Spectral Algorithms for Computing Fair Support Vector Machines

Matt Olfat
UC Berkeley

Anil Aswani
UC Berkeley

Abstract

Classifiers and rating scores are prone to implicitly codifying biases, which may be present in the training data, against protected classes (i.e., age, gender, or race). So it is important to understand how to design classifiers and scores that prevent discrimination in predictions. This paper develops computationally tractable algorithms for designing accurate but fair support vector machines (SVM's). Our approach imposes a constraint on the covariance matrices conditioned on each protected class, which leads to a nonconvex quadratic constraint in the SVM formulation. We develop iterative algorithms to compute fair linear and kernel SVM's, which solve a sequence of relaxations constructed using a spectral decomposition of the nonconvex constraint. Its effectiveness in achieving high prediction accuracy while ensuring fairness is shown through numerical experiments on several data sets.

1 INTRODUCTION

The increasing prevalence of machine learning to automate decision-making systems has drawn societal scrutiny on the approaches used for verification and validation of these systems. In particular, two main concerns have been voiced with regards to the correctness and accuracy of decisions provided. The first is a general lack of interpretability for how predictions are produced by many learning techniques [Ridgeway et al., 1998, Lou et al., 2012, Caruana et al., 2015], and the second is the possibility of perpetuating inequities that may be present in the training data [Podesta et al., 2014, Barocas and Selbst, 2016, Bhandari, 2016]. This paper focuses on the latter: we study how to design fair support vector machines (SVM's), and our goal is to construct a classifier $h(x,t) : \mathbb{R}^p \times \mathbb{R} \rightarrow \{-1,+1\}$ that inputs predictors $x \in \mathbb{R}^p$ and a threshold t, and predicts a label $y \in \{-1,+1\}$, while ensuring fairness with respect to a protected class $z \in \{-1,+1\}$ (e.g., age, gender, or race). We assume there are only two protected classes; however, our formulations generalize to the setting with multiple protected classes.

We make four main contributions. First, we reinterpret two fairness notions using receiver operating characteristic (ROC) curves, which leads to a new visualization for classifier fairness. Second, we capture fairness by defining a constraint on covariance matrices conditioned on protected classes, which leads to a nonconvex quadratic constraint in the SVM formulation. Third, we construct an iterative algorithm that uses a spectral decomposition of the nonconvex constraint to compute fair linear and kernel SVM's; we prove iterates converge to a local minimum. Fourth, we conduct numerical experiments to evaluate our algorithms.

1.1 Fairness Notions for Classifiers

Ensuring classifiers are fair requires quantifying their fairness. However, Friedler et al. [2016] and Kleinberg et al. [2016] showed that no single metric can capture all intuitive aspects of fairness, and so any metric must choose a specific aspect of fairness to quantify. In this paper, we consider arguably the two most popular notions: demographic parity [Calders et al., 2009, Zliobaite, 2015, Zafar et al., 2017] and equal opportunity [Dwork et al., 2012, Hardt et al., 2016]. Precise definitions of these are given in Section 3. These notions are typically considered for a single threshold of the classifier, but here we will consider all possible thresholds. We believe this is more in-line with malicious usage of classifiers in which strategic choice of thresholds can be used to practice discrimination.

1.2 Algorithms to Compute Fair Classifiers

Several approaches have been developed to construct fair classifiers. Some [Zemel et al., 2013, Louizos et al.,

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84. Copyright 2018 by the author(s).

2015] compute transformations of the data to make it independent of the protected class, though this can be too conservative and reduce predictive accuracy more than desired. Another method [Hardt et al., 2016] modifies any classifier to reduce its accuracy with respect to protected classes until fairness is achieved. Several techniques compute a fair classifier at a single threshold [Calders et al., 2009, Cotter et al., 2016]; however, our interest is in classifiers that are fair at all thresholds. The only method we are aware of that tries to compute a fair classifier for all thresholds is that of Zafar et al. [2017], which will be our main comparison.

1.3 Outline

After describing the data and our notation in Section 2, we next define two fairness notions and provide a new ROC visualization of fairness in Section 3. Section 4 derives constraints to improve the fairness of linear and kernel SVM's at all thresholds. This involves nonconvex constraints, and in Section 5 we present iterative algorithms that compute fair linear and kernel SVM's by solving a sequence of convex problems defined using a spectral decomposition. Section 6 conducts numerical experiments using both synthetic and real datasets to demonstrate the efficacy of our approach in computing accurate but fair SVM's.

2 DATA AND NOTATION

Our data consists of 3-tuples (x_i, y_i, z_i) for $i = 1, \dots, n$ points, where $x_i \in \mathbb{R}^p$ are predictors, $y_i \in \{-1, +1\}$ are labels, and $z_i \in \{-1, +1\}$ label a protected class. For a matrix W, the i-th row of W is denoted W_i. Define $X \in \mathbb{R}^{n \times p}$, $Y \in \mathbb{R}^n$, and $Z \in \mathbb{R}^n$ to be matrices and vectors such that $X_i = x_i^\mathsf{T}$, $Y_i = y_i$, and $Z_i = z_i$, respectively.

Let $N = \{i : z_i = -1\}$ be the set of indices for which the protected class is negative, and similarly let $P = \{i : z_i = +1\}$ be the set of indices for which the protected class is positive. We use $\#N$ and $\#P$ for the cardinality of the sets N and P, respectively. Now define X_+ to be a matrix whose rows are x_i^T for $i \in P$, and similarly define X_- to be a matrix whose rows are x_i^T for $i \in N$. Let Σ_+ and Σ_- be the covariance matrices of $[x_i|z_i = +1]$ and $[x_i|z_i = -1]$, respectively.

Next let $K(x, x') : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ be a kernel function, and consider the notation

$$K(X, X') = \begin{bmatrix} K(X_1, X_1') & K(X_1, X_2') & \cdots \\ K(X_2, X_1') & K(X_2, X_2') & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} \quad (1)$$

Recall that the essence of the *kernel trick* is to replace $x_i^\mathsf{T} x_j$ with $K(x_i, x_j)$, and so the benefit of the matrix notation given in (1) is that it allows us to replace $X(X')^\mathsf{T}$ with $K(X, X')$ as part of the kernel trick.

