Too Relaxed to Be Fair
Supplementary Material

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

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

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

Then, we choose a lower bound $\lambda_{\min}$ and an upper bound $\lambda_{\max}$, so that $\mathrm{sign}\Big(\mathrm{DDP}\Big(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_{\min})\Big)\Big) \neq \mathrm{sign}\Big(\mathrm{DDP}\Big(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_{\max})\Big)\Big)$. If $\mathrm{R}_{\widehat{\mathrm{DDP}}}(f)$ is chosen correctly then setting $\lambda_{\min} = 0$ and $\lambda_{\max} = 1$ 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 $\lambda_{\min} = 0$ and $\lambda_{\max} = 1$ this choice often ensures that $\mathrm{sign}(\mathrm{DDP}(f(\lambda_{\min}))) \neq \mathrm{sign}(\mathrm{DDP}(f(\lambda_{\max})))$. 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 $\mathrm{sign}\Big(\mathrm{DDP}\Big(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_{\min})\Big)\Big) \neq \mathrm{sign}\Big(\mathrm{DDP}\Big(f^{\beta}_{\widehat{\mathcal{D}}_{\mathcal{Z}}}(\lambda_{\max})\Big)\Big)$ is respected.

[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.

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 $\beta > 0$, λ_{min} and λ_{max} the lower and upper bounds for λ, 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 λ_{min} and λ_{max}, or choose a new fairness regularizer $\text{R}_{\widehat{\text{DDP}}}(f)$.
end if

Satisfying the conditions of Theorem 1. 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 $\frac{d\hat{L}(f^0)}{df} + \lambda\frac{d\text{R}_{\widehat{\text{DDP}}}(f^0)}{df} + \beta\frac{d\Omega(f^0)}{df} = \mathbf{0}$ 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 $\text{R}_{\widehat{\text{DDP}}}(f)$. 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 datasets in Figures 1– 6. 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 7 and 8. 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

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

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 1. 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 2. 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 3. 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.

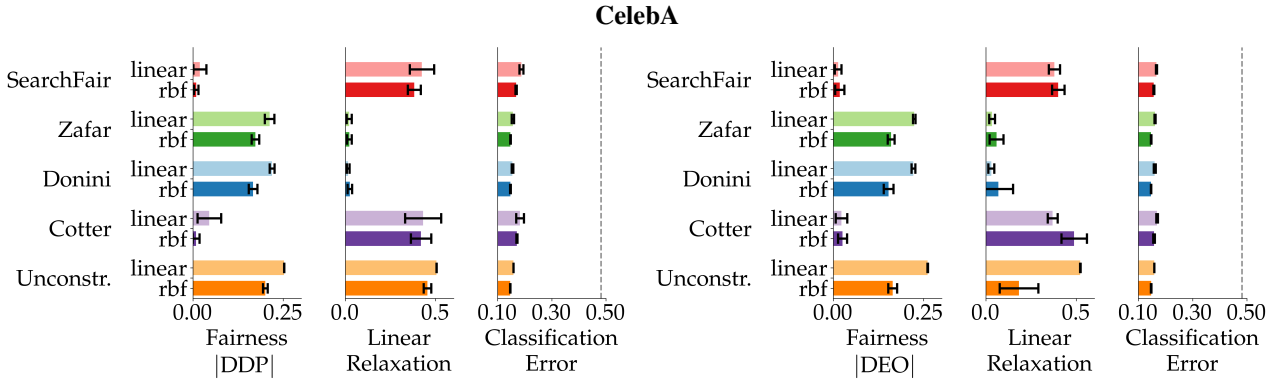
Compas–Figure 4. The Compas dataset (Larson et al., 2016) contains 7214 points with 53 features, such as name, age, degree of crime, and number of prior crimes. We use the same pre-processing as Zafar et al. (2017a) and, in particular, we select the same 5 features. The goal is to predict if a defendant 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 5. 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 6. 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.

