to the relaxation or thanks to the choice of hyperparameters (we already evoked this issue in Figure 3). We believe that it is better to have a method that is guaranteed to find a fair classifier for any given family of models and does not rely on a complex cross validation procedure.

7. Conclusion

In this paper, we have shown that existing approaches to learn fair and accurate classifiers have many shortcomings. They use loose relaxations of the fairness constraint and guarantees that relate the relaxed fairness to the true fairness of the solutions are either missing or not sufficient. We empirically demonstrated how these approaches can produce undesirable models. If "fair machine learning" is supposed to be employed in real applications in society, we need algorithms that actually find fair solutions, and ideally come with guarantees. We made a first step in this direction by proposing SearchFair, an approach that uses convex relaxations to learn a classifier that is guaranteed to be fair.

Acknowledgements

This work has been supported by the German Research Foundation through the Cluster of Excellence "Machine Learning – New Perspectives for Science" (EXC 2064/1 number 390727645), the BMBF Tübingen AI Center (FKZ: 01IS18039A), and the International Max Planck Research School for Intelligent Systems (IMPRS-IS). It has also been supported by the ACADEMICS grant of the IDEXLYON, project of the Université de Lyon, PIA operated by ANR-16-IDEX-0005. The authors also thank Luca Rendsburg for helpful discussions.

References

Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J., and Wallach, H. A reductions approach to fair classification. In *International Conference on Machine Learning*, 2018.

Balcan, M.-F., Blum, A., and Srebro, N. Improved guarantees for learning via similarity functions. In *Conference on Learning Theory*, 2008.

Calders, T. and Verwer, S. Three naive Bayes approaches for discrimination-free classification. In *International Conference on Data Mining and Knowledge Discovery*, 2010.

Calders, T., Kamiran, F., and Pechenizkiy, M. Building classifiers with independency constraints. In *International Conference on Data Mining Workshops*, 2009.

Calmon, F., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., and Varshney, K. R. Optimized pre-processing for discrimination prevention. In *International Conference on Neural Information Processing Systems*, 2017.

Chzhen, E., Denis, C., Hebiri, M., Oneto, L., and Pontil, M. Leveraging labeled and unlabeled data for consistent fair binary classification. In *International Conference on Neural Information Processing Systems*, 2019.

Cotter, A., Jiang, H., and Sridharan, K. Two-player games for efficient non-convex constrained optimization. In *International Conference on Algorithmic Learning Theory*, 2019.

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J., and Pontil, M. Empirical risk minimization under fairness constraints. In *International Conference on Neural Information Processing Systems*, 2018.

Dua, D. and Graff, C. UCI machine learning repository, 2017.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Innovations in Theoretical Computer Science Conference*, 2012.

Elstrodt, J. *Maß-und Integrationstheorie*, volume 7. Springer, 1996.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and removing disparate impact. In *International Conference on Knowledge Discovery and Data Mining*, 2015.

Goh, G., Cotter, A., Gupta, M., and Friedlander, M. P. Satisfying real-world goals with dataset constraints. In *International Conference in Neural Information Processing Systems*, 2016.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *International Conference on Neural Information Processing Systems*, 2016.

Kamiran, F. and Calders, T. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 2012.

Kamiran, F., Calders, T., and Pechenizkiy, M. Discrimination aware decision tree learning. In *International Conference on Data Mining*, 2010.

Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. Fairness-aware classifier with prejudice remover regularizer. In *Machine Learning and Knowledge Discovery in Databases*, 2012.

Kohavi, R. and Becker, B. Scaling up the accuracy of naive-bayes classifiers: a decision-tree hybrid, 1996.

Larson, J., Mattu, S., Kirchner, L., and Angwin, J. How we analyzed the compas recidivism algorithm. *ProPublica*, 2016.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *International Conference on Computer Vision*, 2015.

Redmond, M. A. and Baveja, A. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, 2002.

Woodworth, B., Gunasekar, S., Ohannessian, M. I., and Srebro, N. Learning non-discriminatory predictors. In *Conference on Learning Theory*, 2017.

Wu, Y., Zhang, L., and Wu, X. On convexity and bounds of fairness-aware classification. In *The World Wide Web Conference*, 2019.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *International Conference on World Wide Web*, 2017a.

Zafar, M. B., Valera, I., Rogriguez, M. G., and Gummadi, K. P. Fairness Constraints: Mechanisms for Fair Classification. In *International Conference on Artificial Intelligence and Statistics*, 2017b.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. Learning fair representations. In *International Conference on Machine Learning*, 2013.

Zliobaite, I., Kamiran, F., and Calders, T. Handling conditional discrimination. In *International Conference on Data Mining*, 2011.

This supplementary gives technical details about SearchFair in Section A and details on our experiments in Section B. The proof of the vanishing intercept problem, the derivations of the equivalent formulations of DDP and the proofs of all theorems are given in Section C.

A. SearchFair: A Binary Search Framework for Fairness

This section presents technical details about SearchFair that were omitted in the main paper. We present our binary search based algorithm in Algorithm 1.

Recall that

$$f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda) = \operatorname*{arg\,min}_{f\in\mathcal{F}} \hat{L}(f) + \lambda \mathrm{R}_{\widehat{\mathrm{DDP}}}(f) + \beta\Omega(f)\,.$$

Then, we choose a lower bound λ_{min} and an upper bound λ_{max}, so that $\operatorname{sign}\Big(\mathrm{DDP}\Big(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_{\text{min}})\Big)\Big) \neq \operatorname{sign}\Big(\mathrm{DDP}\Big(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_{\text{max}})\Big)\Big)$. If $\mathrm{R}_{\widehat{\mathrm{DDP}}}(f)$ is chosen correctly then setting $\lambda_{\text{min}} = 0$ and $\lambda_{\text{max}} = 1$ usually works. The number of iterations C is used to control how close to 0 the fairness measure should be. Note that, instead of a number of iterations, it is also possible to choose a stopping criterion, for example when DDP falls below a threshold.

Example of convex relaxation. One example of a convex relaxation is to use the bounds proposed by Wu et al. (2019). When no fairness regularizer is used, we evaluate the fairness of the resulting classifier and choose an approximation accordingly. More precisely, with $\lambda = 0$ if $\mathrm{DDP}(f(\lambda)) > 0$ we use the upper bound with hinge loss:

$$\mathrm{R}_{\widehat{\mathrm{DDP}}}(f) = \frac{1}{n}\sum_{(x,s,y)\in\widehat{\mathcal{D}}_{\mathcal{Z}}}\left[\frac{\mathbb{I}_{s=1}}{p_1}\max\left(0, 1+f(x)\right) + \frac{\mathbb{I}_{s=-1}}{1-p_1}\max\left(0, 1-f(x)\right) - 1\right].$$

If $\mathrm{DDP}(f(\lambda)) < 0$, we use the negative lower bound with hinge loss:

$$\mathrm{R}_{\widehat{\mathrm{DDP}}}(f) = -\frac{1}{n}\sum_{(x,s,y)\in\widehat{\mathcal{D}}_{\mathcal{Z}}}\left[\frac{\mathbb{I}_{s=1}}{p_1}\min\left(1, f(x)\right) - \frac{\mathbb{I}_{s=-1}}{1-p_1}\min\left(1, -f(x)\right) + 1\right].$$

With $\lambda_{\text{min}} = 0$ and $\lambda_{\text{max}} = 1$ this choice often ensures that $\operatorname{sign}(\mathrm{DDP}(f(\lambda_{\text{min}}))) \neq \operatorname{sign}(\mathrm{DDP}(f(\lambda_{\text{max}})))$. We use this approach in all our experiments in the paper.

Note that we give an example where the relaxations are in fact upper and lower bounds of the DDP score. However, we want to stress that any convex *approximation* would work as long as the condition $\operatorname{sign}\Big(\mathrm{DDP}\Big(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_{\text{min}})\Big)\Big) \neq \operatorname{sign}\Big(\mathrm{DDP}\Big(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_{\text{max}})\Big)\Big)$ is respected.

Satisfying the conditions of Theorem 3. The strong convexity of the optimization problem (condition (i)) can be ensured by choosing a strongly convex regularization term (we adopt this strategy in our experiments).

Satisfying conditions (ii) to (v) mainly depends on our choice of function class $\mathcal{F}$. For example, linear classifiers satisfy all the conditions as long as $\mathcal{X}$ is bounded (which is the case in most machine learning applications) and the classifier $f^0(x) = \mathbf{0}^T x$, where $\mathbf{0}$ is the vector of all zeros, is not part of the set of learnable functions $\mathcal{F}_\Lambda$ (otherwise condition (v) would be violated). To verify that $f^0 \notin \mathcal{F}_\Lambda$, it is sufficient to verify that the equation $\frac{d\hat{L}(f^0)}{df} + \lambda\frac{d\mathrm{R}_{\widehat{\mathrm{DDP}}}(f^0)}{df} + \beta\frac{d\Omega(f^0)}{df} = \mathbf{0}$ with β fixed has no solutions for $\lambda \in [\lambda_{\text{min}}, \lambda_{\text{max}}]$. Note that, in practice, this is usually easy to verify and can be achieved by correctly choosing $\mathrm{R}_{\widehat{\mathrm{DDP}}}(f)$. Note that the similarity-based classifiers that we use in our experiments are a particular form of linear classifiers and thus satisfy conditions (ii) to (v).

Finally, condition (vi) depends on the data distribution and should be satisfied for most non-degenerate problems.

B. Missing details on the Experiments

In this section we provide missing details on the experiments. First, we shortly describe the toy dataset in Figure 1 of the main paper. Second, we describe the pre-processing step and each dataset. Then, we present detailed results on all 6

Algorithm 1 SearchFair: A binary search framework for fairness

```
**Input**: A set $\widehat{\mathcal{D}}_{\mathcal{Z}} = (x_i, s_i, y_i)_{i=1}^n$ of $n$ labelled examples, a regularization parameter $\beta > 0$, $\lambda_{\text{min}}$ and $\lambda_{\text{max}}$ the lower and upper bounds for $\lambda$, a convex fairness regularizer $\text{R}_{\widehat{\text{DDP}}}(\cdot)$, a number of iterations $C$.
**Output**: A fair classifier.
if $\text{DDP}\Big(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_{\text{min}})\Big) > 0$ and $\text{DDP}\Big(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_{\text{max}})\Big) < 0$ then
    $\lambda_+ = \lambda_{\text{min}}$ and $\text{DDP}_+ = \text{DDP}\Big(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_{\text{min}})\Big)$.
    $\lambda_- = \lambda_{\text{max}}$ and $\text{DDP}_- = \text{DDP}\Big(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_{\text{max}})\Big)$.
    search_possible = True
else if $\text{DDP}\Big(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_{\text{min}})\Big) < 0$ and $\text{DDP}\Big(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_{\text{max}})\Big) > 0$ then
    $\lambda_- = \lambda_{\text{min}}$ and $\text{DDP}_- = \text{DDP}\Big(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_{\text{min}})\Big)$.
    $\lambda_+ = \lambda_{\text{max}}$ and $\text{DDP}_+ = \text{DDP}\Big(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_{\text{max}})\Big)$.
    search_possible = True
else
    search_possible = False
end if
if search_possible then
    for $c = 1, \ldots, C$ do
        $\lambda = \frac{1}{2}(\lambda_- + \lambda_+)$
        $\text{DDP}_\lambda = \text{DDP}\Big(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda)\Big)$.
        if $\text{DDP}_\lambda > 0$ then
            $\lambda_+ = \lambda$ and $\text{DDP}_+ = \text{DDP}_\lambda$.
        else
            $\lambda_- = \lambda$ and $\text{DDP}_- = \text{DDP}_\lambda$.
        end if
    end for
    if $|\text{DDP}_-| < |\text{DDP}_+|$ then
        return $f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_-)$
    else
        return $f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_+)$
    end if
else
    Either choose new values for $\lambda_{\text{min}}$ and $\lambda_{\text{max}}$, or choose a new fairness regularizer $\text{R}_{\widehat{\text{DDP}}}(f)$.
end if
```

datasets in Figures 6– 11. The overall trend of the results is the same as described in the main paper and we discuss the results of each dataset in the corresponding figure. Lastly, we present the results of the experiments under two different hyperparameters selection methods in Figures 12 and 13. For all results, each experiment has been repeated 10 times and we report the average and standard deviation of classification error and absolute fairness scores DDP and DEO. Furthermore, we report the value of the Donini linear relaxation (Section 3.1) on the test set to show the discrepancy between the true and relaxed fairness (this metric was omitted in the main paper).

Toy dataset–Figure 1 in main paper. The toy dataset set in Figure 1 of the main paper consists of 600 points (for the sake of readability, we only plot a random subset of 400 examples). We draw the points from different Gaussian distributions. For the protected sensitive attribute (the dots), we sample 150 points with negative label from a Gaussian with mean $\mu_1 = [2, -2]$ and covariance matrix $\Sigma_1 = [[1, 0], [0, 1]]$, and another 150 points for the positive class from the mixture of two Gaussians, with $\mu_2 = [3, -1]$ and $\Sigma_2 = [[1, 0], [0, 1]]$ and $\mu_3 = [1, 4]$ and $\Sigma_3 = [[0.5, 0], [0, 0.5]]$. For the unprotected sensitive attribute (the crosses), we draw 150 points with positive label from a Gaussian with $\mu_4 = [2.5, 2.5]$ and $\Sigma_4 = [[1, 0], [0, 1]]$, and 150 points with negative label from a Gaussian with $\mu_5 = [4.5, -1.5]$ and $\Sigma_5 = [[1, 0], [0, 1]]$.