Last, we define some additional notation. Let $[n] = \{1, \dots, n\}$, and note $\mathbf{1}(u)$ is the indicator function. A positive semidefinite matrix U is denoted $U \succeq 0$. If U, V are vectors of equal dimension, then the notation $U \circ V$ refers to their element-wise product: $(U \circ V)_i = U_i \cdot V_i$. Also, $\mathbf{e}$ is the vector whose entries are all 1.

3 ROC Visualization of Fairness

3.1 Demographic Parity

One popular notion of fairness is that predictions of the label y are independent of the protected class z. This definition is typically stated [Calders et al., 2009, Zliobaite, 2015, Zafar et al., 2017] in terms of a single threshold, though it can be generalized to multiple thresholds. We say that a classifier $h(x, t)$ has demographic parity at level Δ (abbreviated as DP-Δ) if

$$\Big| \mathbb{P}\big[h(x,t) = +1 \big| z = +1\big] - \mathbb{P}\big[h(x,t) = +1 \big| z = -1\big] \Big| \leq \Delta, \ \forall t \in \mathbb{R}. \quad (2)$$

To understand this, note $\mathbb{P}\big[h(x,t) = +1\big|z = +1\big]$ is the true positive rate when predicting the protected class at threshold t, while $\mathbb{P}\big[h(x,t) = +1\big|z = -1\big]$ is the false positive rate when predicting the protected class at threshold t. So the intuition is that a classifier is DP-Δ if its true positive and false positive rates with respect to its ability to predict the protected class are approximately (up to Δ deviation) equal at all threshold levels.

Reinterpreted, demographic parity requires that predictions of the classifier cannot reveal information about the protected class any better (up to Δ deviation) than random guessing. DP-Δ is in fact equivalent to requiring that the ROC curve for the classifier $h(x, t)$ in predicting z is within Δ of the *line of no-discrimination*, which is the line that is achievable by biased random guessing. More visually, Figure 1 shows how DP-Δ can be seen using an ROC curve.

3.2 Equal Opportunity

Demographic parity has been criticized as too strict [Dwork et al., 2012, Hardt et al., 2016], and so another notion of fairness has been proposed in which predictions of the label y are independent of the protected class z, when the true label is positive (i.e., $y = +1$). In this definition, we must interpret $y = +1$ as a better label than $y = -1$; for instance, $y = -1$ may be a loan default, while $y = +1$ is full repayment

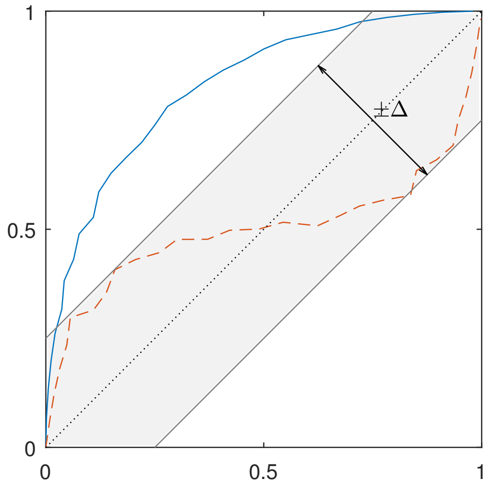


Figure 1: A visual representation of our notion of fairness. Here, the solid blue line is the ROC curve for the y label and the dotted red line the ROC curve for the protected z label. Δ refers to the maximum distance of the latter from the diagonal, which represents a perfect lack of predictability.

of a loan. This definition is typically stated [Hardt et al., 2016] in terms of a single threshold, though it can be generalized to multiple thresholds. We say that a classifier $h(x,t)$ has equal opportunity with level Δ (abbreviated as EO-Δ) if

$$\Big|\mathbb{P}\big[h(x,t)=+1\big|z=+1,y=+1\big]-\\ \mathbb{P}\big[h(x,t)=+1\big|z=-1,y=+1\big]\Big|\leq\Delta,\ \forall t\in\mathbb{R}. \quad (3)$$

To understand this, note $\mathbb{P}\big[h(x,t)=+1\big|z=+1,y=+1\big]$ is the true positive rate conditioned on $y=+1$ when predicting the protected class at threshold t, while $\mathbb{P}\big[h(x,t)=+1\big|z=-1,y=+1\big]$ is the false positive rate conditioned on $y=+1$ when predicting the protected class at threshold t. So the intuition is that a classifier is EO-Δ if its false positive rates and true positive rates are approximately (up to Δ deviation) equal at all threshold levels for the protected class, when conditioned on $y=+1$.

Reinterpreted, equal opportunity requires that predictions of the classifier cannot reveal information about the protected class any better (up to Δ deviation) than random guessing, when the true label is positive. EO-Δ is equivalent to requiring that the ROC curve for the classifier $h(x,t)$ in predicting z conditioned on $y=+1$ is within Δ of the line of no-discrimination. Figure 1 shows how DP-Δ can be seen using an ROC curve.

4 FAIRNESS CONSTRAINTS

In this section, we derive several fairness constraints for linear SVM. The kernel trick is used to convert these constraints for use in kernel SVM. We will focus on presenting formulations for demographic parity, though all of our formulations easily generalize to equal opportunity by simply conditioning on $y=+1$.