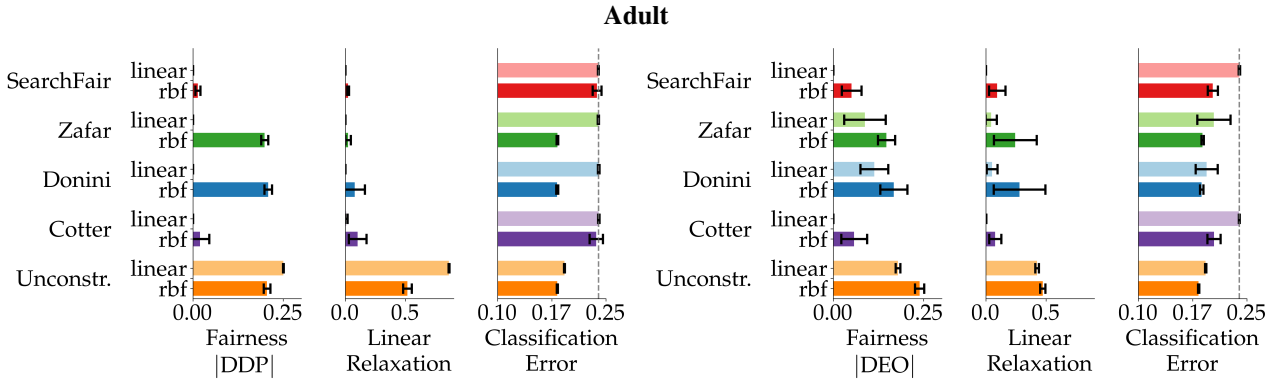
Cross Validation Procedure. We report the results for different cross validation procedures as discussed in the main paper. In Figure 7 we use a procedure called NVP proposed by (Donini et al., 2018). In a first step, we exclude the hyperparameters with an accuracy score that is lower than 90% of the 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 8 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.



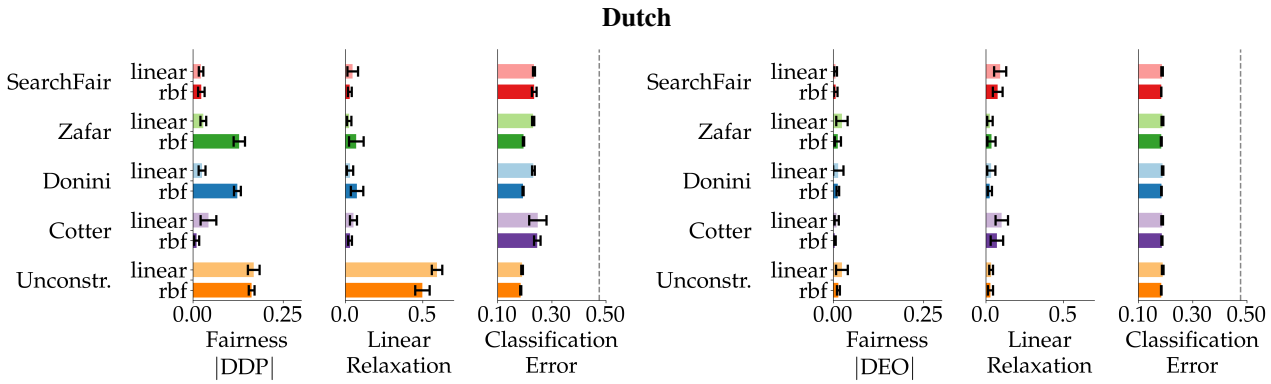
FAIRNESS NOTION		Demographic Parity			Equality of Opportunity		
	Kernel	|DDP|	|LR|	Error	|DEO|	|LR|	Error
SearchFair	linear	0.02 ± 0.02	0.42 ± 0.07	0.19 ± 0.01	0.01 ± 0.01	0.38 ± 0.01	0.17 ± 0.00
	rbf	0.01 ± 0.01	0.38 ± 0.04	0.17 ± 0.00	0.02 ± 0.01	0.40 ± 0.04	0.16 ± 0.00
Zafar	linear	0.21 ± 0.01	0.02 ± 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 1. **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.