General setup. For all the methods except Cotter, as the set of functions $\mathcal{F}$, we use the similarity-based classifiers that were presented in the main paper. As similarities, we consider both the linear and the rbf kernel. As reasonable points, we use a random subset of 70% (at most 1000) of the training examples. As regularization term we use a squared ℓ_2 norm (which is a strongly convex function). The loss function in the empirical risk is the hinge loss, that is

$$\ell(f(x)\,,y) = \max\left(0, 1 - f(x)\,y\right).$$

For the linear version of Cotter et al. (2019), we use the approach suggested in their example on the Adult dataset. We use a single-layer neural network where the input size is the number of features. The parameters are then learned using the RATEMINIMIZATIONPROBLEM provided by the package TENSORFLOW-CONSTRAINED-OPTIMIZATION. In order to use more complex classifiers based on the rbf kernel, we precompute the kernel matrix between the training points and the reasonable points. Then, the input size of the single-layer neural network is set to the number of reasonable points. For both linear and complex classifiers, no further regularization is used. However, to obtain reasonable and stable results, the number of epochs has to be carefully chosen. We use between 1000 and 5000 epochs depending on the dataset, and for the minibatch size we use the default of 200 points.

We pre-process the datasets by normalizing and centering continuous variables. For categorical values, we use a one-hot encoding. We select a fixed number of randomly selected points for training, and use the rest of the points for testing.

CelebA–Figure 6. The CelebA dataset (Liu et al., 2015) contains $202,599$ images of celebrity faces from the web. In addition to the image data, there exist 40 binary attribute labels describing the content of the images, such as 'Black Hair', 'Bald', and 'Eyeglasses'. We use 38 of those descriptions as features, the sex as the sensitive attribute, and the attribute 'Smiling' as the class label. We use $10,000$ randomly selected points for training.

Adult–Figure 7. The Adult dataset (Kohavi & Becker, 1996) contains data from the U.S. 1994 Census database. There are $48,842$ instances with 14 features, among others age and education, including the two sensitive attributes sex and race. We apply the pre-processing of Wu et al. (2019): we consider sex with values male and female as the sensitive attribute and use 9 features for training, dropping FNLWGT, EDUCATION, CAPITAL-GAIN, CAPITAL-LOSS. The goal is to predict the income: $y = 1$ if it is more than fifty thousand U.S. Dollars, $y = -1$ otherwise. We use $10,000$ randomly selected points for training.

Dutch–Figure 8. The Dutch dataset (Zliobaite et al., 2011) contains data from the 2001 Netherlands Census and consists of $60,420$ data points which are characterized by 12 features. We use gender as the sensitive attribute and predict *low income* or *high income* as it is determined by occupation. Hence, we learn with the remaining 10 features. We use $10,000$ randomly selected points for training.

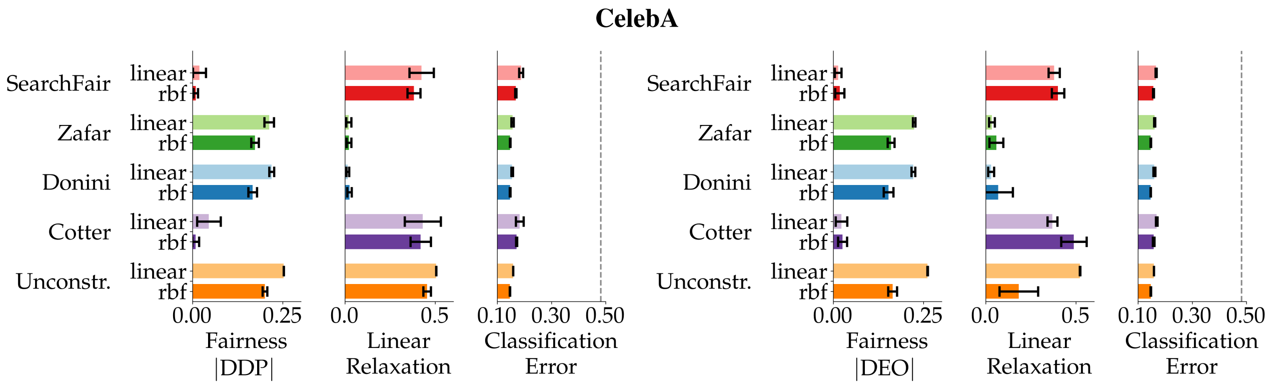
Compas–Figure 9. The Compas dataset (Larson et al., 2016) contains 7214 points with 53 features, such as name, age, degree of crime, and number of prior crimes. We use the same pre-processing as Zafar et al. (2017a) and, in particular, we select the same 5 features. The goal is to predict if a defendant has been arrested again within two years of the decision. The sensitive attribute is race. It has been changed to a binary attribute with the values 'White' and 'NonWhite'. We use $5,000$ randomly selected points for training.

Communities and Crime–Figure 10. This dataset includes socio-economic data of 1994 communities in the United States (Redmond & Baveja, 2002). It consists of 128 attributes, of which we drop the name of the state, county, and community, and features with missing values. Overall, we drop 29 features. We use the attribute RACEPCTWHITE to construct a binary sensitive attribute. A community with a percentage of white residents higher than the mean 0.75 obtains the sensitive label 1, otherwise the label is -1. The goal of this data set is to predict the number of violent crimes. We binarize the label by splitting VIOLENTCRIMESPERPOP at the mean of 0.24. We use $1,500$ randomly selected points for training.

German Credit–Figure 11. There are 1000 records of german applicants for a credit with 20 attributes (Dua & Graff, 2017). The goal is to classify the applicants in creditworthy or not creditworthy. The categorical feature 'personal status' is changed into the binary feature sex. We use it as the sensitive attribute and use the other 19 features for training. We use 700 randomly selected points for training.

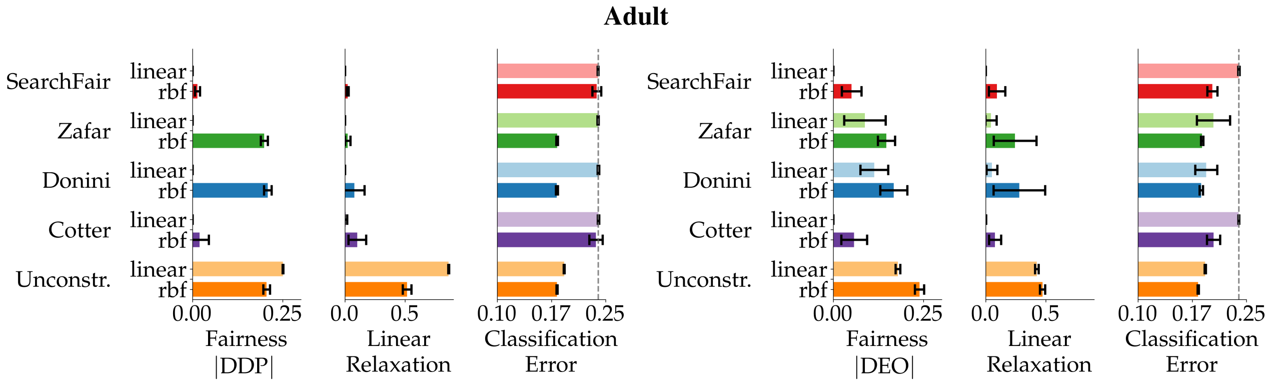
Cross Validation Procedure. We report the results for different cross validation procedures as discussed in the main paper. In Figure 12 we use a procedure called NVP proposed by (Donini et al., 2018). In a first step, we exclude the hyperparameters with an accuracy score that is lower than 90% of the best accuracy score. Then, we choose the set of hyperparameters with the best average fairness score. Finally, we use these hyperparameters to train on the whole train set.

In Figure 13 we report the results when we use a given fairness threshold. We shortlist all hyperparameters with an absolute fairness score lower than 0.05 and, among them, choose the hyperparameters with the highest accuracy score. We report average and standard deviation of classification error and absolute fairness scores DDP and DEO over 10 repetitions. Note that we also report results for the approach by Cotter et al. (2019) for comparison, even though the linear version does not tune any hyperparameters. Using the rbf kernel on the other hand, we need to tune the width of the kernel.



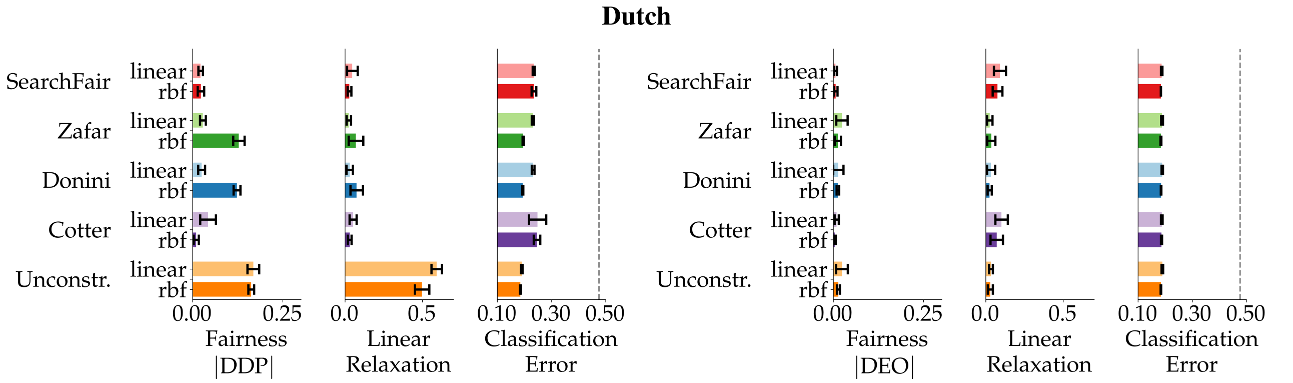
FAIRNESS NOTION		Demographic Parity			Equality of Opportunity		
	Kernel	$\lvert$DDP$\rvert$	$\lvert$LR$\rvert$	Error	$\lvert$DEO$\rvert$	$\lvert$LR$\rvert$	Error
SearchFair	linear	0.02 ± 0.02	0.42 ± 0.07	0.19 ± 0.01	0.01 ± 0.01	0.38 ± 0.01	0.17 ± 0.00
	rbf	0.01 ± 0.01	0.38 ± 0.04	0.17 ± 0.00	0.02 ± 0.01	0.40 ± 0.04	0.16 ± 0.00
Zafar	linear	0.21 ± 0.01	0.02 ± 0.01	0.16 ± 0.00	0.22 ± 0.00	0.04 ± 0.02	0.16 ± 0.00
	rbf	0.17 ± 0.01	0.02 ± 0.01	0.15 ± 0.00	0.16 ± 0.01	0.06 ± 0.04	0.15 ± 0.00
Donini	linear	0.22 ± 0.01	0.02 ± 0.01	0.15 ± 0.00	0.22 ± 0.01	0.03 ± 0.02	0.16 ± 0.00
	rbf	0.17 ± 0.01	0.03 ± 0.01	0.15 ± 0.00	0.15 ± 0.01	0.07 ± 0.08	0.15 ± 0.00
Cotter	linear	0.05 ± 0.03	0.43 ± 0.10	0.18 ± 0.01	0.02 ± 0.02	0.37 ± 0.03	0.17 ± 0.00
	rbf	0.01 ± 0.01	0.42 ± 0.06	0.18 ± 0.00	0.03 ± 0.01	0.49 ± 0.07	0.16 ± 0.00
Unconstrained	linear	0.25 ± 0.00	0.51 ± 0.00	0.16 ± 0.00	0.26 ± 0.00	0.52 ± 0.00	0.16 ± 0.00
	rbf	0.20 ± 0.01	0.46 ± 0.02	0.15 ± 0.00	0.16 ± 0.01	0.18 ± 0.11	0.15 ± 0.00
Constant	–	0.00 ± 0.00	–	0.48 ± 0.00	0.00 ± 0.00	–	0.48 ± 0.00

Figure 6. **CelebA.** The grey dashed vertical line depicts the classification error of the constant classifier, which is perfectly fair for both DDP and DEO. Overall, SearchFair is the only method with both low DDP and DEO, while linear Cotter is only slightly worse for DDP. Additionally, SearchFair and Cotter exhibit high values for the linear relaxations which might imply that this relaxation is not suitable here. This is confirmed by the fact that the competing methods have low relaxation values with high DDP and DEO values.