4.1 Constraints for Linear SVM

We first study constraints that can be used to ensure fairness with respect to z for a linear SVM

$$h(x,b)=\mathrm{sign}(x^{\mathsf{T}}w+b), \quad (4)$$

where $w\in\mathbb{R}^p$, $b\in\mathbb{R}$ are coefficients to predict the label y. Consider the following generic linear SVM formulation

$$\begin{aligned}\min\ & \textstyle\sum_{i=1}^n u_i+\lambda\|w\|_2^2\\ \text{s.t. }\ & y_i(x_i^{\mathsf{T}}w+b)\geq 1-u_i, && \text{for } i\in[n]\\ & u_i\geq 0, && \text{for } i\in[n]\\ & G(w,\theta)\leq 0.\end{aligned} \quad (5)$$

where $\lambda\in\mathbb{R}$ is a tuning parameter that can be chosen using cross-validation, and $G(w,\theta)\leq 0$ is a fairness constraint with the fairness level controlled by the (possibly vector-valued) parameter θ. We will next consider several possibilities for $G(w,\theta)\leq 0$.

4.1.1 Indicator Constraints

The definition of DP-Δ in (2) uses probabilities, which are are not available as data. Fortunately, we can use empirical fractions of events as an approximation:

$$\Big|\tfrac{1}{\#P}\textstyle\sum_{i\in P}\mathbf{1}(\mathrm{sign}(x_i^{\mathsf{T}}w+t)=+1)-\\ \tfrac{1}{\#N}\textstyle\sum_{i\in N}\mathbf{1}(\mathrm{sign}(x_i^{\mathsf{T}}w+t)=+1)\Big|\leq\Delta,\ \forall t\in\mathbb{R}. \quad (6)$$

Standard arguments using VC dimension [Wainwright, 2017] can be used to show that (6) bounds (3) with high probability. However, (6) is also difficult to include in the linear SVM as the fairness constraint $G(w,\theta)\leq 0$ since it involves the discontinuous and nonconvex sign$(\cdot)$ function, and is infinite-dimensional since it must hold for all $t\in\mathbb{R}$.

4.1.2 Integer Constraints

An initial idea to incorporate (6) in the linear SVM is based on recent empirical results [Miyashiro and Takano, 2015, Bertsimas et al., 2016] where mixed integer programming (MIP) was used to exactly solve cardinality constrained linear regression. Specifically, we can approximate the indicator constraints (6) as the following mixed-integer linear inequalities

$$\begin{aligned} -\Delta \leq \tfrac{1}{\#P}\textstyle\sum_{i\in P} v_i(t) - \tfrac{1}{\#N}\sum_{i\in N} v_i(t) \leq \Delta \\ -M\cdot(1-v_i(t)) \leq x_i^\mathsf{T} w + t \leq M\cdot v_i(t) \\ v_i(t)\in\{0,1\} \end{aligned} \qquad (7)$$

for $t\in\{t_1,\ldots,t_k\}$; where $M>0$ is a large constant, and $\{t_1,\ldots,t_k\}$ is a fixed set of values. This removes the sign$(\cdot)$ function, and it ensures a finite number of constraints. However, we found in numerical experiments using Gurobi [2016] and Mosek [2017] that computing a fair linear SVM with the above constraints was prohibitively slow except for very small data sets.

4.1.3 Convex Relaxation

We next derive a convex relaxation of the indicator constraints (6). Let $M>0$ be a constant, and consider $-M\leq u\leq M$. Then convex upper bounds for the indicators functions are $\mathbf{1}(\text{sign}(u)=+1)\leq 1+u/M$ and $-\mathbf{1}(\text{sign}(u)=+1)\leq -u/M$. Using these upper bounds on (6) leads to the following convex relaxation:

$$-d \leq \big(\tfrac{1}{\#P}\textstyle\sum_{i\in P} x_i - \tfrac{1}{\#N}\sum_{i\in N} x_i\big)^\mathsf{T} w \leq d, \qquad (8)$$

where $d=M\cdot(\Delta-1)$. There are three important remarks about this convex relaxation. The first is that the threshold t does not appear, which means this is simply a linear constraint. The second is that the bound M can be subsumed into the parameter d used to control the fairness level, meaning that this relaxation is practically independent of the bound M. The third is that this convex relaxation is equivalent to the fairness constraint proposed by Zafar et al. [2017], though they derive this using a correlation argument.

4.1.4 Covariance Constraints

The convex relaxation (8) is fairly weak, and so it is relevant to ask whether constraints can be added to (8) to improve the fairness level Δ of the resulting linear SVM. Instead of using convex lifts [Lasserre, 2001, Gouveia et al., 2013, Chandrasekaran and Jordan, 2013] to tighten the above constraint, our approach is inspired by an information-theoretic bound and has an intuitive geometric interpretation.

Specifically, consider the two conditional distributions $D_+=[x|z=+1]$ and $D_-=[x|z=-1]$. Equation (2) can be interpreted as the Kolmogorov-Smirnov (KS) distance between the distributions of the "score" functions $w^\mathsf{T}D_+$ and $w^\mathsf{T}D_-$. Since KS distance is upper-bounded by the total variation distance, Pinsker's inequality [Massart, 2007] implies $\Delta\leq\sqrt{\frac{1}{2}\mathcal{KL}(w^\mathsf{T}D_+\|w^\mathsf{T}D_-)}$, where $\mathcal{KL}(\cdot\|\cdot)$ is the Kullback-Leibler divergence. In the sepcial case where $D_+\sim\mathcal{N}(\mu_+,\Sigma_+)$ and $D_-\sim\mathcal{N}(\mu_-,\Sigma_-)$, we have

$$\Delta\leq\sqrt{\frac{1}{4}\left(\frac{\sigma_-}{\sigma_+}+\frac{1}{\sigma_+}(m_+-m_-)^2+\ln\frac{\sigma_+}{\sigma_-}-1\right)}, \qquad (9)$$

where $m_+=w^\mathsf{T}\mu_+, m_-=w^\mathsf{T}\mu_-, \sigma_+=w^\mathsf{T}\Sigma_+ w$ and $\sigma_-=w^\mathsf{T}\Sigma_- w$ [Kullback, 1997].

