Fair Classification and Social Welfare

Lily Hu
lilyhu@g.harvard.edu
Harvard University
Cambridge, MA

Yiling Chen
yiling@seas.harvard.edu
Harvard University
Cambridge, MA

ABSTRACT

Now that machine learning algorithms lie at the center of many important resource allocation pipelines, computer scientists have been unwittingly cast as partial social planners. Given this state of affairs, important questions follow. How do leading notions of fairness as defined by computer scientists map onto longer-standing notions of social welfare? In this paper, we present a welfare-based analysis of fair classification regimes. Our main findings assess the welfare impact of fairness-constrained empirical risk minimization programs on the individuals and groups who are subject to their outputs. We fully characterize the ranges of $\Delta\epsilon$ perturbations to a fairness parameter ϵ in a fair Soft Margin SVM problem that yield better, worse, and neutral outcomes in utility for individuals and by extension, groups. Our method of analysis allows for fast and efficient computation of "fairness-to-welfare" solution paths, thereby allowing practitioners to easily assess whether and which fair learning procedures result in classification outcomes that make groups better-off. Our analyses show that applying stricter fairness criteria codified as parity constraints can worsen welfare outcomes for *both* groups. More generally, always preferring "more fair" classifiers does not abide by the Pareto Principle—a fundamental axiom of social choice theory and welfare economics. Recent work in machine learning has rallied around these notions of fairness as critical to ensuring that algorithmic systems do not have disparate negative impact on disadvantaged social groups. By showing that these constraints often fail to translate into improved outcomes for these groups, we cast doubt on their effectiveness as a means to ensure fairness and justice.

ACM Reference Format:
Lily Hu and Yiling Chen. 2020. Fair Classification and Social Welfare. In *Conference on Fairness, Accountability, and Transparency (FAT* '20), January 27–30, 2020, Barcelona, Spain.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3351095.3372857

1 INTRODUCTION

In his 1979 Tanner Lectures, Amartya Sen noted that since nearly all egalitarian theories are founded on an equality of *some* sort, the heart of the issue rests on clarifying the "equality of what?" problem [1]. The field of fair machine learning has not escaped this essential question. Does machine learning have an obligation to assure probabilistic equality of outcomes across various social groups [2, 3]? Or does it simply owe an equality of treatment [4]? Does fairness demand that individuals (or groups) be subject to equal mistreatment rates [5, 6]? Or does being fair refer only to avoiding some intolerable level of algorithmic error?

Currently, the task of accounting for fair machine learning cashes out in the comparison of myriad metrics—probability distributions, error likelihoods, classification rates—sliced up every way possible to reveal the range of inequalities that may arise before, during, and after the learning process. But as shown in work by Chouldechova [7] and Kleinberg et al. [8], fundamental statistical incompatibilities rule out any solution that can satisfy all parity metrics. Fairness-constrained loss minimization offers little guidance on its own for choosing among the fairness desiderata, which appear incommensurable and result in different impacts on different individuals and groups. We are thus left with the harsh but unavoidable task of adjudicating between these measures and methods. How ought we decide? For a given application, who actually benefits from the operationalization of a certain fairness constraint? This is a basic but critical question that must be answered if we are to understand the impact that fairness constraints have on classification outcomes. Much research in fairness has been motivated by the well-documented negative impacts that these systems can have on already structurally disadvantaged groups. But do fairness constraints as currently formulated in fact earn their reputation as serving to improve the welfares of marginalized social groups?

When algorithms are adopted in social environments—consider, for example, the use of predictive systems in the financial services industry—classification outcomes directly bear on individuals' material well-beings. We, thus, view predictions as *resource allocations* awarded to individuals and by extension, to various social groups. In this paper, we build out a method of analysis that takes in generic fair learning regimes and analyzes them from a welfare perspective.

Our main contributions, presented in Section 3, are methodological as well as substantive in the field of algorithmic fairness. We show that how "fair" a classifier is—how well it accords with a group parity constraint such as "equality of opportunity" or "balance for false positives"—does not neatly translate into statements about different groups' welfares are affected. Drawing on techniques from parametric programming and finding a SVM's regularization path, our method of analysis finds the optimal ϵ-fair Soft-Margin SVM solution for all values of a fairness tolerance parameter $\epsilon \in [0, 1]$. We track the welfares of individuals and groups as a function of ϵ and identify those ranges of ϵ values that support solutions that are Pareto-dominated by neighboring ϵ values. Further, the algorithmic implementation of our analyses is computationally efficient, with a complexity on the same order as current standard SVM solvers that fit a single SVM model, and is thus practical as a procedure that translates fairness constraints into welfare effects for all ϵ.