FAIRNESS NOTION		Demographic Parity			Equality of Opportunity		
	Kernel	|DDP|	|LR|	Error	|DEO|	|LR|	Error
SearchFair	linear	0.00 ± 0.00	0.00 ± 0.00	0.24 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.24 ± 0.00
	rbf	0.01 ± 0.01	0.02 ± 0.01	0.24 ± 0.01	0.05 ± 0.03	0.09 ± 0.07	0.20 ± 0.01
Zafar	linear	0.00 ± 0.00	0.00 ± 0.00	0.24 ± 0.00	0.09 ± 0.06	0.04 ± 0.05	0.20 ± 0.02
	rbf	0.20 ± 0.01	0.02 ± 0.02	0.18 ± 0.00	0.15 ± 0.02	0.24 ± 0.18	0.19 ± 0.00
Donini	linear	0.00 ± 0.00	0.00 ± 0.00	0.24 ± 0.00	0.11 ± 0.04	0.05 ± 0.05	0.19 ± 0.02
	rbf	0.21 ± 0.01	0.08 ± 0.08	0.18 ± 0.00	0.17 ± 0.04	0.28 ± 0.21	0.19 ± 0.00
Cotter	linear	0.00 ± 0.00	0.01 ± 0.01	0.24 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.24 ± 0.00
	rbf	0.02 ± 0.03	0.10 ± 0.07	0.24 ± 0.01	0.06 ± 0.04	0.08 ± 0.05	0.20 ± 0.01
Unconstrained	linear	0.25 ± 0.00	0.86 ± 0.00	0.19 ± 0.00	0.18 ± 0.01	0.43 ± 0.02	0.19 ± 0.00
	rbf	0.21 ± 0.01	0.52 ± 0.04	0.18 ± 0.00	0.24 ± 0.01	0.47 ± 0.02	0.18 ± 0.00
Constant	–	0.00 ± 0.00	–	0.24 ± 0.00	0.00 ± 0.00	–	0.24 ± 0.00

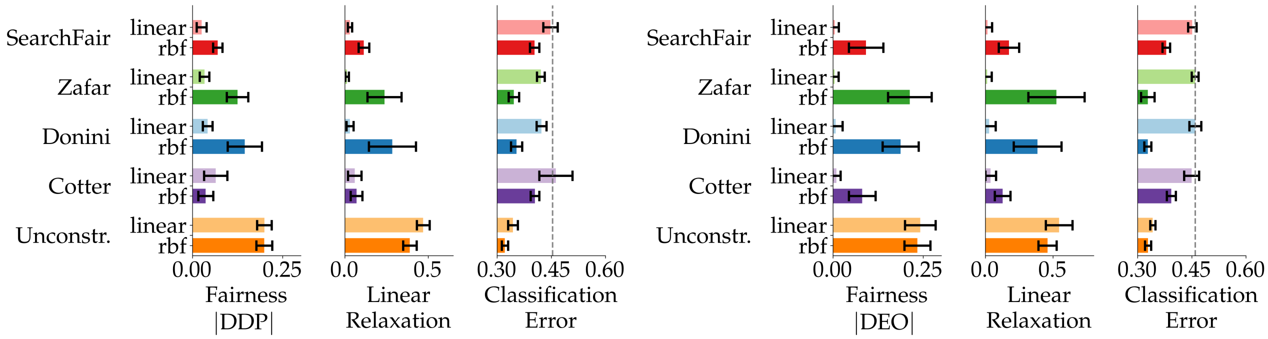
Figure 7. **Adult.** The grey dashed vertical line depicts the classification error of the constant classifier, which is perfectly fair for both DDP and DEO. All the methods tend to learn the constant classifier to obtain a DDP fair model with the linear kernel. With the rbf kernel, SearchFair and Cotter obtain low fairness scores (both for DDP and DEO) showing that the fairness of the model learned by the relaxation based baselines can be heavily linked to the complexity of the models. Note that, even though all the fairness methods learn classifiers with a low linear relaxation, their DDP scores vary widely. It confirms that there is no guarantee that a low relaxation value will lead to a fair classifier.



FAIRNESS NOTION		Demographic Parity			Equality of Opportunity		
	Kernel	$\lvert\text{DDP}\rvert$	$\lvert\text{LR}\rvert$	Error	$\lvert\text{DEO}\rvert$	$\lvert\text{LR}\rvert$	Error
SearchFair	linear	0.02 ± 0.01	0.05 ± 0.03	0.23 ± 0.00	0.01 ± 0.00	0.09 ± 0.04	0.19 ± 0.00
	rbf	0.02 ± 0.01	0.03 ± 0.01	0.23 ± 0.01	0.01 ± 0.00	0.08 ± 0.03	0.18 ± 0.00
Zafar	linear	0.03 ± 0.01	0.03 ± 0.01	0.23 ± 0.00	0.02 ± 0.02	0.03 ± 0.02	0.19 ± 0.00
	rbf	0.13 ± 0.02	0.07 ± 0.05	0.20 ± 0.00	0.01 ± 0.01	0.04 ± 0.03	0.18 ± 0.00
Donini	linear	0.03 ± 0.01	0.03 ± 0.02	0.23 ± 0.00	0.01 ± 0.01	0.03 ± 0.03	0.19 ± 0.00
	rbf	0.12 ± 0.01	0.08 ± 0.04	0.19 ± 0.00	0.01 ± 0.00	0.03 ± 0.01	0.18 ± 0.00
Cotter	linear	0.04 ± 0.02	0.05 ± 0.02	0.25 ± 0.03	0.01 ± 0.01	0.10 ± 0.04	0.19 ± 0.00
	rbf	0.01 ± 0.01	0.03 ± 0.01	0.25 ± 0.01	0.00 ± 0.00	0.07 ± 0.04	0.19 ± 0.00
Unconstrained	linear	0.17 ± 0.02	0.59 ± 0.03	0.19 ± 0.00	0.02 ± 0.02	0.03 ± 0.01	0.19 ± 0.00
	rbf	0.16 ± 0.01	0.50 ± 0.05	0.19 ± 0.00	0.01 ± 0.00	0.03 ± 0.02	0.18 ± 0.00
Constant	–	0.00 ± 0.00	–	0.48 ± 0.00	0.00 ± 0.00	–	0.48 ± 0.00

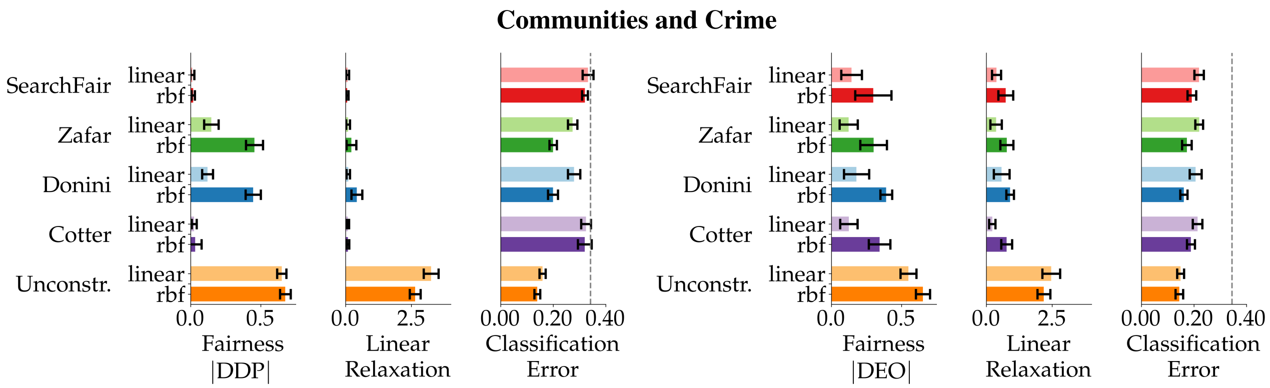
Figure 8. **Dutch.** The grey dashed vertical line depicts the classification error of the constant classifier, which is perfectly fair for both DDP and DEO. In terms of DEO all the methods perform equally well on this dataset as the Unconstrained classifier is already DEO fair. On the other hand, SearchFair and Cotter obtain a low DDP regardless of the complexity of the model. Once again, even though all the fairness methods learn classifiers with a low linear relaxation, their DDP scores vary widely. It confirms that there is no guarantee that a low relaxation value will lead to a fair classifier.

Compas



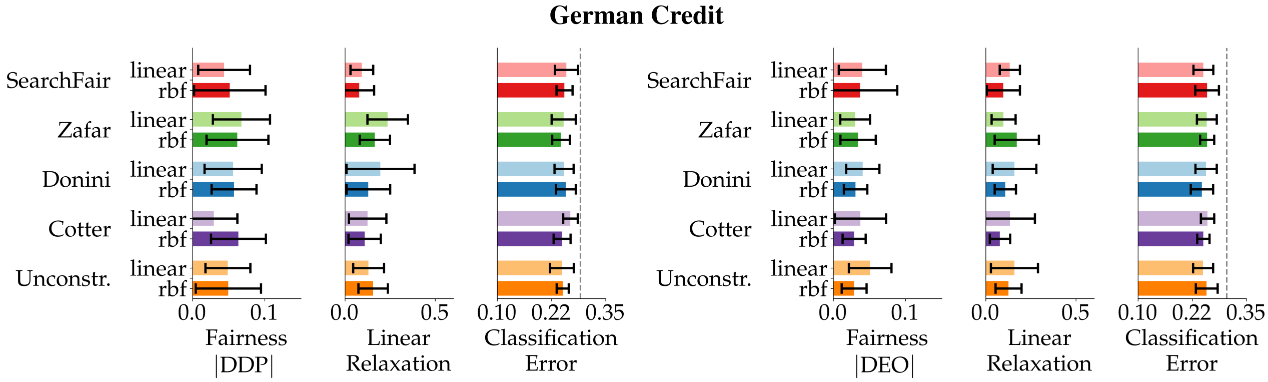
FAIRNESS NOTION	Kernel	Demographic Parity			Equality of Opportunity		
		|DDP|	|LR|	Error	|DEO|	|LR|	Error
SearchFair	linear	0.03 ± 0.01	0.03 ± 0.01	0.45 ± 0.02	0.01 ± 0.01	0.02 ± 0.03	0.45 ± 0.01
	rbf	0.07 ± 0.01	0.11 ± 0.03	0.40 ± 0.01	0.09 ± 0.05	0.17 ± 0.08	0.38 ± 0.01
Zafar	linear	0.03 ± 0.01	0.01 ± 0.01	0.42 ± 0.01	0.00 ± 0.01	0.01 ± 0.03	0.46 ± 0.01
	rbf	0.12 ± 0.03	0.24 ± 0.10	0.35 ± 0.01	0.21 ± 0.06	0.53 ± 0.21	0.33 ± 0.02
Donini	linear	0.04 ± 0.01	0.03 ± 0.02	0.42 ± 0.01	0.01 ± 0.02	0.03 ± 0.05	0.46 ± 0.02
	rbf	0.14 ± 0.05	0.29 ± 0.14	0.35 ± 0.02	0.19 ± 0.05	0.39 ± 0.18	0.33 ± 0.01
Cotter	linear	0.06 ± 0.03	0.06 ± 0.04	0.46 ± 0.05	0.01 ± 0.01	0.04 ± 0.04	0.45 ± 0.02
	rbf	0.04 ± 0.02	0.07 ± 0.04	0.40 ± 0.01	0.08 ± 0.04	0.13 ± 0.06	0.40 ± 0.01
Unconstrained	linear	0.20 ± 0.02	0.47 ± 0.04	0.34 ± 0.01	0.24 ± 0.04	0.55 ± 0.10	0.34 ± 0.01
	rbf	0.20 ± 0.02	0.39 ± 0.04	0.32 ± 0.01	0.23 ± 0.04	0.46 ± 0.07	0.33 ± 0.01
Constant	–	0.00 ± 0.00	–	0.46 ± 0.01	0.00 ± 0.00	–	0.46 ± 0.01

Figure 9. **Compas.** The grey dashed vertical line depicts the classification error of the constant classifier, which is perfectly fair for both DDP and DEO. On this dataset, Zafar and Donini with rbf kernel tend to have high values for the linear relaxation, which is probably due to an overfitting issue (as evoked at the end of Section 3.1 in the main paper). Overall, SearchFair obtains good fairness scores, comparable to Cotter. For the DDP, SearchFair is slightly worse than Cotter with an rbf kernel, but better with a linear kernel. Surprisingly, both methods also have low relaxation values which hints that, on this dataset, this relaxation might be relevant if one could avoid overfitting.