Note the above bound is minimized if $m_+=m_-$ and $\sigma_+=\sigma_-$. In fact, (8) can be interpreted as requiring

$$-d\leq\mathbb{E}(D_+)^\mathsf{T}w-\mathbb{E}(D_-)^\mathsf{T}w\leq d, \qquad (10)$$

or that $-d\leq m_+-m_-\leq d$ in the special case above. Subsequently, we propose constraints to control the difference between conditional variances of $x^\mathsf{T}w+t$, which would make σ_+ close to σ_- in the above special case. Let Σ_+ and Σ_- be the sample covariance matrices for D_+ and D_-, respectively. Then the sample variances of $w^\mathsf{T}D_+$ and $w^\mathsf{T}D_-$ are $w^\mathsf{T}\Sigma_+w$ and $w^\mathsf{T}\Sigma_-w$, respectively. So we specify our covariance constraint as

$$-s\leq w^\mathsf{T}(\Sigma_+-\Sigma_-)w\leq s. \qquad (11)$$

To our knowledge, this constraint has not been previously used to improve the fairness of classifiers. Unfortunately, it is nonconvex because $(\Sigma_+-\Sigma_-)$ is symmetric but typically indefinite (i.e., not positive or negative semidefinite). Hence, computing a linear SVM with this constraint requires further development.

One obvious approach is to lift the constraint (11) and then construct a semidefinite programming (SDP) relaxation [Goemans and Williamson, 1995, Luo et al., 2010]. Specifically, note that (11) is equivalent to

$$\begin{aligned} -s\leq\text{trace}(W(\Sigma_+-\Sigma_-))\leq s \\ U=\begin{bmatrix} W & w\\ w^\mathsf{T} & 1\end{bmatrix}\succeq 0 \\ \text{rank}(U)=1 \end{aligned} \qquad (12)$$

The above is nonconvex, but it can be convexified by dropping the rank$(U)=1$ constraint. However, we found in numerical experiments using the Mosek [2017] solver that the SDP relaxation was weak and did not consistently affect the fairness or accuracy of the SVM. Despite this result, we believe that additional convexification techniques [Kocuk et al., 2016, Madani et al., 2017] can be used to strengthen the quality of the SDP relaxation; we leave the problem of how to design a strengthened SDP relaxation for future work.

4.2 Constraints for Kernel SVM

We next briefly present constraints analogous to (8) and (11) that can be used to ensure fairness with respect to z for a kernel SVM

$$h(x,b)=\text{sign}(K(X,x)^\mathsf{T}(Y\circ\alpha)+b), \qquad (13)$$

where $\alpha \in \mathbb{R}^n$ are coefficients to predict the label y, and $b = \frac{1}{\#I}\sum_{i\in I}(y_i - K(X,x_i)^\mathsf{T}(Y\circ\alpha))$ with the set of indices $I = \{i : 0 < \alpha_i < \lambda\}$. Consider the following generic kernel SVM formulation

$$\begin{aligned} \min\ & (Y\circ\alpha)^\mathsf{T}K(X,X)(Y\circ\alpha) - \textstyle\sum_{i=1}^n \alpha_i \\ \text{s.t.}\ & Y^\mathsf{T}\alpha = 0 \\ & 0 \le \alpha_i \le \lambda, \qquad \text{for } i \in [n] \\ & H(w,\theta) \le 0. \end{aligned} \tag{14}$$

where $\lambda \in \mathbb{R}$ is a tuning parameter that can be chosen using cross-validation, and $H(w,\theta) \le 0$ is a fairness constraint with the fairness level controlled by the (possibly vector-valued) parameter θ. We will next consider several possibilities for $H(w,\theta) \le 0$.

First, we note that we may use a convex relaxation on indicator functions akin to that presented in Sections 4.1.2 and 4.1.3 to obtain the linear constraint

$$\begin{aligned} -d \le \tfrac{1}{\#P}\textstyle\sum_{i\in P} K(X,x_i)^\mathsf{T}(Y\circ\alpha) + \\ -\tfrac{1}{\#N}\textstyle\sum_{i\in N} K(X,x_i)^\mathsf{T}(Y\circ\alpha) \le d. \end{aligned} \tag{15}$$

Note again that threshold t does not appear.

Next, our covariance constraint can be rewritten for the kernel SVM by first recalling that the kernel SVM with $K(x,x') = x^\mathsf{T}x'$ generates the same classifier as directly solving a linear SVM. Thus, we have the relationship $w = X^\mathsf{T}(Y\circ\alpha)$ in this special case. Next, observe that

$$\begin{aligned} \Sigma_+ &= \tfrac{1}{\#P}X_+^\mathsf{T}\big(\mathbb{I} - \tfrac{1}{\#P}\mathbf{e}\mathbf{e}^\mathsf{T}\big)X_+ \\ \Sigma_- &= \tfrac{1}{\#N}X_-^\mathsf{T}\big(\mathbb{I} - \tfrac{1}{\#N}\mathbf{e}\mathbf{e}^\mathsf{T}\big)X_- \end{aligned} \tag{16}$$

So if apply the kernel trick to our covariance constraint (11) with the above relationships, then the resulting covariance constraint for kernel SVM becomes

$$-s \le (Y\circ\alpha)^\mathsf{T}(S_+ - S_-)(Y\circ\alpha) \le s, \tag{17}$$

where we have

$$\begin{aligned} S_+ &= \tfrac{1}{\#P}K(X,X_+)\big(\mathbb{I} - \tfrac{1}{\#P}\mathbf{e}\mathbf{e}^\mathsf{T}\big)K(X,X_+)^\mathsf{T} \\ S_- &= \tfrac{1}{\#N}K(X,X_-)\big(\mathbb{I} - \tfrac{1}{\#N}\mathbf{e}\mathbf{e}^\mathsf{T}\big)K(X,X_-)^\mathsf{T}. \end{aligned} \tag{18}$$

Similar to (11), the constraint (17) is a nonconvex quadratic constraint because $(S_+ - S_-)$ is symmetric but typically indefinite.

5 SPECTRAL ALGORITHM

The covariance constraints are conceptually promising, but they result in nonconvex optimization problems. Here, we describe an iterative algorithm to effectively solve the SVM with our covariance constraints.