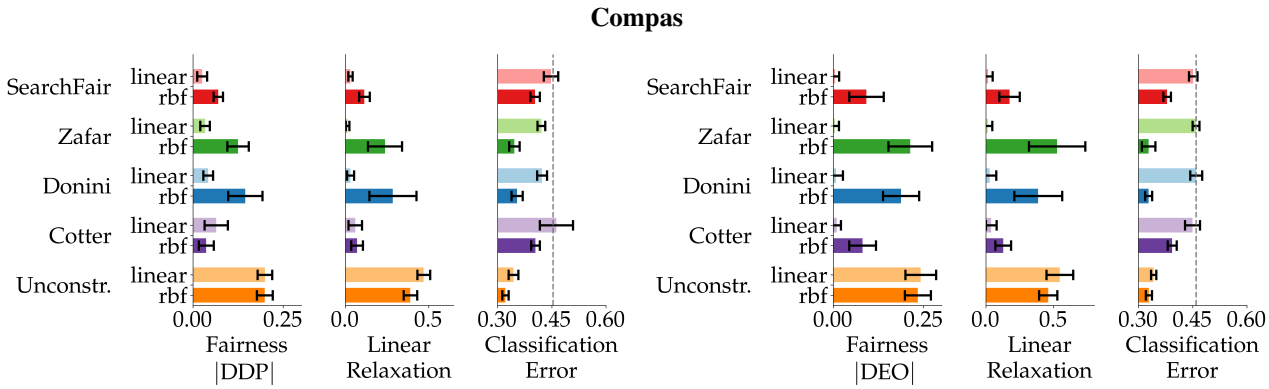
FAIRNESS NOTION		Demographic Parity			Equality of Opportunity		
	Kernel	|DDP|	|LR|	Error	|DEO|	|LR|	Error
SearchFair	linear	0.00 ± 0.00	0.00 ± 0.00	0.24 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.24 ± 0.00
	rbf	0.01 ± 0.01	0.02 ± 0.01	0.24 ± 0.01	0.05 ± 0.03	0.09 ± 0.07	0.20 ± 0.01
Zafar	linear	0.00 ± 0.00	0.00 ± 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 2. **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.



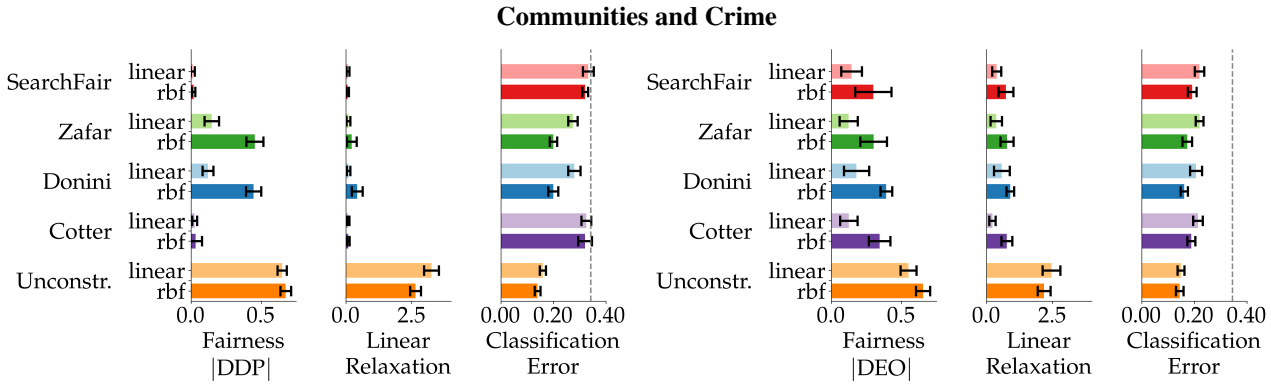
FAIRNESS NOTION		Demographic Parity			Equality of Opportunity		
	Kernel	|DDP|	|LR|	Error	|DEO|	|LR|	Error
SearchFair	linear	0.02 ± 0.01	0.05 ± 0.03	0.23 ± 0.00	0.01 ± 0.00	0.09 ± 0.04	0.19 ± 0.00
	rbf	0.02 ± 0.01	0.03 ± 0.01	0.23 ± 0.01	0.01 ± 0.00	0.08 ± 0.03	0.18 ± 0.00
Zafar	linear	0.03 ± 0.01	0.03 ± 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 3. **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.