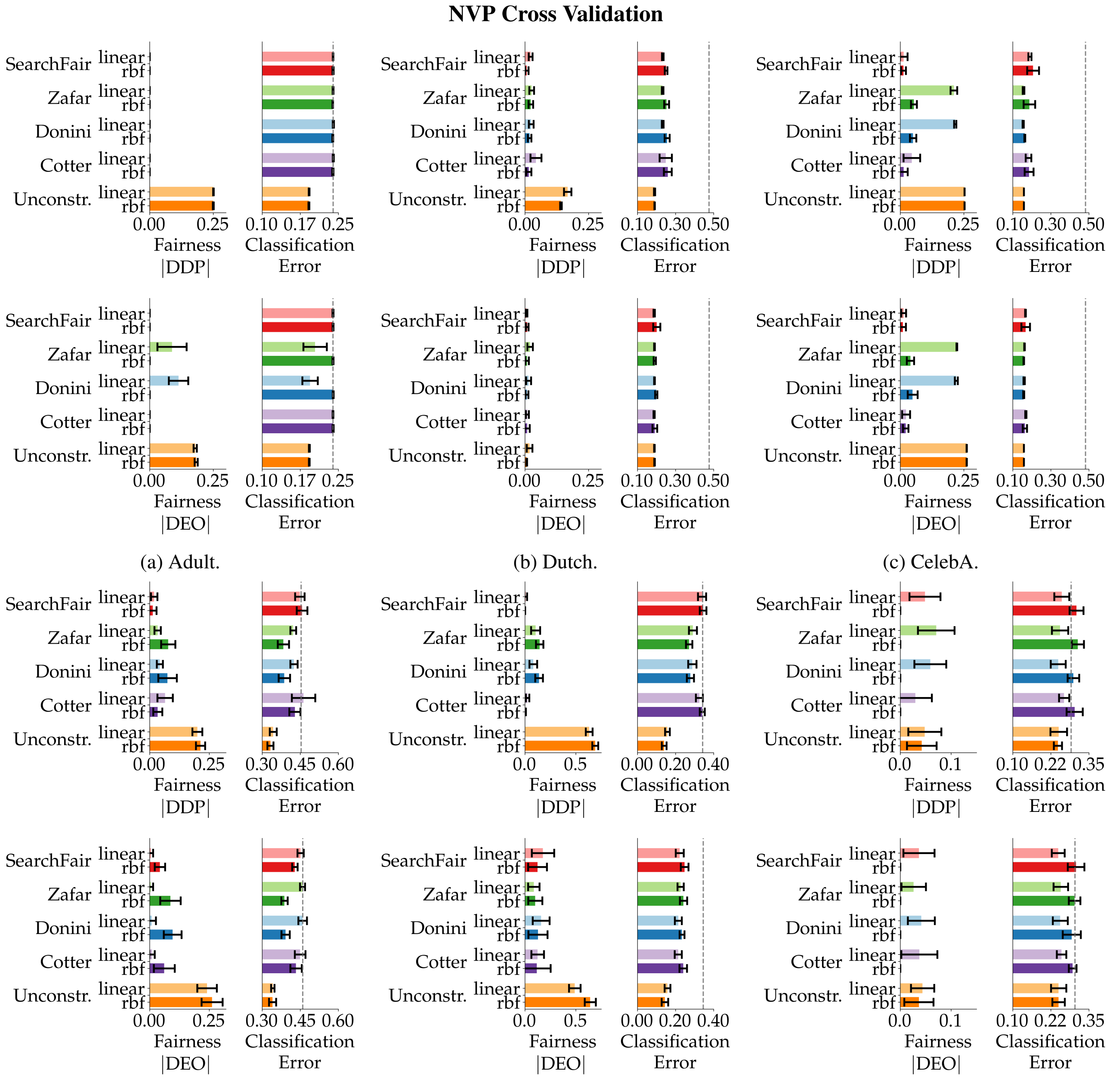
FAIRNESS NOTION		**Demographic Parity**			**Equality of Opportunity**		
	Kernel	|DDP|	|LR|	Error	|DEO|	|LR|	Error
SearchFair	linear	0.01 ± 0.01	0.07 ± 0.05	0.33 ± 0.02	0.14 ± 0.07	0.38 ± 0.17	0.22 ± 0.02
	rbf	0.02 ± 0.01	0.06 ± 0.05	0.32 ± 0.01	0.30 ± 0.13	0.73 ± 0.28	0.19 ± 0.02
Zafar	linear	0.15 ± 0.05	0.09 ± 0.07	0.27 ± 0.02	0.12 ± 0.06	0.37 ± 0.22	0.22 ± 0.01
	rbf	0.45 ± 0.06	0.22 ± 0.19	0.20 ± 0.01	0.30 ± 0.10	0.78 ± 0.25	0.17 ± 0.02
Donini	linear	0.12 ± 0.04	0.10 ± 0.07	0.28 ± 0.02	0.18 ± 0.09	0.58 ± 0.30	0.21 ± 0.02
	rbf	0.45 ± 0.05	0.42 ± 0.21	0.20 ± 0.02	0.39 ± 0.04	0.90 ± 0.14	0.16 ± 0.01
Cotter	linear	0.02 ± 0.02	0.09 ± 0.04	0.32 ± 0.02	0.12 ± 0.06	0.22 ± 0.12	0.21 ± 0.02
	rbf	0.03 ± 0.05	0.09 ± 0.05	0.32 ± 0.03	0.34 ± 0.08	0.77 ± 0.21	0.19 ± 0.02
Unconstrained	linear	0.65 ± 0.03	3.25 ± 0.28	0.16 ± 0.01	0.55 ± 0.06	2.46 ± 0.34	0.15 ± 0.01
	rbf	0.67 ± 0.04	2.64 ± 0.21	0.14 ± 0.01	0.65 ± 0.05	2.19 ± 0.24	0.14 ± 0.01
Constant	–	0.00 ± 0.00	–	0.34 ± 0.01	0.00 ± 0.00	–	0.34 ± 0.01

Figure 10. **Communities and Crime.** The grey dashed vertical line depicts the classification error of the constant classifier, which is perfectly fair for both DDP and DEO. Overall, all the fairness methods perform similarly well in terms of DEO. For DDP, only SearchFair and Cotter are able to learn a fair classifier for both the linear and rbf kernel. Once again, one can notice that a low linear relaxation might or might not imply a DDP fair classifier. Indeed, the DDP scores of Zafar, Donini, and SearchFair are very different while their linear relaxation scores are all close to 0.



FAIRNESS NOTION		Demographic Parity			Equality of Opportunity		
	Kernel	|DDP|	|LR|	Error	|DEO|	|LR|	Error
SearchFair	linear	0.04 ± 0.04	0.09 ± 0.06	0.26 ± 0.03	0.04 ± 0.03	0.13 ± 0.06	0.25 ± 0.02
	rbf	0.05 ± 0.05	0.08 ± 0.08	0.25 ± 0.02	0.04 ± 0.05	0.10 ± 0.09	0.26 ± 0.03
Zafar	linear	0.07 ± 0.04	0.24 ± 0.11	0.25 ± 0.03	0.03 ± 0.02	0.10 ± 0.07	0.26 ± 0.02
	rbf	0.06 ± 0.04	0.17 ± 0.08	0.25 ± 0.02	0.03 ± 0.02	0.17 ± 0.12	0.26 ± 0.02
Donini	linear	0.06 ± 0.04	0.20 ± 0.19	0.25 ± 0.02	0.04 ± 0.02	0.16 ± 0.12	0.26 ± 0.02
	rbf	0.06 ± 0.03	0.13 ± 0.12	0.26 ± 0.02	0.03 ± 0.02	0.11 ± 0.06	0.25 ± 0.03
Cotter	linear	0.03 ± 0.03	0.13 ± 0.10	0.27 ± 0.02	0.04 ± 0.04	0.13 ± 0.14	0.26 ± 0.02
	rbf	0.06 ± 0.04	0.11 ± 0.09	0.25 ± 0.02	0.03 ± 0.02	0.08 ± 0.06	0.25 ± 0.01
Unconstrained	linear	0.05 ± 0.03	0.13 ± 0.09	0.25 ± 0.03	0.05 ± 0.03	0.16 ± 0.13	0.25 ± 0.02
	rbf	0.05 ± 0.05	0.16 ± 0.08	0.25 ± 0.01	0.06 ± 0.03	0.13 ± 0.07	0.26 ± 0.03
Constant	–	0.00 ± 0.00	–	0.29 ± 0.02	0.00 ± 0.00	–	0.30 ± 0.02

Figure 11. **German Credit.** The grey dashed vertical line depicts the classification error of the constant classifier, which is perfectly fair for both DDP and DEO. This is the smallest dataset out of the 6 with 700 training examples and 300 test examples. This explains the large standard deviations. For this particular dataset, SearchFair does not bring any significant improvement in terms of fairness compared to the baselines. We believe that it is due to a slight overfitting issue since the dataset is so small. Nevertheless, SearchFair is not worse that the other baselines as all the methods perform comparably.



(a) Adult.

(b) Dutch.

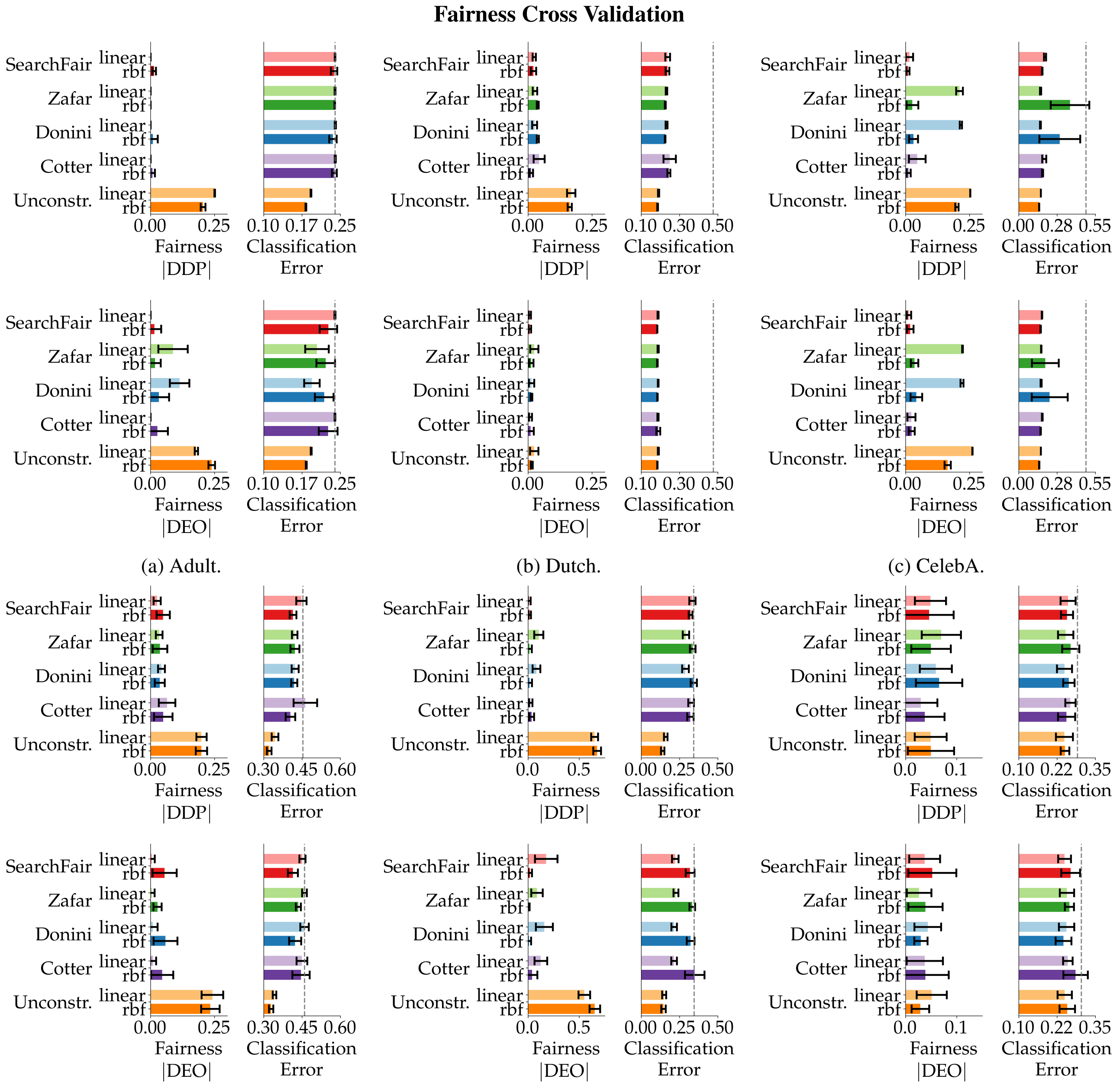
(c) CelebA.

(d) Compas.

(e) Communities and Crime.

(f) German Credit.

Figure 12. We use a procedure called NVP (Donini et al., 2018), where we choose the set of hyperparameters with the best average fairness score while having an accuracy above a given threshold. Overall, using this procedure greatly improves the performances of the fairness baselines. Hence, on most datasets, they now obtain classifiers that are as fair as the ones learned by SearchFair and Cotter. Nevertheless, there is no guarantee that the method will succeed and it indeed fails for both DDP and DEO on CelebA (linear kernel), and for DEO on Adult (linear kernel). The fact that NVP succeeds for the rbf kernel and sometimes fails for the linear kernel hints that NVP is a good way to address the complexity issue of the linear relaxations but that it does not solve the other shortcommings. The grey dashed vertical line depicts the classification error of the constant classifier, which is perfectly fair for both DDP and DEO.



(a) Adult.

(b) Dutch.

(c) CelebA.

(d) Compas.

(e) Communities and Crime.

(f) German.

Figure 13. We use another cross validation procedure, where we shortlist all hyperparameters with an absolute fairness score lower than 0.05 and, among them, choose the hyperparameters with the highest accuracy score. The results are very similar to the ones of NVP presented in Figure 12 and the same conclusions can be drawn. In particular, it seems to solve the complexity issue of linear relaxations with rbf kernel but can still fail when using the linear kernel (for both DDP and DEO on CelebA, and for DEO on Adult). The grey dashed vertical line depicts the classification error of the constant classifier, which is perfectly fair for both DDP and DEO.