5.1 Linear SVM

Though we could compute the linear SVM with the covariance constraint (11) using successive linearization, a better approach is possible through careful design of the algorithm: Our key observation regarding (11) is that $(\Sigma_+ - \Sigma_-)$ is symmetric, which means it can be diagonalized by an orthogonal matrix:

$$\Sigma_+ - \Sigma_- = \textstyle\sum_{i=1}^p \zeta_i v_i v_i^\mathsf{T}, \tag{19}$$

where $\zeta_i \in \mathbb{R}$ and the v_i form an orthonormal basis. Now let $I_{\zeta+} = \{i : \zeta_i > 0\}$ and $I_{\zeta-} = \{i : \zeta_i < 0\}$, and define the positive semidefinite matrices

$$\begin{aligned} U_{\zeta+} &= \ \ \textstyle\sum_{i\in I_{\zeta+}} \zeta_i v_i v_i^\mathsf{T} \\ U_{\zeta-} &= -\textstyle\sum_{i\in I_{\zeta-}} \zeta_i v_i v_i^\mathsf{T} \end{aligned} \tag{20}$$

This means that the function

$$w^\mathsf{T}(\Sigma_+ - \Sigma_-)w = w^\mathsf{T}U_{\zeta+}w - w^\mathsf{T}U_{\zeta-}w \tag{21}$$

in the covariance constraint (11) is the difference of the two convex functions $w^\mathsf{T}U_{\zeta+}w$ and $w^\mathsf{T}U_{\zeta-}w$.

There is an important point to note regarding the practical importance of the spectral decomposition we performed above. The function in the covariance constraint (11) can alternatively be written as the difference of the two convex functions $w^\mathsf{T}\Sigma_+w$ and $w^\mathsf{T}\Sigma_-w$. However, using this alternative decomposition yields an algorithm where the convexified subproblems are weaker relaxations than the convex subproblems generated using the spectral decomposition. As a result, our algorithm given below is ultimately more effective because it employs the spectral decomposition.

Consequently, the constrained convex-concave procedure [Tuy, 1995, Yuille and Rangarajan, 2002, Smola et al., 2005] can be used to design an algorithm for our setting. We opt to use a penalized form of the quadratic constraint in our spectral algorithm to ensure feasibility always holds. Let $w_k \in \mathbb{R}^p$ be a fixed point, and consider the optimization problem where the concave terms are linearized:

$$\begin{aligned} \min\ & \textstyle\sum_{i=1}^n u_i + \lambda\|w\|_2^2 + \mu\cdot t \\ \text{s.t.}\ & y_i(x_i^\mathsf{T}w + b) \ge 1 - u_i, \qquad \text{for } i\in[n] \\ & u_i \ge 0, \qquad \text{for } i \in [n] \\ & -d \le \big(\tfrac{1}{\#P}\textstyle\sum_{i\in P} x_i - \tfrac{1}{\#N}\sum_{i\in N} x_i\big)^\mathsf{T} w \le d \\ & w^\mathsf{T}U_{\zeta+}w - w_k^\mathsf{T}U_{\zeta-}w_k - 2w_k^\mathsf{T}U_{\zeta-}^\mathsf{T}(w - w_k) \le t \\ & w^\mathsf{T}U_{\zeta-}w - w_k^\mathsf{T}U_{\zeta+}w_k - 2w_k^\mathsf{T}U_{\zeta+}^\mathsf{T}(w - w_k) \le t \end{aligned} \tag{22}$$

Our spectral algorithm for computing a fair linear SVM consists of the constrained CCP adapted to the problem of computing a linear SVM with the linear

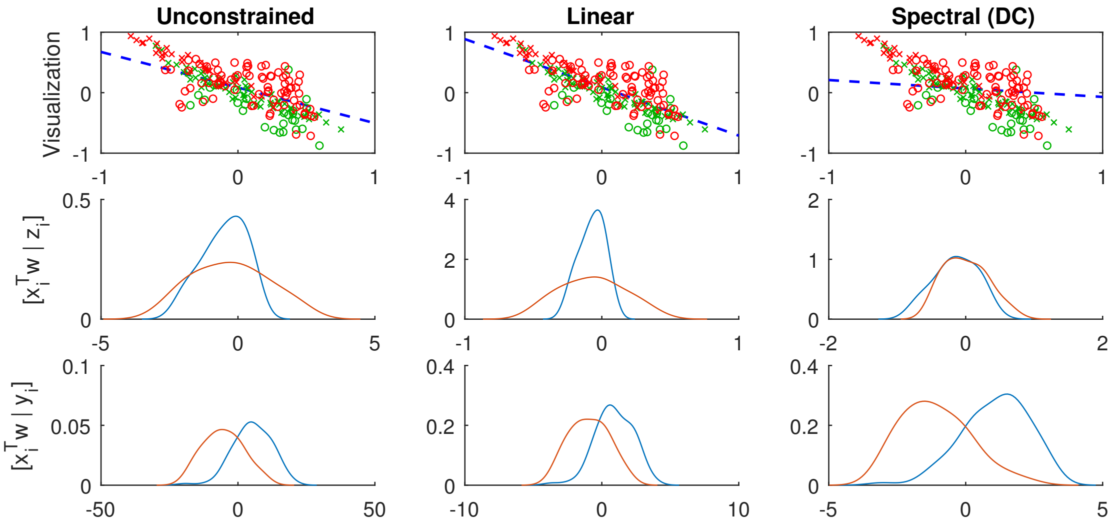


Figure 2: A comparison of the unconstrained, linear and spectral SVM methodologies on two-dimensional data. The first row visualizes the data, as well as the optimal support vectors from each of the methodologies. The second and third rows show the density of $x_i^\intercal w$ conditioned on the z and y variables, respectively. In each case, the blue curve represents the density for $z_i = 1$ ($y_i = 1$) and the red curve the density for $z_i = -1$ ($y_i = -1$).

constraint (8) and covariance constraint (11): We initialize w_0 by solving a linear SVM with only the linear constraint (8), and then compute successive w_k by solving (22). This produces a local minimum.