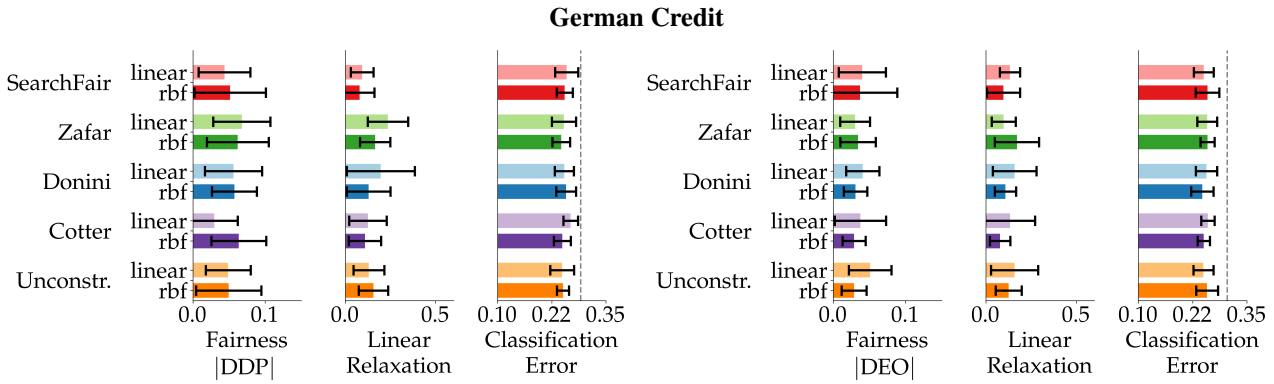
FAIRNESS NOTION		Demographic Parity			Equality of Opportunity		
	Kernel	|DDP|	|LR|	Error	|DEO|	|LR|	Error
SearchFair	linear	0.03 ± 0.01	0.03 ± 0.01	0.45 ± 0.02	0.01 ± 0.01	0.02 ± 0.03	0.45 ± 0.01
	rbf	0.07 ± 0.01	0.11 ± 0.03	0.40 ± 0.01	0.09 ± 0.05	0.17 ± 0.08	0.38 ± 0.01
Zafar	linear	0.03 ± 0.01	0.01 ± 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 4. **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.