C. Proofs

This section contains all the proofs and derivations that were omitted in the main paper. First, we shortly show that the intercept does not influence the value of the linear fairness relaxation. Second, we present equivalent formulations of the DDP in Section C.2. Finally, Theorem 3 is proved in Section C.3, Corollary 2 is proved in Section C.4, and Theorem 4 is proved in Section C.5.

C.1. Proof of vanishing intercept

In the main paper, we claim that for a classifier $f(x) = g(x) + b$ with b the intercept, it holds that $\mathrm{LR}_{\widehat{\mathrm{DDP}}}(f) = \mathrm{LR}_{\widehat{\mathrm{DDP}}}(g)$ for any b. First, we consider the formulation of Donini et al. (2018) with $C\left(s, \widehat{\mathcal{D}}_{\mathcal{Z}}\right) = \frac{s}{\hat{p}_s}$. Recall that $\hat{p}_s = \frac{n_s}{n}$, where n_s is the number of samples with sensitive attribute s. We need to show that $\frac{1}{n}\sum_{(x,s,y)\in\widehat{\mathcal{D}}_{\mathcal{Z}}} C\left(s, \widehat{\mathcal{D}}_{\mathcal{Z}}\right) = 0$.

$$\begin{aligned}\frac{1}{n}\sum_{(x,s,y)\in\widehat{\mathcal{D}}_{\mathcal{Z}}}\frac{s}{\hat{p}_s} &= \sum_{(x,s,y)\in\widehat{\mathcal{D}}_{\mathcal{Z}}}\frac{s}{n_s}\\ &= \sum_{(x,s=1,y)\in\widehat{\mathcal{D}}_{\mathcal{Z}}}\frac{1}{n_1} - \sum_{(x,s=-1,y)\in\widehat{\mathcal{D}}_{\mathcal{Z}}}\frac{1}{n_{-1}} = 0.\end{aligned}$$

Second, we consider Zafar et al. (2017b) with $C\left(s, \widehat{\mathcal{D}}_{\mathcal{Z}}\right) = \frac{1}{\hat{p}_1(1-\hat{p}_1)}\left(\frac{s+1}{2} - \hat{p}_1\right)$.

$$\begin{aligned}\frac{1}{n}\sum_{(x,s,y)\in\widehat{\mathcal{D}}_{\mathcal{Z}}}\frac{1}{\hat{p}_1(1-\hat{p}_1)}\left(\frac{s+1}{2}-\hat{p}_1\right) &= n\sum_{(x,s,y)\in\widehat{\mathcal{D}}_{\mathcal{Z}}}\frac{1}{n_1 n_{-1}}\left(\frac{s+1}{2}-\frac{n_1}{n}\right)\\ &= \frac{n}{n_1 n_{-1}}\left(\sum_{(x,s=1,y)\in\widehat{\mathcal{D}}_{\mathcal{Z}}}\left(1-\frac{n_1}{n}\right) - \sum_{(x,s=-1,y)\in\widehat{\mathcal{D}}_{\mathcal{Z}}}\frac{n_1}{n}\right)\\ &= \frac{n}{n_1 n_{-1}}\left(\sum_{(x,s=1,y)\in\widehat{\mathcal{D}}_{\mathcal{Z}}}\left(\frac{n_{-1}}{n}\right) - \frac{n_1 n_{-1}}{n}\right) = 0.\end{aligned}$$

C.2. Equivalent Formulations of DDP

In the main paper, we use several equivalent formulations of DDP depending on the method that we consider. We detail their derivations here. Note that these derivations are analogous for equivalent DEO formulations. First, recall the original DDP formulation:

$$\mathrm{DDP}(f) = \underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{E}}\left[\mathbb{I}_{f(x)>0}|s=1\right] - \underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{E}}\left[\mathbb{I}_{f(x)>0}|s=-1\right].$$

We can rewrite it to obtain the formulation of Zafar et al. (2017b) and Zafar et al. (2017a). Recall that $\mathcal{S} = \{-1, 1\}$ and that

$p_1 = \mathbb{P}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}\left(s=1\right)$, then:

$$\begin{aligned}
\mathrm{DDP}(f) &= \mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}\left[\mathbb{I}_{f(x)>0}|s=1\right] - \mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}\left[\mathbb{I}_{f(x)>0}|s=-1\right] \\
\Leftrightarrow \qquad &= \mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}\left[\frac{s+1}{2}\mathbb{I}_{f(x)>0}|s=1\right] - \mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}\left[\frac{1-s}{2}\mathbb{I}_{f(x)>0}|s=-1\right] \\
\Leftrightarrow \qquad &= \frac{1}{p_1}\mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}\left[\frac{s+1}{2}\mathbb{I}_{f(x)>0}\right] - \frac{1}{1-p_1}\mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}\left[\frac{1-s}{2}\mathbb{I}_{f(x)>0}\right] \\
\Leftrightarrow \qquad &= \mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}\left[\left(\frac{s+1}{2p_1} - \frac{1-s}{2(1-p_1)}\right)\mathbb{I}_{f(x)>0}\right] \\
\Leftrightarrow \qquad &= \mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}\left[\left(\frac{(s+1)(1-p_1)-(1-s)p_1}{2p_1(1-p_1)}\right)\mathbb{I}_{f(x)>0}\right] \\
\Leftrightarrow \qquad &= \mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}\left[\left(\frac{s+1-2p_1}{2p_1(1-p_1)}\right)\mathbb{I}_{f(x)>0}\right] \\
\Leftrightarrow \qquad &= \mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}\left[\frac{1}{p_1(1-p_1)}\left(\frac{s+1}{2} - p_1\right)\mathbb{I}_{f(x)>0}\right].
\end{aligned}$$

We can also rewrite it to obtain the formulation of Donini et al. (2018). Recall that $p_s = \mathbb{P}_{(x',s',y')\sim\mathcal{D}_\mathcal{Z}}\left(s'=s\right)$, then:

$$\begin{aligned}
\mathrm{DDP}(f) &= \mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}\left[\mathbb{I}_{f(x)>0}|s=1\right] - \mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}\left[\mathbb{I}_{f(x)>0}|s=-1\right] \\
\Leftrightarrow \qquad &= \mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}\left[s\mathbb{I}_{f(x)>0}|s=1\right] + \mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}\left[s\mathbb{I}_{f(x)>0}|s=-1\right] \\
\Leftrightarrow \qquad &= \mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}\left[s\mathbb{I}_{f(x)>0}|s=1\right]\frac{p_1}{p_1} + \mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}\left[s\mathbb{I}_{f(x)>0}|s=-1\right]\frac{1-p_1}{1-p_1} \\
\Leftrightarrow \qquad &= \mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}\left[\frac{s}{p_s}\mathbb{I}_{f(x)>0}|s=1\right]p_1 + \mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}\left[\frac{s}{p_s}\mathbb{I}_{f(x)>0}|s=-1\right](1-p_1) \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Law of total expectation.)} \\
\Leftrightarrow \qquad &= \mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}\left[\frac{s}{p_s}\mathbb{I}_{f(x)>0}\right].
\end{aligned}$$

Finally, we can rewrite it to obtain the formulation of Wu et al. (2019) as follows:

$$\begin{aligned}
\mathrm{DDP}(f) &= \mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}\left[\mathbb{I}_{f(x)>0}|s=1\right] - \mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}\left[\mathbb{I}_{f(x)>0}|s=-1\right] \\
\Leftrightarrow \qquad &= \mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}\left[\frac{s}{p_s}\mathbb{I}_{f(x)>0}\right] \qquad\qquad \text{(Formulation of Donini et al. (2018).)} \\
\Leftrightarrow \qquad &= \mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}\left[\frac{\mathbb{I}_{s=1}}{p_1}\mathbb{I}_{f(x)>0} - \frac{\mathbb{I}_{s=-1}}{1-p_1}\mathbb{I}_{f(x)>0}\right] \\
\Leftrightarrow \qquad &= \mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}\left[\frac{\mathbb{I}_{s=1}}{p_1}\mathbb{I}_{f(x)>0} + \frac{\mathbb{I}_{s=-1}}{1-p_1}\mathbb{I}_{f(x)<0} - 1\right].
\end{aligned}$$

C.3. Proof of Theorem 1 and Continuity of DEO$\left(f^{\beta}_{\widehat{\mathcal{D}}_\mathcal{Z}}(\lambda)\right)$

Recall that our main optimization problem is:

$$f^{\beta}_{\widehat{\mathcal{D}}_\mathcal{Z}}(\lambda) = \mathop{\arg\min}_{f\in\mathcal{F}} \hat{L}(f) + \lambda \mathrm{R}_{\widehat{\mathrm{DDP}}}(f) + \beta\Omega(f)\,. \tag{7}$$

To prove Theorem 1, that is to show the continuity of the function $\lambda \mapsto \mathrm{DDP}\left(f^{\beta}_{\widehat{\mathcal{D}}_\mathcal{Z}}(\lambda)\right)$, we need technical Lemmas 1 and 2. The first one shows that the function $\lambda \mapsto f^{\beta}_{\widehat{\mathcal{D}}_\mathcal{Z}}(\lambda)$ is continuous. The second one shows that for particular function classes, $f \mapsto \mathbb{P}_{x\sim\mathcal{D}_\mathcal{X}}\left[f(x) \leq 0\right]$ is a continuous function. Before proving them, we first recall the definition of a m-strongly convex function.

Definition 2 (m-strongly convex functions). *A function $f : X \mapsto \mathbb{R}$ is called m-strongly convex with parameter $m > 0$ if for all $x, y \in X$ and $t \in [0, 1]$*

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) - \frac{m}{2}t(1-t)\left\|x - y\right\|_2^2 .$$

We can now prove our two technical lemmas.

Lemma 1 (Continuity of $\lambda \mapsto f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda)$). *Assume that Optimization Problem 7 is m-strongly convex and that $R_{\widehat{DDP}}(f)$ is bounded in the interval $[-B, B]$. Given a training set $\widehat{\mathcal{D}}_{\mathcal{Z}}$ and a regularization parameter $\beta > 0$, the function:*

$$\lambda \mapsto f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda)$$

is continuous and there exists a constant $C = \sqrt{\frac{8B}{m}}$ such that:

$$\left\| f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda) - f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda') \right\|_{\mathcal{F}} \leq C\sqrt{|\lambda - \lambda'|}.$$

Proof. Let $g^{\lambda}(f) = \hat{L}(f) + \lambda \mathrm{R}_{\widehat{\mathrm{DDP}}}(f) + \beta\Omega(f)$ and $g^{\lambda'}(f) = g^{\lambda}(f) + \varepsilon \mathrm{R}_{\widehat{\mathrm{DDP}}}(f)$ with $\varepsilon > 0$ and $\varepsilon = \lambda' - \lambda$. For the sake of readability, for the remainder of the proof, we write $f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda)$ as $f(\lambda)$. Since Optimization Problem 7 is m-strongly convex, it holds that:

$$\begin{aligned} g^{\lambda}(tf(\lambda) + (1-t)f(\lambda')) &+ \varepsilon \mathrm{R}_{\widehat{\mathrm{DDP}}}(tf(\lambda) + (1-t)f(\lambda')) \\ &\leq tg^{\lambda}(f(\lambda)) + (1-t)g^{\lambda}(f(\lambda')) + t\varepsilon \mathrm{R}_{\widehat{\mathrm{DDP}}}(f(\lambda)) + (1-t)\varepsilon \mathrm{R}_{\widehat{\mathrm{DDP}}}(f(\lambda')) \\ &\quad - \frac{m}{2}t(1-t)\left\|f(\lambda) - f(\lambda')\right\|_{\mathcal{F}}^2 . \end{aligned}$$

