A Particularizing the Fairness Constraints for SVM

One can specialize our fair classifier formulation proposed in (4) as:

Linear SVM. A linear SVM distinguishes among classes using a linear hyperplane $\boldsymbol{\theta}^T\mathbf{x} = 0$. In this case, the parameter vector $\boldsymbol{\theta}$ of the *fair* linear SVM can be found by solving the following quadratic program:

$$\begin{array}{ll} \text{minimize} & \|\boldsymbol{\theta}\|^2 + C\sum_{i=1}^n \xi_i \\ \text{subject to} & y_i\boldsymbol{\theta}^T\mathbf{x}_i \geq 1 - \xi_i, \forall i \in \{1,\ldots,n\} \\ & \xi_i \geq 0, \forall i \in \{1,\ldots,n\}, \\ & \frac{1}{N}\sum_{i=1}^N (\mathbf{z}_i - \bar{\mathbf{z}})\,\boldsymbol{\theta}^T\mathbf{x}_i \leq \mathbf{c}, \\ & \frac{1}{N}\sum_{i=1}^N (\mathbf{z}_i - \bar{\mathbf{z}})\,\boldsymbol{\theta}^T\mathbf{x}_i \geq -\mathbf{c}, \end{array} \tag{9}$$

where $\boldsymbol{\theta}$ and ξ are the variables, $\|\boldsymbol{\theta}\|^2$ corresponds to the margin between the *support vectors* assigned to different classes, and $C\sum_{i=1}^n \xi_i$ penalizes the number of data points falling inside the margin.

Nonlinear SVM. In a nonlinear SVM, the decision boundary takes the form $\boldsymbol{\theta}^T\Phi(\mathbf{x}) = 0$, where $\Phi(\cdot)$ is a nonlinear transformation that maps every feature vector $\mathbf{x}$ into a higher dimensional transformed feature space. Similarly as in the case of a linear SVM, one may think of finding the parameter vector $\boldsymbol{\theta}$ by solving a constrained quadratic program similar to the one defined by Eq. (9). However, the dimensionality of the transformed feature space can be large, or even infinite, making the corresponding optimization problem difficult to solve. Fortunately, we can leverage the *kernel trick* [Schölkopf and Smola, 2002] both in the original optimization problem and the fairness inequalities, and resort instead to the dual form of the problem, which can be solved efficiently. In particular, the dual form is given by:

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^N \alpha_i + \sum_{i=1}^N \alpha_i y_i (g_{\boldsymbol{\alpha}}(\mathbf{x}_i) + h_{\boldsymbol{\alpha}}(\mathbf{x}_i)) \\ \text{subject to} & \alpha_i \geq 0, \forall i \in \{1,\ldots,N\}, \\ & \sum_{i=1}^N \alpha_i y_i = 0, \\ & \frac{1}{N}\sum_{i=1}^N (\mathbf{z}_i - \bar{\mathbf{z}})\, g_{\boldsymbol{\alpha}}(\mathbf{x}_i) \leq \mathbf{c}, \\ & \frac{1}{N}\sum_{i=1}^N (\mathbf{z}_i - \bar{\mathbf{z}})\, g_{\boldsymbol{\alpha}}(\mathbf{x}_i) \geq -\mathbf{c}, \end{array} \tag{10}$$

where $\boldsymbol{\alpha}$ are the dual variables, $g_{\boldsymbol{\alpha}}(\mathbf{x}_i) = \sum_{j=1}^N \alpha_j y_j k(\mathbf{x}_i, \mathbf{x}_j)$ can still be interpreted as a signed distance to the decision boundary in the transformed feature space, and $h_{\boldsymbol{\alpha}}(\mathbf{x}_i) = \sum_{j=1}^N \alpha_j y_j \frac{1}{C}\delta_{ij}$, where $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$, otherwise. Here, $k(\mathbf{x}_i, \mathbf{x}_j) = \langle\phi(\mathbf{x}_i), \phi(\mathbf{x}_j)\rangle$ denotes the inner product between a pair of transformed feature vectors and is often called the kernel function.