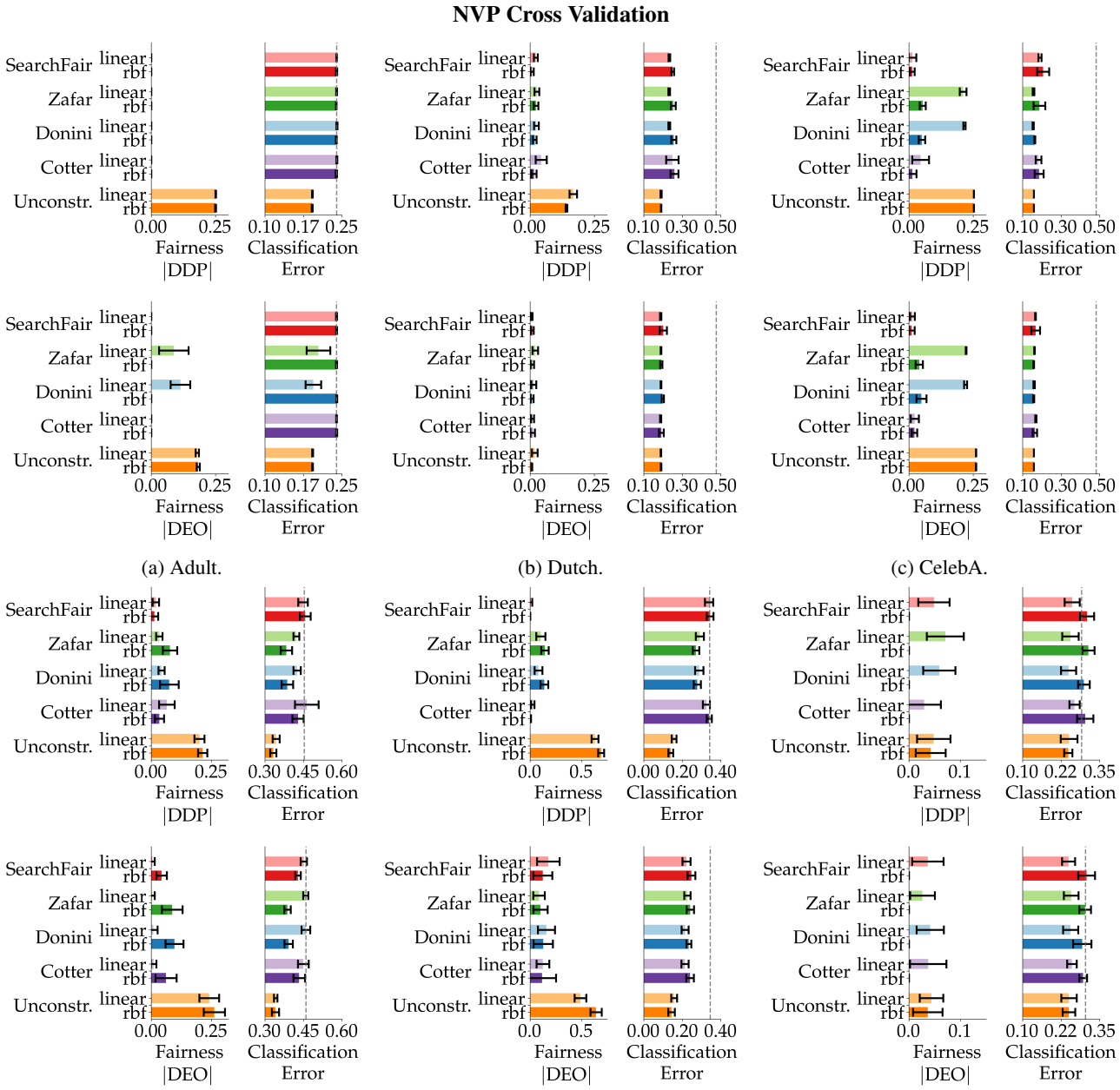
FAIRNESS NOTION		Demographic Parity			Equality of Opportunity		
	Kernel	|DDP|	|LR|	Error	|DEO|	|LR|	Error
SearchFair	linear	0.01 ± 0.01	0.07 ± 0.05	0.33 ± 0.02	0.14 ± 0.07	0.38 ± 0.17	0.22 ± 0.02
	rbf	0.02 ± 0.01	0.06 ± 0.05	0.32 ± 0.01	0.30 ± 0.13	0.73 ± 0.28	0.19 ± 0.02
Zafar	linear	0.15 ± 0.05	0.09 ± 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 5. **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