In particular, for $t = \frac{1}{2}$:

$$\begin{aligned} & \frac{m}{8}\left\|f(\lambda) - f(\lambda')\right\|_{\mathcal{F}}^2 \leq \frac{1}{2}g^{\lambda}(f(\lambda)) + \frac{1}{2}g^{\lambda}(f(\lambda')) + \frac{1}{2}\varepsilon \mathrm{R}_{\widehat{\mathrm{DDP}}}(f(\lambda)) + \frac{1}{2}\varepsilon \mathrm{R}_{\widehat{\mathrm{DDP}}}(f(\lambda')) \\ & \qquad - g^{\lambda}\left(\frac{1}{2}f(\lambda) + \frac{1}{2}f(\lambda')\right) - \varepsilon \mathrm{R}_{\widehat{\mathrm{DDP}}}\left(\frac{1}{2}f(\lambda) + \frac{1}{2}f(\lambda')\right) \\ \Leftrightarrow \quad & \frac{m}{8}\left\|f(\lambda) - f(\lambda')\right\|_{\mathcal{F}}^2 \leq \frac{1}{2}g^{\lambda}(f(\lambda)) - \frac{1}{2}g^{\lambda}\left(\frac{1}{2}f(\lambda) + \frac{1}{2}f(\lambda')\right) \\ & \qquad + \frac{1}{2}g^{\lambda}(f(\lambda')) + \frac{1}{2}\varepsilon \mathrm{R}_{\widehat{\mathrm{DDP}}}(f(\lambda')) - \frac{1}{2}g^{\lambda}\left(\frac{1}{2}f(\lambda) + \frac{1}{2}f(\lambda')\right) \\ & \qquad - \frac{1}{2}\varepsilon \mathrm{R}_{\widehat{\mathrm{DDP}}}\left(\frac{1}{2}f(\lambda) + \frac{1}{2}f(\lambda')\right) \\ & \qquad + \frac{1}{2}\varepsilon \mathrm{R}_{\widehat{\mathrm{DDP}}}(f(\lambda)) - \frac{1}{2}\varepsilon \mathrm{R}_{\widehat{\mathrm{DDP}}}\left(\frac{1}{2}f(\lambda) + \frac{1}{2}f(\lambda')\right) \\ \Leftrightarrow \quad & \frac{m}{8}\left\|f(\lambda) - f(\lambda')\right\|_{\mathcal{F}}^2 \leq \frac{1}{2}g^{\lambda}(f(\lambda)) - \frac{1}{2}g^{\lambda}\left(\frac{1}{2}f(\lambda) + \frac{1}{2}f(\lambda')\right) \\ & \qquad + \frac{1}{2}g^{\lambda'}(f(\lambda')) - \frac{1}{2}g^{\lambda'}\left(\frac{1}{2}f(\lambda) + \frac{1}{2}f(\lambda')\right) \\ & \qquad + \frac{1}{2}\varepsilon \mathrm{R}_{\widehat{\mathrm{DDP}}}(f(\lambda)) - \frac{1}{2}\varepsilon \mathrm{R}_{\widehat{\mathrm{DDP}}}\left(\frac{1}{2}f(\lambda) + \frac{1}{2}f(\lambda')\right) . \end{aligned}$$

Since $f(\lambda)$ and $f(\lambda')$ respectively minimize $g^{\lambda}(f)$ and $g^{\lambda'}(f)$, it holds that

$$\begin{aligned} \frac{1}{2}g^{\lambda}(f(\lambda)) - \frac{1}{2}g^{\lambda}\left(\frac{1}{2}f(\lambda) + \frac{1}{2}f(\lambda')\right) &\leq 0 \\ \frac{1}{2}g^{\lambda'}(f(\lambda')) - \frac{1}{2}g^{\lambda'}\left(\frac{1}{2}f(\lambda) + \frac{1}{2}f(\lambda')\right) &\leq 0 \end{aligned}$$

which, in turns, implies

$$
\begin{aligned}
& & \frac{m}{8}\left\|f(\lambda)-f(\lambda')\right\|_{\mathcal{F}}^{2} &\leq \frac{1}{2}\varepsilon \mathrm{R}_{\widehat{\mathrm{DDP}}}(f(\lambda)) - \frac{1}{2}\varepsilon \mathrm{R}_{\widehat{\mathrm{DDP}}}\left(\frac{1}{2}f(\lambda)+\frac{1}{2}f(\lambda')\right) \\
&\Leftrightarrow & \left\|f(\lambda)-f(\lambda')\right\|_{\mathcal{F}}^{2} &\leq \frac{8}{2m}\varepsilon \mathrm{R}_{\widehat{\mathrm{DDP}}}(f(\lambda)) - \frac{8}{2m}\varepsilon \mathrm{R}_{\widehat{\mathrm{DDP}}}\left(\frac{1}{2}f(\lambda)+\frac{1}{2}f(\lambda')\right) \\
& & & \qquad\qquad\qquad\qquad (\mathrm{R}_{\widehat{\mathrm{DDP}}}(f) \in [-B,B]) \\
&\Leftrightarrow & \left\|f(\lambda)-f(\lambda')\right\|_{\mathcal{F}}^{2} &\leq \frac{8B}{m}\varepsilon \\
& & & \qquad\qquad\qquad\qquad (\varepsilon \leq |\lambda'-\lambda|) \\
&\Rightarrow & \left\|f(\lambda)-f(\lambda')\right\|_{\mathcal{F}}^{2} &\leq \frac{8B}{m}\left|\lambda'-\lambda\right| \\
&\Rightarrow & \left\|f(\lambda)-f(\lambda')\right\|_{\mathcal{F}} &\leq \sqrt{\frac{8B}{m}}\sqrt{|\lambda'-\lambda|}.
\end{aligned}
$$

Choosing $C=\sqrt{\frac{8B}{m}}$ concludes the proof. □

Lemma 2 (Continuity of $f \mapsto \mathbb{P}_{x\sim\mathcal{D}_{\mathcal{X}}}[f(x)\leq 0]$). *Let $\mathcal{F}$ be a space of real valued functions $f:\mathcal{X}\to\mathbb{R}$. Assume that the following conditions hold:*
(i) there exists a metric ρ such that $(\mathcal{F},\rho)$ is a metric space,
(ii) $\forall x\in\mathcal{X}$, the function $g(f): f\mapsto f(x)$ is continuous,
(iii) $\forall f\in\mathcal{F}$, f is Lebesgue measurable and the set $\{x : x\in\mathcal{X}, f(x)=0\}$ is a Lebesgue null set,
(iv) the probability density functions $f_{\mathcal{X}}$ is Lebesgue-measurable.
We have that:

$$\mathop{\mathbb{P}}_{x\sim\mathcal{D}_{\mathcal{X}}}[f(x)\leq 0]$$

is a continuous function in $f\in\mathcal{F}$.

Proof. We have that:

$$\mathop{\mathbb{P}}_{x\sim\mathcal{D}_{\mathcal{X}}}[f(x)\leq 0] = \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathcal{X}}}\left[\mathbb{I}_{f(x)\leq 0}\right] = \int_{\mathcal{X}} \mathbb{I}_{f(x)\leq 0} f_{\mathcal{X}}(x)\,dx = \int_{\mathcal{X}} h(f,x)\,dx.$$

To show that this function is continuous, we apply Theorem 5.6 in Elstrodt (1996). To this extent, we need to show that all the conditions hold.

- **Condition a:** $\forall f\in\mathcal{F}$, $h(f,\cdot)\in\mathcal{L}^1$.
  The function $f(x)\mapsto\mathbb{I}_{f(x)\leq 0}$ is Borel measurable and the function f is Lebesgue measurable. By composition, the function $x\mapsto\mathbb{I}_{f(x)\leq 0}$ is also Lebesgue measurable. As the product of two Lebesgue measurable functions, h is also Lebesgue measurable. Furthermore, we have:

  $$\int_{\mathcal{X}} |h(f,x)|\,dx \leq \int_{\mathcal{X}} f_{\mathcal{X}}(x)\,dx = 1 < \infty$$

  which is the desired condition.

- **Condition b:** $h(\cdot,x)$ is continuous in $f_0\in\mathcal{F}$ for μ-almost all $x\in\mathcal{X}$.
  Since $\forall x\in\mathcal{X}, g(f): f\mapsto f(x)$ is continuous in f_0, $\mathbb{I}_{f(x)\leq 0}$ is also a continuous function in f_0 expect for the set $\{x : x\in\mathcal{X}, f(x)=0\}$ which is a Lebesgue null set.

- **Condition c:** There exists a neighbourhood U of f_0 and an integrable function $u:\mathcal{X}\to[0,\infty)$ such that $\forall f\in U$ we have $h(f,\cdot)\leq u$ μ-a.e..
  Taking $u=f_{\mathcal{X}}$ satisfy the condition with $U=\mathcal{F}$.

Since all the conditions hold, we have that $\mathbb{P}_{x\sim\mathcal{D}_\mathcal{X}}\left[f(x)\leq 0\right]$ is continuous at f_0. Furthermore, given our assumptions on $\mathcal{F}$, this remains true $\forall f_0\in\mathcal{F}$. This concludes the proof. □

We are now ready to prove Theorem 3.

Theorem 3 (Continuity of DDP$\left(f^{\beta}_{\widehat{\mathcal{D}}_\mathcal{Z}}(\lambda)\right)$). *Let $\mathcal{F}$ be a function space, we define the set of learnable functions as $\mathcal{F}_\Lambda=\left\{f\in\mathcal{F}:\exists\lambda\geq 0, f=f^{\beta}_{\widehat{\mathcal{D}}_\mathcal{Z}}(\lambda)\right\}$. Assume that the following conditions hold:*
(i) *Optimization Problem 7 is m-strongly convex in f,*
(ii) *for all $f\in\mathcal{F}$, $R_{\widehat{DDP}}(f)$ is bounded in the interval $[-B,B]$,*
(iii) *there exists a metric ρ such that $(\mathcal{F}_\Lambda,\rho)$ is a metric space,*
(iv) *$\forall x\in\mathcal{X}$, the function $g(f):f\mapsto f(x)$ is continuous,*
(v) *$\forall f\in\mathcal{F}_\Lambda$, f is Lebesgue measurable and the sets $\{x:(x,s,y)\in\mathcal{Z},s=1,f(x)=0\}$ and $\{x:(x,s,y)\in\mathcal{Z},s=-1,f(x)=0\}$ are Lebesgue null sets,*
(vi) *the probability density functions $f_{\mathcal{Z}|s=1}$ and $f_{\mathcal{Z}|s=-1}$ are Lebesgue-measurable.*
Then, the function $\lambda\mapsto DDP\left(f^{\beta}_{\widehat{\mathcal{D}}_\mathcal{Z}}(\lambda)\right)$ is continuous.

Proof. Recall that DDP is defined as follows:

$$\mathrm{DDP}(f)=\underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}|s=1}}{\mathbb{P}}\left[f(x)>0\right]-\underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}|s=-1}}{\mathbb{P}}\left[f(x)>0\right].$$

Applying Lemma 2, we have that $c:\mathcal{F}_\Lambda\to\mathbb{R}$ defined as $c(f)=\mathbb{P}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}|s=1}}\left[f(x)>0\right]$ and $c':\mathcal{F}_\Lambda\to\mathbb{R}$ defined as $c'(f)=\mathbb{P}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}|s=-1}}\left[f(x)>0\right]$ are continuous functions. It implies that the function $q:\mathcal{F}_\Lambda\to\mathbb{R}$ defined as $q(f)=\mathrm{DDP}(f)$ is continuous.

Then, using Lemma 1 and recalling that the composition of two continuous functions is also continuous gives the theorem. □

We use the same proof technique to prove the continuity of DEO as stated in the next theorem. The main differences are in conditions (v) and (vi) where we only need to consider the positively labelled examples.

Theorem 1.1 (Continuity of DEO$\left(f^{\beta}_{\widehat{\mathcal{D}}_\mathcal{Z}}(\lambda)\right)$). *Let $\mathcal{F}$ be a function space, we define the set of learnable functions as $\mathcal{F}_\Lambda=\left\{f\in\mathcal{F}:\exists\lambda\geq 0, f=f^{\beta}_{\widehat{\mathcal{D}}_\mathcal{Z}}(\lambda)\right\}$. Assume that the following conditions hold:*
(i) *Optimization Problem 7 is m-strongly convex in f,*
(ii) *for all $f\in\mathcal{F}$, $R_{\widehat{DDP}}(f)$ is bounded in the interval $[-B,B]$,*
(iii) *there exists a metric ρ such that $(\mathcal{F}_\Lambda,\rho)$ is a metric space,*
(iv) *$\forall x\in\mathcal{X}$, the function $g(f):f\mapsto f(x)$ is continuous,*
(v) *$\forall f\in\mathcal{F}_\Lambda$, f is Lebesgue measurable and the sets $\{x:(x,s,y)\in\mathcal{Z},y=1,s=1,f(x)=0\}$ and $\{x:(x,s,y)\in\mathcal{Z},y=1,s=-1,f(x)=0\}$ are Lebesgue null sets,*
(vi) *the probability density functions $f_{\mathcal{Z}|y=1,s=1}$ and $f_{\mathcal{Z}|y=1,s=-1}$ are Lebesgue-measurable.*
Then the function $\lambda\mapsto DEO\left(f^{\beta}_{\widehat{\mathcal{D}}_\mathcal{Z}}(\lambda)\right)$ is continuous.

Proof. Recall that DEO is defined as follows:

$$\mathrm{DEO}(f)=\underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}|y=1,s=1}}{\mathbb{P}}\left[f(x)>0\right]-\underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}|y=1,s=-1}}{\mathbb{P}}\left[f(x)>0\right].$$

Applying Lemma 2, we have that $c:\mathcal{F}_\Lambda\to\mathbb{R}$ defined as $c(f)=\mathbb{P}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}|y=1,s=1}}\left[f(x)>0\right]$ and $c':\mathcal{F}_\Lambda\to\mathbb{R}$ defined as $c'(f)=\mathbb{P}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}|y=1,s=-1}}\left[f(x)>0\right]$ are continuous functions. It implies that the function $q:\mathcal{F}_\Lambda\to\mathbb{R}$ defined as $q(f)=\mathrm{DEO}(f)$ is continuous.