FAT '20, January 27–30, 2020, Barcelona, Spain*

ACM ISBN 978-1-4503-6936-7/20/02…$15.00
https://doi.org/10.1145/3351095.3372857

Our substantive results show that a classifier that abides by a stricter fairness standard does not necessarily issue improved outcomes for the disadvantaged group. In particular, we prove two results: first, starting at any nonzero ϵ-fair optimal SVM solution, we express the range of $\Delta\epsilon < 0$ perturbations that tighten the fairness constraint and lead to classifier-output allocations that are weakly Pareto dominated by those issued by the "less fair" original classifier. Second, there are nonzero ϵ-fair optimal SVM solutions, such that there exist $\Delta\epsilon < 0$ perturbations that yield classifications that are strongly Pareto dominated by those issued by the "less fair" original classifier. We demonstrate these findings on the Adult dataset. In general, our results show that when notions of fairness rest entirely on leading parity-based notions, always preferring more fair machine learning classifiers does not accord with the Pareto Principle, an axiom typically seen as fundamental in social choice theory and welfare economics.

The purposes of our paper are twofold. The first is simply to encourage a welfare-centric understanding of algorithmic fairness. Whenever machine learning is deployed within important social and economic processes, concerns for fairness arise when societal ideals are in tension with a decision-maker's interests. Most leading methodologies have focused on optimization of utility or welfare to the vendor but have rarely awarded those individuals and groups who are subject to these systems the same kind of attention to welfare effects. Our work explicitly focuses its analysis on the latter.

We also seek to highlight the limits of conceptualizing fairness only in terms of group-based parity measures. Our results show that at current, making a system "more fair" as defined by popular metrics can harm the vulnerable social populations that were ostensibly meant to be served by the imposition of such constraints. Though the Pareto Principle is not without faults, the frequency with which "more fair" classification outcomes are welfare-wise dominated by "less fair" ones occurs is troublesome and should lead scholars to reevaluate popular methodologies by which we understand the impact of machine learning on different social populations.

1.1 Related Work

Research in fair machine learning has largely centered on computationally defining "fairness" as a property of a classifier and then showing that techniques can be invented to satisfy such a notion [2–5, 5, 6, 9–18]. Since most methods are meant to apply to learning problems generally, many such notions of fairness center on parity-based metrics about a classifier's behavior on various legally protected social groups rather than on matters of welfare.

Most of the works that do look toward a welfare-based framework for interpreting appeals to fairness sit at the intersection of computing and economics. Mullainathan [19] also makes a comparison between policies as set by machine learning systems and policies as set by a social planner. He argues that systems that make explicit their description of a global welfare function are less likely to perpetrate biased outcomes and are more successful at ameliorating social inequities. Heidari et al. [20] propose using social welfare functions as fairness constraints on loss minimization programs. They suggest that a learner ought to optimize her classifier while in Rawls' original position. As a result, their approach to social welfare is closely tied with considerations of risk. Rather than integrate social welfare functions into the supervised learning pipeline, we claim that the result of an algorithmic classification system can itself be considered a welfare-impacting allocation. Thus, our work simply takes a generic ϵ-fair learning problem as-is, and then considers the welfare implications of its full path of outcomes for all $\epsilon \in [0, 1]$ on individuals as well as groups. Attention to the potential harms of machine learning systems is not new, of course. Within the fairness literature, Corbett-Davies & Goel [21] and Liu et al. [22] devote most of their analyses to the person-impacting effects of algorithmic systems. We agree that these effects are relevant to the question of fairness, but our results differ in their methodological focus: we introduce a technique that derives the full range of welfare effects achieved by a fair classification algorithm.

The techniques that we use to translate fair learning outcomes into welfare paths are related to a number of existing works. The proxy fairness constraint in our instantiation of the ϵ-fair SVM problem original appeared in Zafar et al.'s work on restricting the disparate impact of machine classifiers [5]. Their research introduces this particular proxy fairness constrained program and shows that it can be efficiently solved and well approximates target fairness constraints. We use the constraint to demonstrate our overall findings about the effect of fairness criteria on individual and group welfares. We share some of the preliminary formulations of the fair SVM problem with Donini et al. [17] though they focus on the statistical and fairness guarantees of the generalized ERM program. Lastly, though work on tuning hyperparameters of SVMs and the solution paths that result seem far afield from questions of fairness and welfare, our analysis on the effect of $\Delta\epsilon$ fairness perturbations on welfare take advantage of methods in that line of work [23–27].