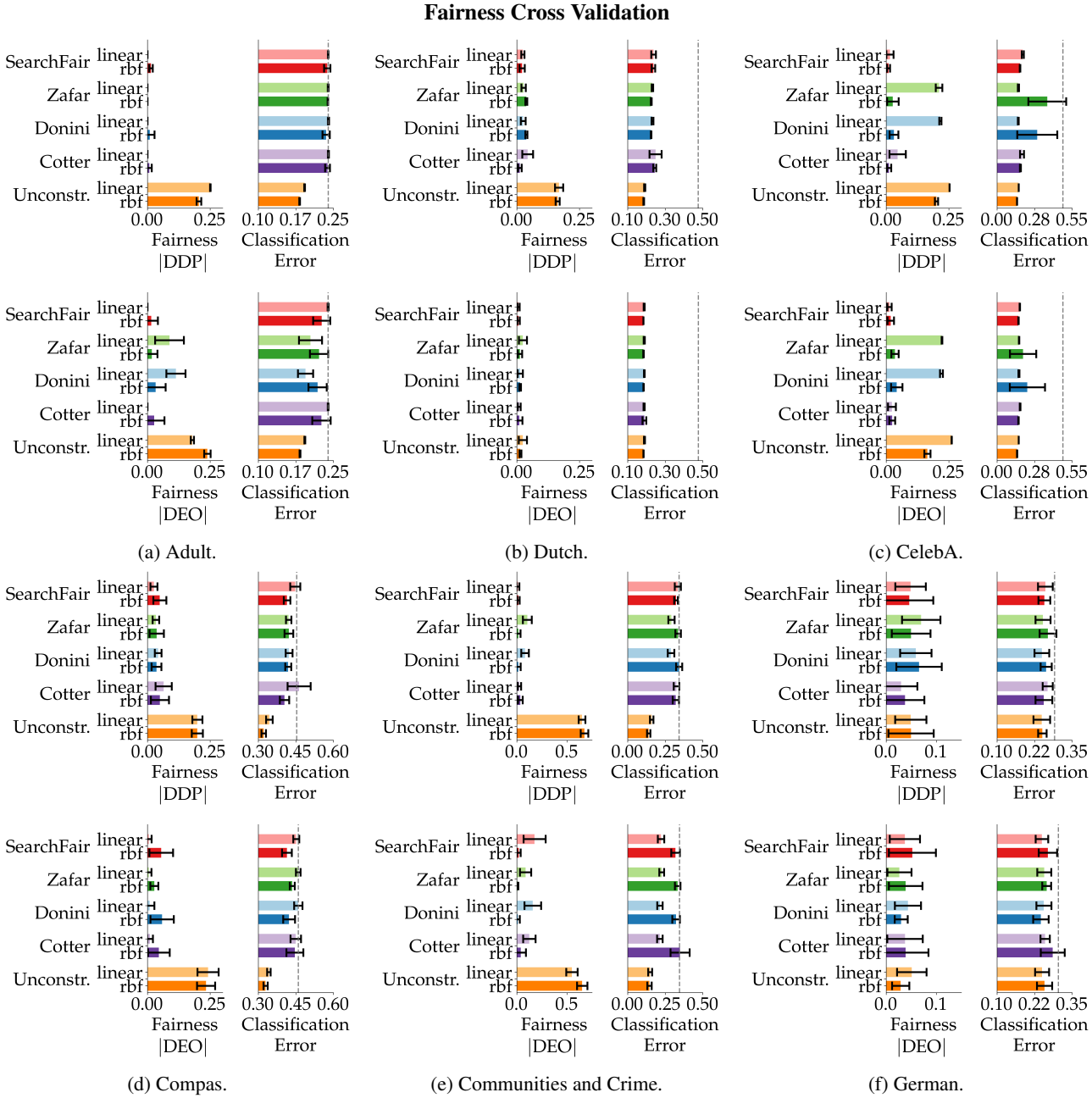
Figure 6. **German Credit.** The grey dashed vertical line depicts the classification error of the constant classifier, which is perfectly fair for both DDP and DEO. This is the smallest dataset out of the 6 with 700 training examples and 300 test examples. This explains the large standard deviations. 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 7. 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 8. 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 7 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 1 is proved in Section C.3, Corollary 1 is proved in Section C.4, and Theorem 2 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 $\text{LR}_{\widehat{\text{DDP}}}(f) = \text{LR}_{\widehat{\text{DDP}}}(g)$ 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:

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

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

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

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

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

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

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

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

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:

$$
f^\beta_{\widehat{\mathcal{D}}_\mathcal{Z}}(\lambda) = \arg\min_{f\in\mathcal{F}} \hat{L}(f) + \lambda \mathrm{R}_{\widehat{\mathrm{DDP}}}(f) + \beta\Omega(f)\,. \tag{1}
$$

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, $f \mapsto \mathbb{P}_{x\sim\mathcal{D}_\mathcal{X}}[f(x)\leq 0]$ is a continuous function. Before proving them, we first recall the definition of a m-strongly convex function.