Theorem 1 (Smola et al. [2005])**.** *The spectral algorithm defined above for computing a fair linear SVM gives iterates w_k that converge to a local minimum.*

This theorem is simply an application of a theorem by Smola et al. [2005], and the constraint qualification required by this theorem trivially holds in our case because all of our convex constraints are linear.

5.2 Kernel SVM

We can apply a similar argument to the kernel SVM formulation by instead defining the decomposition $S_+ - S_- = \sum_{i=1}^{p} \xi_i \nu_i \nu_i^\intercal$, where $\xi_i \in \mathbb{R}$ and the ν_i form an orthonormal basis. Letting $I_{\xi+} = \{i : \xi_i > 0\}$ and $I_{\xi-} = \{i : \xi_i < 0\}$, we may define the positive semidefinite matrices

$$\begin{aligned} U_{\xi+} &= \sum_{i\in I_{\xi+}} \xi_i \nu_i \nu_i^\intercal \\ U_{\xi-} &= -\sum_{i\in I_{\xi-}} \xi_i \nu_i \nu_i^\intercal. \end{aligned} \tag{23}$$

Finally, we may implement the constrained CCP [Tuy, 1995, Yuille and Rangarajan, 2002, Smola et al., 2005] with penalized quadratic term in the same manner as above. We use a penalized form of the quadratic constraint in our spectral algorithm to ensure that feasibility always holds.

Letting $\alpha_k \in \mathbb{R}^n$ be a fixed point, we obtain the analogous kernel formulation

$$\begin{aligned} \min\ & (Y\circ\alpha)^\intercal K(X,X)(Y\circ\alpha) - \textstyle\sum_{i=1}^{n}\alpha_i + \mu\cdot t \\ \text{s.t.}\ & Y^\intercal\alpha = 0 \\ & 0 \le \alpha_i \le \lambda, \qquad \text{for } i\in[n] \\ & -d \le \tfrac{1}{\#P}\textstyle\sum_{i\in P} K(X,x_i)^\intercal(Y\circ\alpha) + \\ & \qquad - \tfrac{1}{\#N}\textstyle\sum_{i\in N} K(X,x_i)^\intercal(Y\circ\alpha) \le d \\ & (Y\circ\alpha)^\intercal U_{\xi+}(Y\circ\alpha) - (Y\circ\alpha_k)^\intercal U_{\xi-}(Y\circ\alpha_k) + \\ & \qquad - 2(Y\circ\alpha_k)^\intercal U_{\zeta-}^\intercal(Y\circ(\alpha-\alpha_k)) \le t \\ & (Y\circ\alpha)^\intercal U_{\xi-}(Y\circ\alpha) - (Y\circ\alpha_k)^\intercal U_{\xi+}(Y\circ\alpha_k) + \\ & \qquad - 2(Y\circ\alpha_k)^\intercal U_{\zeta+}^\intercal(Y\circ(\alpha-\alpha_k)) \le t \end{aligned} \tag{24}$$

We initialize w_0 by solving a kernel SVM with only the linear constraint (15), and then computing successive w_k by solving (24). This produces a local minimum.

Theorem 2 (Smola et al. [2005])**.** *The spectral algorithm defined above for computing a fair kernel SVM gives iterates w_k that converge to a local minimum.*

5.3 Generalization to Linear Learners

Our spectral algorithm can be applied to any linear classifier or linear regression model. For instance, in logistic regression we model the conditional probability of y_i given x_i, as $\mathbb{P}[y_i|x_i] = (1+\exp(-y_i(w^\intercal x_i+b)))^{-1}$. The maximum likelihood estimate (MLE) is computed by solving $\min \sum_{i=1}^{n} \exp(-y_i(w^\intercal x_i + b))$, and we can improve the fairness of the MLE by adding our constraints (8) and (11). This leads to the following formulation for fair logistic regression:

$$\begin{aligned} \min\ & \textstyle\sum_{i=1}^n \exp(-y_i(w^\top x_i + b)) \\ \text{s.t.}\ & -d \le \left(\tfrac{1}{\#P}\textstyle\sum_{i\in P} x_i - \tfrac{1}{\#N}\textstyle\sum_{i\in N} x_i\right)^\top w \le d \\ & -s \le w^\top(\Sigma_+ - \Sigma_-)w \le s \end{aligned} \tag{25}$$

Since (11) is nonconvex, we use our spectral algorithm to find a local minimum of the above problem. Let $w_k \in \mathbb{R}^p$ be a fixed point, and consider the optimization problem where the concave terms are linearized:

$$\begin{aligned} \min\ & \textstyle\sum_{i=1}^n \exp(-y_i(w^\top x_i + b)) + \mu \cdot t \\ \text{s.t.}\ & -d \le \left(\tfrac{1}{\#P}\textstyle\sum_{i\in P} x_i - \tfrac{1}{\#N}\textstyle\sum_{i\in N} x_i\right)^\top w \le d \\ & w^\top U_{\zeta+} w - w_k^\top U_{\zeta-} w_k - 2w_k^\top U_{\zeta-}^\top (w - w_k) \le t \\ & w^\top U_{\zeta-} w - w_k^\top U_{\zeta+} w_k - 2w_k^\top U_{\zeta+}^\top (w - w_k) \le t \end{aligned} \tag{26}$$

where U_+, U_- are as defined in Section 5.1. Our spectral algorithm is: We initialize w_0 by solving the MLE for logistic regression with only the linear constraint (8), and then computing successive w_k by solving (26). This produces a local minimum as discussed above.