B Additional Experiments

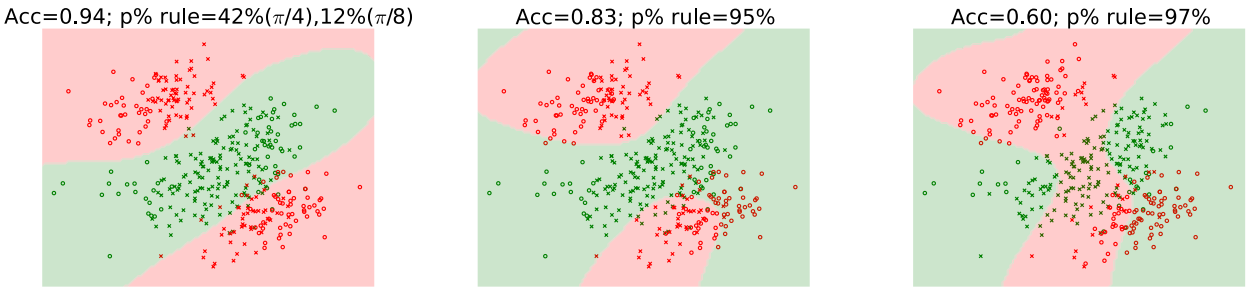
B.1 Experiments on Non-linear Synthetic Data

Here, we illustrate how the decision boundary of an non-linear classifier, a SVM with radial basis function (RBF) kernel, changes under fairness constraints. To this end, we generate 4,000 user binary class labels uniformly at random and assign a 2-dimensional user feature vector per label by drawing samples from $p(x|y = 1, \beta) = \beta N([2;2],[5\ 1;1\ 5]) + (1-\beta)N([-2;-2],[10\ 1;1\ 3])$ if $y = 1$, and $p(x|y = -1, \beta) = \beta N([4;-4],[4\ 4;2\ 5]) + (1-\beta)N([-4;6],[6\ 2;2\ 3])$ otherwise, where $\beta \in \{0, 1\}$ is sampled from $Bernoulli(0.5)$. Then, we generate each user's sensitive attribute z by applying the same rotation as detailed in Section 4.1.

Figure 5 shows the decision boundaries provided by the SVM that maximizes accuracy under fairness constraints with $\mathbf{c} = 0$ for two different correlation values, set by $\phi = \pi/4$ and $\phi = \pi/8$, in comparison with the unconstrained SVM. We observe that, in this case, the decision boundaries provided by the constrained SVMs are very different to the decision boundary provided by the unconstrained SVM, not simple shifts or rotations of the latter, and successfully reverse engineer the mechanism we used to generate the class labels and sensitive attributes.

B.2 Experiments on Real Data

Additional Data Statistics. In this section, we show the distribution of sensitive features and class labels in our real-world datasets.



(a) Unconstrained (b) $\phi = \pi/4$ (c) $\phi = \pi/8$

Figure 5: Decision boundaries for SVM classifier with RBF Kernel trained without fairness constraints (left) and with fairness constraints ($\mathbf{c} = 0$) on two synthetic datasets with different correlation value between sensitive attribute values (crosses vs circles) and class labels (red vs green).

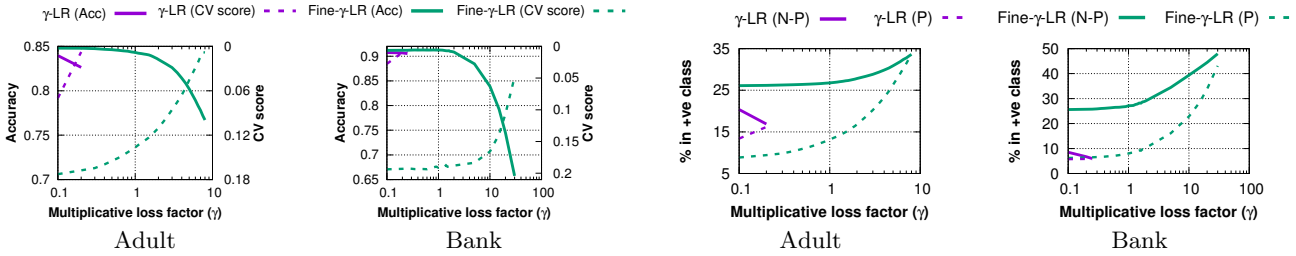
Table 2: Datasets details (binary sensitive attributes: gender and age).

Sensitive Attribute	$y \leq 50K$	$> 50K$	Total
Males	20,988	9,539	30,527
Females	13,026	1,669	14, 695
Total	34,014	11,208	45,222

(a) Adult dataset

Sensitive Attribute	No	Yes	Total
$25 \leq$ age ≤ 60	35,240	3,970	39,210
age < 25 or age > 60	1,308	670	1,978
Total	36,548	4,640	41,188

(b) Bank dataset



Adult Bank Adult Bank

(a) Acc and CV score (b) % in +ve class

Figure 7: [Maximizing fairness under accuracy constraints] Panels in (a) show the accuracy (solid) and CV score value (dashed) against γ. Panels in (b) show the percentage of protected (P, dashed) and non-protected (N-P, solid) users in the positive class against γ.