Definition 1 (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 1 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 1 is m-strongly convex, it holds that:

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

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

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

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

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

$$(\varepsilon \leq |\lambda' - \lambda|)$$

$$\Rightarrow \qquad \|f(\lambda) - f(\lambda')\|_{\mathcal{F}}^2 \leq \frac{8B}{m} |\lambda' - \lambda|$$

$$\Rightarrow \qquad \|f(\lambda) - f(\lambda')\|_{\mathcal{F}} \leq \sqrt{\frac{8B}{m}} \sqrt{|\lambda' - \lambda|}.$$

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 $\mathbb{P}_{x\sim\mathcal{D}_{\mathcal{X}}}[f(x) \leq 0]$ 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 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 1 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 $\lambda \mapsto DDP\Big(f^\beta_{\widehat{\mathcal{D}}_\mathcal{Z}}(\lambda)\Big)$ is continuous.

Proof. Recall that DDP is defined as follows:

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

Applying Lemma 2, we have that $c : \mathcal{F}_\Lambda \to \mathbb{R}$ defined as $c(f) = \mathbb{P}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}\,|\,s=1}}[f(x) > 0]$ and $c' : \mathcal{F}_\Lambda \to \mathbb{R}$ defined as $c'(f) = \mathbb{P}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}\,|\,s=-1}}[f(x) > 0]$ are continuous functions. It implies that the function $q : \mathcal{F}_\Lambda \to \mathbb{R}$ defined as $q(f) = \text{DDP}(f)$ 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$\Big(f^\beta_{\widehat{\mathcal{D}}_\mathcal{Z}}(\lambda)\Big)$). *Let $\mathcal{F}$ be a function space, we define the set of learnable functions as $\mathcal{F}_\Lambda = \Big\{f \in \mathcal{F} : \exists\lambda \geq 0, f = f^\beta_{\widehat{\mathcal{D}}_\mathcal{Z}}(\lambda)\Big\}$. Assume that the following conditions hold:*
(i) *Optimization Problem 1 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 $\lambda \mapsto DEO\Big(f^\beta_{\widehat{\mathcal{D}}_\mathcal{Z}}(\lambda)\Big)$ is continuous.

Proof. Recall that DEO is defined as follows:

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

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

Corollary 1 (Existence of a DDP-fair classifier). *Let $\mathcal{F}$ be a function space, we define the set of learnable functions as $\mathcal{F}_\Lambda = \Big\{f \in \mathcal{F} : \exists\lambda \geq 0, f = f^\beta_{\widehat{\mathcal{D}}_\mathcal{Z}}(\lambda)\Big\}$. Assume that Theorem 1 holds and that there exist two hyperparameters λ_+ and λ_- such that $DDP\Big(f^\beta_{\widehat{\mathcal{D}}_\mathcal{Z}}(\lambda_+)\Big) > 0$ and $DDP\Big(f^\beta_{\widehat{\mathcal{D}}_\mathcal{Z}}(\lambda_-)\Big) < 0$. Then, there exists at least one value $\lambda_0 \in [\min(\lambda_+, \lambda_-), \max(\lambda_+, \lambda_-)]$ such that $DDP\Big(f^\beta_{\widehat{\mathcal{D}}_\mathcal{Z}}(\lambda_0)\Big) = 0$.*

Proof. This corollary is a direct consequence of the intermediate value theorem and the continuity of DDP proven in Theorem 1. □

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 1 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 2

Recall the definition of a good similarity for fairness.