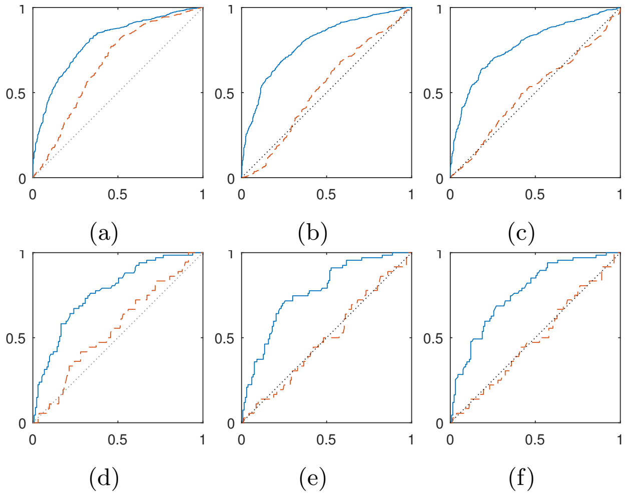
6 NUMERICAL EXPERIMENTS

We use synthetic and real datasets to evaluate the efficacy of our approach. We compare linear SVM's computed using our spectral algorithm (SSVM) to a standard linear SVM (LSVM) and a linear SVM computed using the approach of Zafar et al. [2017] (ZSVM), since this is the only existing approach that to our knowledge is designed to ensure fairness at all thresholds. Numerical implementations are available at `http://github.com/molfat66/FairML`.

6.1 Synthetic Data

Experimental Design We seek to examine the case in which our y and z labels are correlated, and X_N and X_P have differing covariances. Thus, we generate 200 data points where the y and z labels are generated through logit models using two separate sets of randomly generated "true" parameters, with dot product between the logit parameters of y and z of 0.85. The singular values of the covariance matrix of $[X_i|z_i = 1]$ were then skewed to generate the data seen in Figure 2. The empirical correlation of y_i and z_i is 0.45.

Results The results of the three methods using $d = 0.075$ and $\mu = 10$ are shown in the three columns of Figure 2. Here, points in N and P are differentiated by marker shape ("x" and "o", respectively), and points with label $y_i = -1$ and $y_i = 1$ are differentiated by color (red and green, respectively). If w denotes the coefficients computed by each method, then the second row shows the empirical densities of $x_i^\top w$ conditioned on the protected class z_i, and the third row shows the empirical densities of $x_i^\top w$ conditioned on the label y_i. Fairness occurs if the conditional densities in the second row are similar, and prediction accuracy occurs if the densities in the third row are disparate. These results show that the densities of $[x_i^\top w|z_i = +1]$ and $[x_i^\top w|z_i = -1]$ are distinguishable for LSVM and ZSVM, while they are almost identical for SSVM. On the other hand, the densities of $[x_i^\top w|y_i = +1]$ and $[x_i^\top w|y_i = -1]$ are distinct for all three methods.



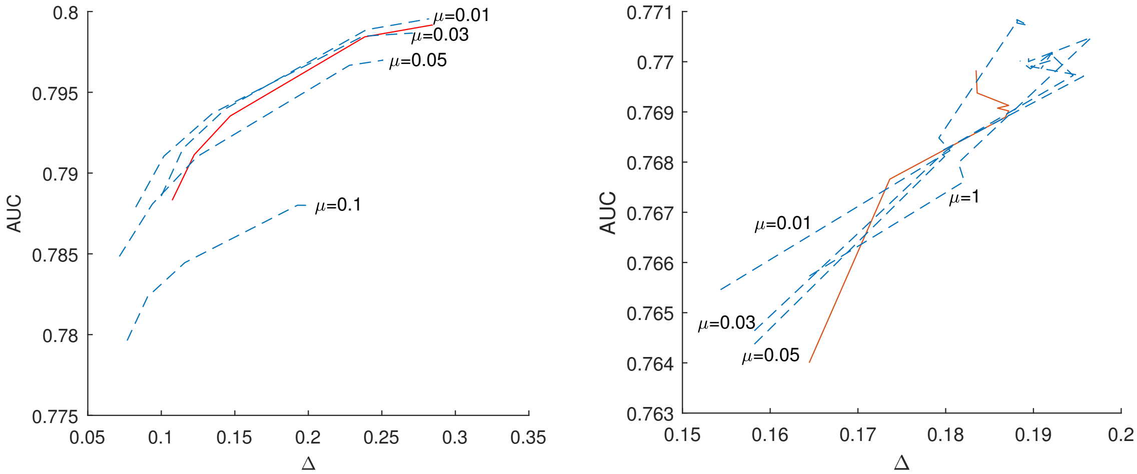
(a) (b) (c)

(d) (e) (f)

Figure 3: ROC plots for the three SVM algorithms on both datasets. In each case, the solid blue line is the ROC curve for the y label and the dotted red line the ROC curve for the protected z label. Figures 3a to 3c show the ROC plots for LSVM, ZSVM, and SSVM on the wine quality data, and Figures 3d to 3f show the same for the German credit data.

6.2 Real World Datasets

Data overview We next use a wine quality dataset [Cortez et al., 2009] and a dataset of German credit card customers [Lichman, 2013]. The first dataset is a compilation of 12 attributes of 6,497 wines (e.g., acidity, residual sugar, alcohol content, and color), as well as a ranking out of 10 that is provided by professional taste-testers. Here, we label $y_i = 1$ when a wine is rated as a 6 or above and $y_i = -1$ otherwise, and we define $z_i = 1$ for white wines and $z_i = -1$ for reds. Notably, all explanatory variables are continuous. The second dataset is a compilation of 20 attributes (e.g., marriage, employment and housing status, number of existing credit lines, and age) of 1000 German applicants for loans. We label $y_i = 1$ for applicants that



(a) Wine quality data.

(b) German credit data.

Figure 4: Comparing the accuracy and fairness of the ZSVM and SSM methods for various d and μ. The solid red line represents results for the ZSVM, and the dotted blue lines denote results for the SSVM for some μ.

defaulted and $y_i = -1$ for applicants that did not default, and let $z_i = 1$ for applicants that are renting a home and $z_i = -1$ for applicants that own their home. Note that a large number of variables are discrete. There is no missing data in either dataset.

Metrics of comparison We compare SSVM and ZSVM based on the tradeoffs that they make between predictive accuracy for y, measured using the area under the ROC curve (AUC), and their fairness with respect z, measured by DP-Δ with respect to z.