Then, using Lemma 1 and recalling that the composition of two continuous functions is also continuous gives the theorem. □

C.4. Proof of Corollary 2 and Existence of a DEO-fair classifier

Corollary 2 (Existence of a DDP-fair classifier). *Let $\mathcal{F}$ be a function space, we define the set of learnable functions as $\mathcal{F}_\Lambda = \left\{f \in \mathcal{F} : \exists \lambda \geq 0, f = f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda)\right\}$. Assume that Theorem 3 holds and that there exist two hyperparameters λ_+ and λ_- such that DDP$\left(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_+)\right) > 0$ and DDP$\left(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_-)\right) < 0$. Then, there exists at least one value $\lambda_0 \in [\min(\lambda_+, \lambda_-), \max(\lambda_+, \lambda_-)]$ such that DDP$\left(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_0)\right) = 0$.*

Proof. This corollary is a direct consequence of the intermediate value theorem and the continuity of DDP proven in Theorem 3. □

Note that one can obtain the same result for DEO.

Corollary 1.1 (Existence of a DEO-fair classifier). *Let $\mathcal{F}$ be a function space, we define the set of learnable functions as $\mathcal{F}_\Lambda = \left\{f \in \mathcal{F} : \exists \lambda \geq 0, f = f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda)\right\}$. Assume that Theorem 3 holds and that there exist two hyperparameters λ_+ and λ_- such that DEO$\left(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_+)\right) > 0$ and DEO$\left(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_-)\right) < 0$. Then, there exists at least one value $\lambda_0 \in [\min(\lambda_+, \lambda_-), \max(\lambda_+, \lambda_-)]$ such that DEO$\left(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_0)\right) = 0$.*

Proof. This corollary is a direct consequence of the intermediate value theorem and the continuity of DEO proven in Theorem 1.1. □

C.5. Proof of Theorem 4

Recall the definition of a good similarity for fairness.

Definition 3 (Good Similarities for Fairness). *A similarity function K is $(\varepsilon, \gamma, \tau)$-good for convex, positive, and decreasing loss ℓ and (μ, ν)-fair for demographic parity if there exists a (random) indicator function $R(x, s, y)$ defining a (probabilistic) set of "reasonable points" such that, given that $\forall x \in \mathcal{X}, g(x) = \mathbb{E}_{(x',s',y') \sim \mathcal{D}_{\mathcal{Z}}}[y' K(x, x') \,|R(x', s', y')]$, the following conditions hold:*

(i) $\underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{E}}\left[\ell\left(\frac{yg(x)}{\gamma}\right)\right] \leq \varepsilon$,

(ii) $\left|\underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}|s=1}}{\mathbb{P}}[g(x) \geq \gamma] - \underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}|s=-1}}{\mathbb{P}}[g(x) \geq \gamma]\right| \leq \mu$,

(iii) $\underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}[|g(x)| \geq \gamma] \geq 1 - \nu$,

(iv) $\underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}[R(x, s, y)] \geq \tau$.

In the following theorem we prove, given a good and fair similarity, the existence of an accurate and fair classifier.

Theorem 4 (Existence of a fair and accurate separator). *Let $K \in [-1, 1]$ be a $(\varepsilon, \gamma, \tau)$-good and (μ, ν)-fair metric for a given convex, positive and decreasing loss ℓ with lipschitz constant L. For any $\varepsilon_1 > 0$ and $0 < \delta < \frac{\gamma\varepsilon_1}{2(L+\ell(0))}$, let $S = \{(x'_1, s'_1, y'_1), \ldots, (x'_d, s'_d, y'_d)\}$ be a set of d examples drawn from $\mathcal{D}_{\mathcal{Z}}$ with*

$$d \geq \frac{1}{\tau}\left[\frac{L^2}{\gamma^2\varepsilon_1^2} + \frac{3}{\delta} + \frac{4L}{\delta\gamma\varepsilon_1}\sqrt{\delta(1-\tau)\log(2/\delta)}\right].$$

Let $\phi^S : \mathcal{X} \to \mathbb{R}^d$ be a mapping defined as $\phi^S_i(x) = K(x, x'_i)$, for all $i \in \{1, \ldots, d\}$. Then, with probability at least $1 - \frac{5}{2}\delta$ over the choice of S, the induced distribution over $\phi^S(\mathcal{X}) \times \mathcal{S} \times \mathcal{Y}$ has a linear separator α such that

$$\underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{E}}\left[\ell\left(\frac{y\left\langle\alpha, \phi^S(x)\right\rangle}{\gamma}\right)\right] \leq \varepsilon + \varepsilon_1.$$

and, with $p_1 = \mathbb{P}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}[s = 1]$,

$$|DDP(\alpha)| \leq \mu + (\nu + 2\delta)\max\left(\frac{1}{p_1}, \frac{1}{1-p_1}\right).$$

Proof. Let $S = \{(x'_1, s'_1, y'_1), \dots, (x'_d, s'_d, y'_d)\}$ be a sample of size d drawn from $\mathcal{D}_\mathcal{Z}$ and let $\phi^S : \mathcal{X} \to \mathbb{R}^d$ be a mapping defined as $\phi^S_i(x) = K(x, x'_i)$, for all $i \in \{1, \dots, d\}$. Recall that $|K(x,x)| \leq 1$ for all x. It implies that $\left\|\phi^S\right\|_\infty \leq 1$. Furthermore, let $\alpha \in \mathbb{R}^d$ be defined as $\alpha_i = \frac{y'_i R\left(x'_i, s'_i, y'_i\right)}{d_1}$ with $d_1 = \sum_i R(x'_i, s'_i, y'_i)$ which ensures that $\|\alpha\|_1 = 1$.

The proof is in two parts. First, we show the bound on the target criterion, that is, given d chosen as in the theorem, we show that

$$\mathbb{P}_{S \sim \mathcal{D}^d_\mathcal{Z}} \left[\mathbb{E}_{(x,s,y) \sim \mathcal{D}_\mathcal{Z}} \left[\ell\left(\frac{y \left\langle \alpha, \phi^S(x) \right\rangle}{\gamma} \right) \right] \leq \varepsilon + \varepsilon_1 \right] \geq 1 - \delta.$$

Second, we show a bound on the true DDP, that is, given d chosen as in the theorem, we show that

$$|\text{DDP}(\alpha)| \leq \mu + \nu \max\left(\frac{1}{p_1}, \frac{1}{1 - p_1} \right)$$

where $p_1 = \mathbb{P}_{(x,s,y) \sim \mathcal{D}_\mathcal{Z}} [s = 1]$.

Bound on the target criterion. For any example $(x, s, y) \sim \mathcal{D}_\mathcal{Z}$, we have

$$y \left\langle \alpha, \phi^S(x) \right\rangle = \frac{\sum_{i=1}^d y y'_i R(x'_i, s'_i, y'_i)\, K(x, x'_i)}{d_1}$$

which is an empirical average of d_1 terms with $R(x'_i, s'_i, y'_i) = 1$ and

$$-1 \leq y y'_i R(x'_i, s'_i, y'_i)\, K(x, x'_i) \leq 1.$$

Using Hoeffding's inequality, we can show that

$$\mathbb{P}_{S \sim \mathcal{D}^d_\mathcal{Z}} \left[y \left\langle \alpha, \phi^S(x) \right\rangle \leq \mathbb{E}_{(x',s',y') \sim \mathcal{D}_\mathcal{Z}} \left[y y' K(x, x') \,|R(x', s', y') \right] - t \right] \leq \exp\left(-\frac{t^2 d_1}{2} \right)$$

which implies that, with probability at least $1 - \frac{\delta^2}{4}$, we have

$$y \left\langle \alpha, \phi^S(x) \right\rangle \geq \mathbb{E}_{(x',s',y') \sim \mathcal{D}_\mathcal{Z}} \left[y y' K(x, x') \,|R(x', s', y') \right] - \sqrt{\frac{2 \log\left(\frac{4}{\delta^2}\right)}{d_1}}.$$

This inequality holds for any $(x, s, y) \sim \mathcal{D}_\mathcal{Z}$ and thus we have that

$$\begin{aligned}
& \mathbb{P}_{S \sim \mathcal{D}^d_\mathcal{Z}} \left[y \left\langle \alpha, \phi^S(x) \right\rangle \leq \mathbb{E}_{(x',s',y') \sim \mathcal{D}_\mathcal{Z}} \left[y y' K(x, x') \,|R(x', s', y') \right] - \sqrt{\frac{2 \log\left(\frac{4}{\delta^2}\right)}{d_1}} \right] \leq \frac{\delta^2}{4} \\
\Rightarrow \quad & \mathbb{E}_{(x,s,y) \sim \mathcal{D}_\mathcal{Z}} \left[\mathbb{P}_{S \sim \mathcal{D}^d_\mathcal{Z}} \left[y \left\langle \alpha, \phi^S(x) \right\rangle \leq \mathbb{E}_{(x',s',y') \sim \mathcal{D}_\mathcal{Z}} \left[y y' K(x, x') \,|R(x', s', y') \right] - \sqrt{\frac{2 \log\left(\frac{4}{\delta^2}\right)}{d_1}} \right] \right] \leq \frac{\delta^2}{4} \\
\Rightarrow \quad & \mathbb{E}_{S \sim \mathcal{D}^d_\mathcal{Z}} \left[\mathbb{P}_{(x,s,y) \sim \mathcal{D}_\mathcal{Z}} \left[y \left\langle \alpha, \phi^S(x) \right\rangle \leq \mathbb{E}_{(x',s',y') \sim \mathcal{D}_\mathcal{Z}} \left[y y' K(x, x') \,|R(x', s', y') \right] - \sqrt{\frac{2 \log\left(\frac{4}{\delta^2}\right)}{d_1}} \right] \right] \leq \frac{\delta^2}{4}.
\end{aligned}$$

Then, applying Markov's inequality, we obtain that

$$\mathbb{P}_{S \sim \mathcal{D}^d_\mathcal{Z}} \left[\mathbb{P}_{(x,s,y) \sim \mathcal{D}_\mathcal{Z}} \left[y \left\langle \alpha, \phi^S(x) \right\rangle \leq \mathbb{E}_{(x',s',y') \sim \mathcal{D}_\mathcal{Z}} \left[y y' K(x, x') \,|R(x', s', y') \right] - \sqrt{\frac{2 \log\left(\frac{4}{\delta^2}\right)}{d_1}} \right] \geq \delta \right] \leq \frac{\delta}{4},$$

which implies

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}_{\mathcal{Z}}^{d}}\left[\mathop{\mathbb{P}}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}\left[y\left\langle\alpha,\phi^{S}(x)\right\rangle\leq\mathop{\mathbb{E}}_{(x',s',y')\sim\mathcal{D}_{\mathcal{Z}}}\left[yy'K(x,x')\,|R(x',s',y')\right]-\sqrt{\frac{2\log\left(\frac{4}{\delta^{2}}\right)}{d_{1}}}\right]\leq\delta\right]\geq1-\frac{\delta}{4}.$$

In other words, with a probability at least $1-\frac{\delta}{4}$ at most δ fraction of points violate

$$y\left\langle\alpha,\phi^{S}(x)\right\rangle\geq\mathop{\mathbb{E}}_{(x',s',y')\sim\mathcal{D}_{\mathcal{Z}}}\left[yy'K(x,x')\,|R(x',s',y')\right]-\sqrt{\frac{2\log\left(\frac{4}{\delta^{2}}\right)}{d_{1}}}.\tag{8}$$

Therefore, let $g(x)=\mathbb{E}_{(x',s',y')\sim\mathcal{D}_{\mathcal{Z}}}\left[y'K(x,x')\,|R(x',s',y')\right]$, with a probability at least $1-\frac{\delta}{4}$ for at least $1-\delta$ fraction of points, which do not violate (8), we have, for our decreasing loss ℓ (an example of decreasing loss is the hinge loss, $\ell(w)=\max\left(0,1-w\right)$):

$$\ell\left(\frac{y\left\langle\alpha,\phi^{S}(x)\right\rangle}{\gamma}\right)\leq\ell\left(\frac{yg(x)-\sqrt{\frac{2\log\left(\frac{4}{\delta^{2}}\right)}{d_{1}}}}{\gamma}\right)$$

(A convex loss is L-lipschitz continuous on any closed sub-interval.)

$$\leq\ell\left(\frac{yg(x)}{\gamma}\right)+L\left|\frac{1}{\gamma}\sqrt{\frac{2\log\left(\frac{4}{\delta^{2}}\right)}{d_{1}}}\right|$$

$$\leq\ell\left(\frac{yg(x)}{\gamma}\right)+\frac{L}{\gamma}\sqrt{\frac{2\log\left(\frac{4}{\delta^{2}}\right)}{d_{1}}}.$$

For at most a δ fraction of points violating (8), we use a bound on the worst case loss derived from its lipschitzness.

$$\ell\left(\frac{y\left\langle\alpha,\phi^{S}(x)\right\rangle}{\gamma}\right)\leq L\left|\frac{y\left\langle\alpha,\phi^{S}(x)\right\rangle}{\gamma}\right|+\ell(0)$$

$$\leq L\max_{x}\frac{\left|\left\langle\alpha,\phi^{S}(x)\right\rangle\right|}{\gamma}+\ell(0)$$

(Cauchy-Schwarz Inequality.)