Definition 2 (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) $\left|\mathop{\mathbb{P}}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}\,|s=1}}[g(x) \geq \gamma] - \mathop{\mathbb{P}}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}\,|s=-1}}[g(x) \geq \gamma]\right| \leq \mu,$

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

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

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

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 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

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

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

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

Proof. Let $S = \{(x'_1, s'_1, y'_1), \ldots, (x'_d, s'_d, y'_d)\}$ 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 $i \in \{1, \ldots, d\}$. 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 $\alpha_i = \frac{y'_i R(x'_i, s'_i, y'_i)}{d_1}$ 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

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

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

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

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

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

Using Hoeffding's inequality, we can show that

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

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

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

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

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

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

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

Then, applying Markov's inequality, we obtain that

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

which implies

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

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

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

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 (2), we have, for our decreasing loss ℓ (an example of decreasing loss is the hinge loss, $\ell(w) = \max(0, 1-w)$):

$$\ell\left(\frac{y\left\langle\alpha,\phi^S(x)\right\rangle}{\gamma}\right) \leq \ell\left(\frac{yg(x) - \sqrt{\frac{2\log\left(\frac{4}{\delta^2}\right)}{d_1}}}{\gamma}\right)$$

(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 (2), 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.)

$$\leq L\max_x \frac{\|\alpha\|_1\left\|\phi^S(x)\right\|_\infty}{\gamma} + \ell(0)$$

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

($\gamma \leq 1$.)

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

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

$$\mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}\left[\ell\left(\frac{y\left\langle\alpha,\phi^S(x)\right\rangle}{\gamma}\right)\right] \leq \mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}\left[\frac{L+\ell(0)}{\gamma}\mathbb{I}_{(x \textit{ violates (2)})}\right.$$
$$\left. + \left(\ell\left(\frac{yg(x)}{\gamma}\right) + \frac{L}{\gamma}\sqrt{\frac{2\log\left(\frac{4}{\delta^2}\right)}{d_1}}\right)\mathbb{I}_{(x \textit{ does not violate (2)})}\right]$$

$$\leq \frac{(L+\ell(0))\,\delta}{\gamma} + \mathop{\mathbb{E}}_{(x,s,y)\sim\mathcal{D}_\mathcal{Z}}\left[\ell\left(\frac{yg(x)}{\gamma}\right)\right] + \frac{L}{\gamma}\sqrt{\frac{2\log\left(\frac{4}{\delta^2}\right)}{d_1}}$$

(By definition of a good similarity.)

$$\leq \frac{(L+\ell(0))\,\delta}{\gamma} + \varepsilon + \frac{L}{\gamma}\sqrt{\frac{2\log\left(\frac{4}{\delta^2}\right)}{d_1}}$$

($\delta < \frac{\gamma\varepsilon_1}{2(L+\ell(0))}$.)

$$\leq \frac{\varepsilon_1}{2} + \varepsilon + \frac{L}{\gamma}\sqrt{\frac{2\log\left(\frac{4}{\delta^2}\right)}{d_1}}.$$

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

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}_{\mathcal{Z}}^{d}}\left[\mathop{\mathbb{P}}_{(x,s,y)\sim\mathcal{D}_{\mathcal{Z}}}\left[\left|\left\langle\alpha,\phi^S(x)\right\rangle-g(x)\right| \geq \gamma\right] \leq \delta\right] \geq 1-\frac{\delta}{2}. \tag{3}$$

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:

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

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

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

which implies

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

Furthermore, we have that

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

This implies that

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

In a similar fashion, we have that

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

and

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

These inequalities imply an upper bound on DDP(α),

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

and, similarly these inequalities imply a lower bound on DDP(α),

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

Then, using Inequality 3 and the union bound, we obtain that

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

In a similar fashion, we also obtain that

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

We can combine both inequalities with the union bound to obtain

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

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

References

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.

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

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.

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.

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