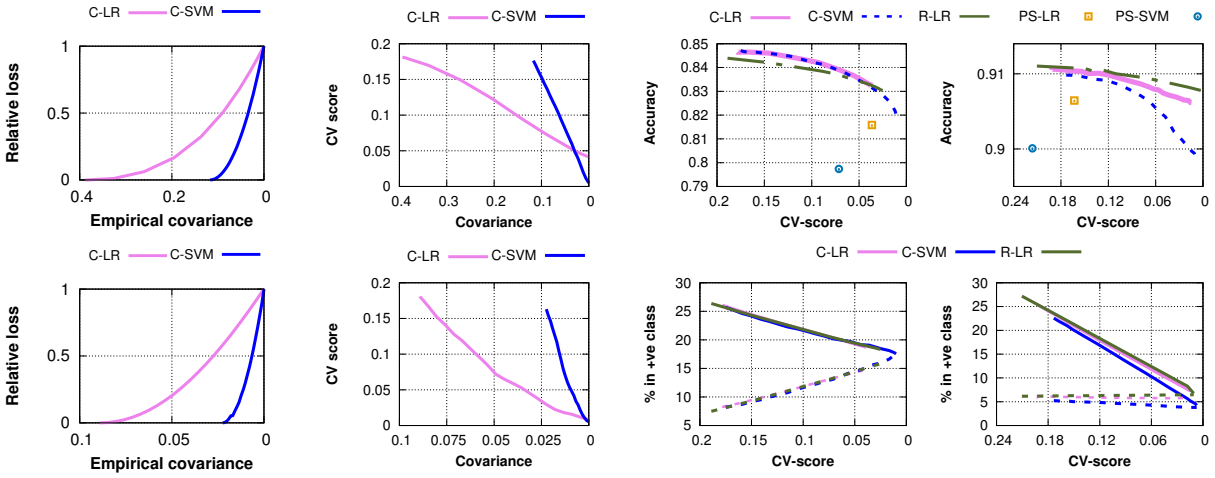
CV score as fairness measure. While evaluating the performance of our method in Section 4, we used p%-rule as the true measure of fairness, since it is a generalization of the 80%-rule advocated by US Equal Employment Opportunity Commission [Biddle, 2005] to quantify disparate impact. We would like to remark that this measure is closely related to another measure of fairness used by some of the previous works [Kamiran and Calders, 2009, Kamishima et al., 2011, Zemel et al., 2013] in this area, referred to as Calder-Verwer (CV) score by Kamishima et al. [2011]. In particular, the CV score is defined as the (absolute value of the) difference between the percentage of users sharing a particular sensitive attribute value that lie on one side of the decision boundary and the percentage of users not sharing that value lying on the same side, *i.e.*, $\big|P(d_{\boldsymbol{\theta}}(\mathbf{x}) \geq 0|z = 0) - P(d_{\boldsymbol{\theta}}(\mathbf{x}) \geq 0|z = 1)\big|$.

In this section, we show that using CV score (instead of p%-rule) as a measure of fairness would yield similar results.

First, we show that constraining the covariance between users' sensitive attributes (Fig. 6a and Figure 6b), and the signed distance from the decision boundary, corresponds to an increasing relative loss and decreasing CV

Table 3: Adult dataset (Non-binary sensitive attribute: race)

Sensitive Attribute	$y \leq 50K$	$> 50K$	Total
American-Indian/Eskimo	382	53	435
Asian/Pacific-Islander	934	369	1,303
White	28,696	10,207	38,903
Black	3,694	534	4,228
Other	308	45	353
Total	34,014	11,208	45,222



(a) Loss vs. Cov. (b) Cov. vs. CV score (c) Single binary sensitive attribute

Figure 6: [Maximizing accuracy under fairness constraints: single, binary sensitive attribute] Panels in (a) show the trade-off between the empirical covariance and the relative loss in accuracy (with respect to the unconstrained classifier), where each pair of (covariance, loss) values is guaranteed to be Pareto optimal by construction. Panels in (b) show the correspondence between the empirical covariance in Eq. 2 and the CV score for classifiers trained under fairness constraints for the Adult (top) and Bank (bottom) datasets. Panels in (c) show the accuracy against CV score value (top) and the percentage of protected (dashed) and non-protected (solid) users in the positive class against the CV score value (bottom).

score (a more fair decision boundary).

Next, we show the performance of different methods based on the CV score (Fig, 6c and 7). The results in Fig.6c and 7 correspond to the ones shown in Fig. 2c and 4, where we took p%-rule as the measure of fairness. It can be seen that both measures of fairness (p%-rule and CV score) provide very similar trades-off in terms of fairness and accuracy.

Notice that according to the definitions provided in Section 2.1, a decreasing CV score corresponds to an increasing p%-rule (and hence, a more fair decision boundary).