$$\leq L\max_{x}\frac{\left\|\alpha\right\|_{1}\left\|\phi^{S}(x)\right\|_{\infty}}{\gamma}+\ell(0)$$

$$\leq\ell(0)+\frac{L}{\gamma}$$

($\gamma\leq1$.)

$$\leq\frac{L+\ell(0)}{\gamma}.$$

Altogether, we obtain with a probability of at least $1-\frac{\delta}{4}$ over S that

$$\begin{aligned}\mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}\left[\ell\left(\frac{y\left\langle\alpha,\phi^S(x)\right\rangle}{\gamma}\right)\right] &\leq \mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}\left[\frac{L+\ell(0)}{\gamma}\mathbb{I}_{(x\text{ violates }(8))}\right.\\ &\qquad\left.+\left(\ell\left(\frac{yg(x)}{\gamma}\right)+\frac{L}{\gamma}\sqrt{\frac{2\log\left(\frac{4}{\delta^2}\right)}{d_1}}\right)\mathbb{I}_{(x\text{ does not violate }(8))}\right]\\ &\leq \frac{(L+\ell(0))\,\delta}{\gamma}+\mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}\left[\ell\left(\frac{yg(x)}{\gamma}\right)\right]+\frac{L}{\gamma}\sqrt{\frac{2\log\left(\frac{4}{\delta^2}\right)}{d_1}}\\ &\hspace{12em}\text{(By definition of a good similarity.)}\\ &\leq \frac{(L+\ell(0))\,\delta}{\gamma}+\varepsilon+\frac{L}{\gamma}\sqrt{\frac{2\log\left(\frac{4}{\delta^2}\right)}{d_1}}\\ &\hspace{12em}\left(\delta<\tfrac{\gamma\varepsilon_1}{2(L+\ell(0))}.\right)\\ &\leq \frac{\varepsilon_1}{2}+\varepsilon+\frac{L}{\gamma}\sqrt{\frac{2\log\left(\frac{4}{\delta^2}\right)}{d_1}}.\end{aligned}$$

Furthermore, the number d_1 of reasonable landmarks follows a binomial distribution $B(d,p)$ with $p\geq\tau$. With our choice of d, we have that

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}_{\mathcal{Z}}^d}\left[\frac{L}{\gamma}\sqrt{\frac{2\log\left(\frac{4}{\delta^2}\right)}{d_1}}\leq\frac{\varepsilon_1}{2}\right]\geq 1-\frac{\delta}{4}.$$

Using the union bound, we obtain with a probability of at least $1-\frac{\delta}{2}$ over S that

$$\mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}\left[\ell\left(\frac{y\left\langle\alpha,\phi^S(x)\right\rangle}{\gamma}\right)\right]\leq\varepsilon+\varepsilon_1.$$

Bound on the fairness criterion For any example $(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}$, we have

$$\left\langle\alpha,\phi^S(x)\right\rangle=\frac{\sum_{i=1}^d y_i'R(x_i',s_i',y_i')\,K(x,x_i')}{d_1},$$

which is an empirical average of d_1 terms with $R(x_i',s_i',y_i')=1$ and

$$-1\leq y_i'R(x_i',s_i',y_i')\,K(x,x_i')\leq 1.$$

Let $g(x)=\mathbb{E}_{(x',s',y')\sim\mathcal{D}_{\mathcal{Z}}}\left[y'K(x,x')\,|R(x',s',y')\right]$. Using the same kind of argument than in the first part of the proof, that is applying Hoeffding's inequality followed by Markov's inequality, we can show that

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}_{\mathcal{Z}}^d}\left[\mathop{\mathbb{P}}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}\left[\left|\left\langle\alpha,\phi^S(x)\right\rangle-g(x)\right|\geq\sqrt{\frac{2\log\left(\frac{8}{\delta^2}\right)}{d_1}}\right]\leq\delta\right]\geq 1-\frac{\delta}{4}.$$

Furthermore, notice that the number d_1 of reasonable landmarks follows a binomial distribution $B(d,p)$ with $p\geq\tau$. With our choice of d, with probability at least $1-\frac{\delta}{4}$ over the choice of S, it implies that

$$\sqrt{\frac{2\log\left(\frac{8}{\delta^2}\right)}{d_1}}\leq\gamma.$$

As a consequence, we have that

$$\underset{S\sim\mathcal{D}_{\mathcal{Z}}^{d}}{\mathbb{P}}\left[\underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}\left[\left|\left\langle\alpha,\phi^{S}(x)\right\rangle-g(x)\right|\geq\gamma\right]\leq\delta\right]\geq 1-\frac{\delta}{2}. \tag{9}$$

To derive a bound on $|\mathrm{DDP}(\alpha)|$, we first derive bounds on $\mathbb{P}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}\left[\left\langle\alpha,\phi^{S}(x)\right\rangle\geq 0\middle|\, s=1\right]$ and $\mathbb{P}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}\left[\left\langle\alpha,\phi^{S}(x)\right\rangle\geq 0\middle|\, s=-1\right]$. Notice that:

$$\begin{aligned}
\underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}\left[\left\langle\alpha,\phi^{S}(x)\right\rangle\geq 0\middle|\, s=1\right] &\geq \underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}\left[g(x)\geq\gamma\cap\left|\left\langle\alpha,\phi^{S}(x)\right\rangle-g(x)\right|\leq\gamma\middle|\, s=1\right]\\
&\geq 1-\underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}\left[g(x)<\gamma\cup\left|\left\langle\alpha,\phi^{S}(x)\right\rangle-g(x)\right|>\gamma\middle|\, s=1\right]\\
&\qquad\qquad\qquad\qquad\text{(Union's bound.)}\\
&\geq 1-\underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}\left[g(x)<\gamma|\, s=1\right]-\underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}\left[\left|\left\langle\alpha,\phi^{S}(x)\right\rangle-g(x)\right|>\gamma\middle|\, s=1\right]\\
&\qquad\qquad\qquad\qquad\left(\mathbb{P}\left[A|B\right]\leq\tfrac{\mathbb{P}[A]}{\mathbb{P}[B]}\right)\\
&\geq \underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}\left[g(x)\geq\gamma|\, s=1\right]-\underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}\left[\left|\left\langle\alpha,\phi^{S}(x)\right\rangle-g(x)\right|>\gamma\middle|\, s=1\right]\\
&\geq \underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}\left[g(x)\geq\gamma|\, s=1\right]-\frac{\mathbb{P}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}\left[\left|\left\langle\alpha,\phi^{S}(x)\right\rangle-g(x)\right|>\gamma\right]}{p_1},
\end{aligned}$$

where $p_1=\mathbb{P}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}\left[s=1\right]$. With a symmetric argument, we have that

$$\underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}\left[\left\langle\alpha,\phi^{S}(x)\right\rangle<0\middle|\, s=1\right]\geq\underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}\left[g(x)\leq-\gamma|\, s=1\right]-\frac{\mathbb{P}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}\left[\left|\left\langle\alpha,\phi^{S}(x)\right\rangle-g(x)\right|>\gamma\right]}{p_1},$$

which implies

$$\begin{aligned}
&1-\underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}\left[\left\langle\alpha,\phi^{S}(x)\right\rangle\geq 0\middle|\, s=1\right]\geq\underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}\left[g(x)\leq-\gamma|\, s=1\right]-\frac{\mathbb{P}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}\left[\left|\left\langle\alpha,\phi^{S}(x)\right\rangle-g(x)\right|>\gamma\right]}{p_1}\\
\Leftrightarrow\quad&\underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}\left[\left\langle\alpha,\phi^{S}(x)\right\rangle\geq 0\middle|\, s=1\right]\leq 1-\underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}\left[g(x)\leq-\gamma|\, s=1\right]+\frac{\mathbb{P}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}\left[\left|\left\langle\alpha,\phi^{S}(x)\right\rangle-g(x)\right|>\gamma\right]}{p_1}\\
\Leftrightarrow\quad&\underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}\left[\left\langle\alpha,\phi^{S}(x)\right\rangle\geq 0\middle|\, s=1\right]\leq\underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}\left[g(x)\geq-\gamma|\, s=1\right]+\frac{\mathbb{P}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}\left[\left|\left\langle\alpha,\phi^{S}(x)\right\rangle-g(x)\right|>\gamma\right]}{p_1}.
\end{aligned}$$

Furthermore, we have that

$$\begin{aligned}
\underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}\left[g(x)\geq-\gamma|\, s=1\right]&\leq\underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}\left[-\gamma\leq g(x)\leq\gamma\cup g(x)\geq\gamma|\, s=1\right]\\
&\qquad\text{(Using the union bound and by definition of a good similarity.)}\\
&\leq\frac{\nu}{p_1}+\underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}\left[g(x)\geq\gamma|\, s=1\right].
\end{aligned}$$

This implies that

$$\underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}\left[\left\langle\alpha,\phi^{S}(x)\right\rangle\geq 0\middle|\, s=1\right]\leq\frac{\nu}{p_1}+\underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}\left[g(x)\geq\gamma|\, s=1\right]+\frac{\mathbb{P}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}\left[\left|\left\langle\alpha,\phi^{S}(x)\right\rangle-g(x)\right|>\gamma\right]}{p_1}.$$

In a similar fashion, we have that

$$\underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}\left[\left\langle\alpha,\phi^{S}(x)\right\rangle\geq 0\middle|\, s=-1\right]\geq\underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}\left[g(x)\geq\gamma|\, s=-1\right]-\frac{\mathbb{P}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}\left[\left|\left\langle\alpha,\phi^{S}(x)\right\rangle-g(x)\right|>\gamma\right]}{1-p_1}$$

and

$$\underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}\left[\left\langle\alpha,\phi^{S}(x)\right\rangle\geq 0\middle|\, s=-1\right]\leq\frac{\nu}{1-p_1}+\underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}\left[g(x)\geq\gamma|\, s=-1\right]+\frac{\mathbb{P}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}\left[\left|\left\langle\alpha,\phi^{S}(x)\right\rangle-g(x)\right|>\gamma\right]}{1-p_1}.$$

These inequalities imply an upper bound on $\mathrm{DDP}(\alpha)$,

$$\begin{aligned}
\mathrm{DDP}(\alpha) &= \underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}\left[\left\langle\alpha,\phi^S(x)\right\rangle \geq 0\middle|\, s=1\right] - \underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}\left[\left\langle\alpha,\phi^S(x)\right\rangle \geq 0\middle|\, s=-1\right] \\
&\leq \frac{\nu}{p_1} + \underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}\left[g(x)\geq\gamma|\, s=1\right] + \frac{\mathbb{P}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}\left[\left|\left\langle\alpha,\phi^S(x)\right\rangle - g(x)\right| > \gamma\right]}{p_1} \\
&\quad - \underset{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}{\mathbb{P}}\left[g(x)\geq\gamma|\, s=-1\right] + \frac{\mathbb{P}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}\left[\left|\left\langle\alpha,\phi^S(x)\right\rangle - g(x)\right| > \gamma\right]}{1-p_1} \\
&\hspace{20em}\text{(By definition of a good similarity.)} \\
&\leq \frac{\nu}{p_1} + \mu + \frac{\mathbb{P}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}\left[\left|\left\langle\alpha,\phi^S(x)\right\rangle - g(x)\right| > \gamma\right]}{p_1} + \frac{\mathbb{P}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}\left[\left|\left\langle\alpha,\phi^S(x)\right\rangle - g(x)\right| > \gamma\right]}{1-p_1}
\end{aligned}$$

and, similarly these inequalities imply a lower bound on $\mathrm{DDP}(\alpha)$,

$$\mathrm{DDP}(\alpha) \geq -\frac{\nu}{1-p_1} - \mu - \frac{\mathbb{P}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}\left[\left|\left\langle\alpha,\phi^S(x)\right\rangle - g(x)\right| > \gamma\right]}{p_1} - \frac{\mathbb{P}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}\left[\left|\left\langle\alpha,\phi^S(x)\right\rangle - g(x)\right| > \gamma\right]}{1-p_1}.$$

Then, using Inequality 9 and the union bound, we obtain that

$$\underset{S\sim\mathcal{D}_{\mathcal{Z}}^d}{\mathbb{P}}\left[\mathrm{DDP}(\alpha) \leq \frac{\nu}{p_1} + \mu + \frac{\delta}{p_1} + \frac{\delta}{1-p_1}\right] \geq 1-\delta$$

In a similar fashion, we also obtain that

$$\underset{S\sim\mathcal{D}_{\mathcal{Z}}^d}{\mathbb{P}}\left[\mathrm{DDP}(\alpha) \geq -\frac{\nu}{1-p_1} - \mu - \frac{\delta}{p_1} - \frac{\delta}{1-p_1}\right] \geq 1-\delta$$

We can combine both inequalities with the union bound to obtain

$$\underset{S\sim\mathcal{D}_{\mathcal{Z}}^d}{\mathbb{P}}\left[|\mathrm{DDP}(\alpha)| \leq \mu + (\nu+2\delta)\max\left(\frac{1}{p_1},\frac{1}{1-p_1}\right)\right] \geq 1-2\delta$$

Using the union one last time to combine the fairness bound and the target criterion bound gives the theorem. □