Experimental Design We conducted five rounds of cross-validation on each dataset and computed the average AUC and average Δ, using a 70-30 training-testing split. Within each round, we first apply 5-fold cross-validation on LSVM to choose the λ that maximizes AUC, and this value of λ was used with both SSVM and ZSVM to minimize the impact of cross-validation on the comparison between methods. We varied d over the values 0, 0.001, 0.002, 0.005, 0.01, 0.025, 0.05 and 0.1. And for SSVM we tried several values of μ, which are shown in our plots.

Results Figure 3 shows representative examples of ROC curves for both datasets from one instance of cross-validation. Both ZSVM and SSVM improve fairness with respect to LSVM while maintaining high accuracy, and SSVM ensures an even stricter level of fairness than LSVM while keeping high accuracy. The tradeoff curves between prediction accuracy and fairness are shown in Figure 4. Increasing d generally jointly decreases fairness and increases accuracy, while small increases in μ for our SSVM can often improve both fairness and accuracy. Large increases in μ generally increase fairness but decrease accuracy. Note that setting $\mu = 0$ leads to the curve for SSVM to align with the curve of ZSVM, since they are equivalent when $\mu = 0$. Also note Δ is more sensitive than AUC to changes in d, μ, which implies we are able control fairness without losing much predictive accuracy.

7 Conclusion

We considered multi-threshold notions of fairness for classifiers, and designed a nonconvex constraint to improve the fairness of linear and kernel SVM's under all thresholds. We developed an iterative optimization algorithm (that uses a spectral decomposition) to handle our nonconvex constraint in the resulting problem to compute the SVM, and empirically compared our approach to standard linear SVM and an SVM with a linear fairness constraint using both synthetic and real data. We found that our method can strictly improve the fairness of classifiers for all thresholding values with little loss in accuracy; in fact, some of our results even showed a slight increase in accuracy with increasing fairness. Our work opens the door for further research in a number of areas, including hierarchies of fairness constraints considering subsequent moments of the data, and theoretical guarantees on the fairness of such classification methods.

Acknowledgements

This work was supported by the UC Berkeley Center for Long-Term Cybersecurity.

References

S. Barocas and A. D. Selbst. Big data's disparate impact. *California Law Review*, 104, 2016.

D. Bertsimas, A. King, and R. Mazumder. Best subset selection via a modern optimization lens. *Annals of Statistics*, 44(2):813–852, 2016.

E. Bhandari. Big data can be used to violate civil rights laws, and the FTC agrees, 2016.

T. Calders, F. Kamiran, and M. Pechenizkiy. Building classifiers with independency constraints. In *Data mining workshops, 2009. ICDMW'09. IEEE international conference on*, pages 13–18. IEEE, 2009.

R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1721–1730. ACM, 2015.

V. Chandrasekaran and M. I. Jordan. Computational and statistical tradeoffs via convex relaxation. *Proceedings of the National Academy of Sciences*, 110 (13):E1181–E1190, 2013.

P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553, 2009.

A. Cotter, M. Friedlander, G. Goh, and M. Gupta. Satisfying real-world goals with dataset constraints. *arXiv preprint arXiv:1606.07558*, 2016.

C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226. ACM, 2012.

S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236*, 2016.

M. X. Goemans and D. P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM (JACM)*, 42(6):1115–1145, 1995.

J. Gouveia, P. A. Parrilo, and R. R. Thomas. Lifts of convex sets and cone factorizations. *Mathematics of Operations Research*, 38(2):248–264, 2013.

Gurobi. Gurobi optimizer reference manual, 2016. URL `http://www.gurobi.com`.

M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016.

J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

B. Kocuk, S. S. Dey, and X. A. Sun. Inexactness of sdp relaxation and valid inequalities for optimal power flow. *IEEE Transactions on Power Systems*, 31(1): 642–651, 2016.

S. Kullback. *Information theory and statistics*. Courier Corporation, 1997.

J. B. Lasserre. Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization*, 11(3):796–817, 2001.

M. Lichman. UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`.

Y. Lou, R. Caruana, and J. Gehrke. Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 150–158. ACM, 2012.

C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

Z.-Q. Luo, W.-K. Ma, A. M.-C. So, Y. Ye, and S. Zhang. Semidefinite relaxation of quadratic optimization problems. *IEEE Signal Processing Magazine*, 27(3):20–34, 2010.

R. Madani, S. Sojoudi, G. Fazelnia, and J. Lavaei. Finding low-rank solutions of sparse linear matrix inequalities using convex optimization. *SIAM Journal on Optimization*, 27(2):725–758, 2017.

P. Massart. *Concentration inequalities and model selection*, volume 6. Springer, 2007.

R. Miyashiro and Y. Takano. Mixed integer second-order cone programming formulations for variable selection in linear regression. *European Journal of Operational Research*, 247(3):721–731, 2015.

Mosek. The MOSEK optimization toolbox for MATLAB manual, 2017.

J. Podesta, P. Pritzker, E. Moniz, J. Holdren, and J. Zients. *Big data: Seizing opportunities, preserving values*. Executive Office of the President, 2014.

G. Ridgeway, D. Madigan, T. Richardson, and J. O'Kane. Interpretable boosted naïve bayes classification. In *KDD*, pages 101–104, 1998.

A. J. Smola, S. Vishwanathan, and T. Hofmann. Kernel methods for missing variables. In *AISTATS*, 2005.

H. Tuy. Dc optimization: theory, methods and algorithms. In *Handbook of global optimization*, pages 149–216. Springer, 1995.

M. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint.* 2017.

A. L. Yuille and A. Rangarajan. The concave-convex procedure (cccp). In *Advances in neural information processing systems*, pages 1033–1040, 2002.

M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 2017.

R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 325–333, 2013.

I. Zliobaite. On the relation between accuracy and fairness in binary classification. *arXiv preprint arXiv:1505.05723*, 2015.