2 PROBLEM FORMALIZATION

Our framework and results are motivated by those algorithmic use cases in which considerations of fairness and welfare stand alongside those of efficiency. Because our paper connects machine classification and notions of algorithmic fairness with conceptions of social welfare, we first provide an overview of the notation and assumptions that feature throughout our work.

In the empirical loss minimization problem, a learner seeks a classifier h that issues the most accurate predictions when trained on set of n data points $\{\mathbf{x}_i, z_i, y_i\}_{i=1}^n$. Each triple gives an individual's feature vector $\mathbf{x}_i \in \mathcal{X}$, protected class attribute $z_i \in \{0, 1\}$, and true label $y_i \in \{-1, +1\}$.[1] A classifier that assigns an incorrect label $h(\mathbf{x}_i) \neq y_i$ incurs a penalty.

The empirical risk minimizing predictor is given by

$$h^* := \arg\min_{h \in \mathcal{H}} \sum_{i=1}^{n} \ell(h(\mathbf{x_i}), y_i)$$

where hypothesis $h : \mathcal{X} \to \mathbb{R}$ gives a learner's model, the loss function $\ell : \mathbb{R} \times \{-1, +1\} \to \mathbb{R}$ gives the penalty incurred by a prediction, and $\mathcal{H}$ is the hypothesis class under the learner's consideration. Binary classification systems issue predictions $h(\mathbf{x}) \in \{-1, +1\}$.

Notions of fairness have been formalized in a variety of ways in the machine learning literature. Though Dwork et al.'s [4] initial conceptualization remains prominent and influential, much work

[1] Though individuals in a dataset will typically be coded with many protected class attributes, in this paper we will consider only a single sensitive attribute of focus.

has since defined fairness as a parity notion applied across different protected class groups [3, 5, 7, 8, 17, 18]. The following definition gives the general form of these types of fairness criteria.

Definition 2.1. A classifier h satisfies a general group-based notion of *ϵ-fairness* if

$$|\mathbb{E}[g(\ell, h, \mathbf{x}_i, y_i)|\mathcal{E}_{\mathbf{z}_i=1}] - \mathbb{E}[g(\ell, h, \mathbf{x}_i, y_i)|\mathcal{E}_{\mathbf{z}_i=0}]| \leq \epsilon \qquad (1)$$

where g is some function of classifier h performance, and $\mathcal{E}_{\mathbf{z}_i=0}$ and $\mathcal{E}_{\mathbf{z}_i=1}$ are events that occur with respect to groups $z = 0$ and $z = 1$ respectively.

Further specifications of the function g and the events $\mathcal{E}$ instantiate particular group-based fairness notions. For example, when $g(\ell, h, \mathbf{x}_i, y_i) = h(\mathbf{x}_i)$ and $\mathcal{E}_{\mathbf{z}_i}$ refers to the events in which $y_i = +1$ for each group $\mathbf{z}_i \in \{0, 1\}$, Definition 2.1 gives an ϵ-approximation of *equality of opportunity* [3]. When $g(\ell, h, \mathbf{x}_i, y_i) = \ell(h(\mathbf{x}_i), y_i)$ and $\mathcal{E}_{\mathbf{z}_i}$ refers to all classification events for each group $\mathbf{z}_i$, Definition 2.1 gives the notion of ϵ-approximation of *overall error rate balance* [7]. Notice that as ϵ increases, the constraint loosens, and the solution is considered "less fair." As ϵ decreases, the fairness constraint becomes more strict, and the solution is considered "more fair."

Mapping classification outcomes to changes in individuals' welfares gives a useful method of analysis for many data-based algorithmic systems that are involved in resource distribution pipelines. In particular, we consider tools that issue outcomes uniformly ranked, or preferred, by those individuals who are the subjects of the system. That is, individuals agree on which outcome is preferred. Examples of such systems abound: applicants for credit generally want to be found eligible; candidates for jobs generally want to be hired, or at least ranked highly in their pool. These realms are precisely those in which fairness considerations are urgent and where fairness-adjusted learning methods are most likely to be adopted.

3 WELFARE IMPACTS OF FAIRNESS CONSTRAINTS

The central inquiry of our work asks how fairness constraints as popularized in the algorithmic fairness community relate to welfare-based analyses that are dominant in economics and policy-making circles. Do fairness-adjusted optimization problems actually make marginalized groups better-off in terms of welfare? In this section, we work from an empirical risk minimization (ERM) program with generic fairness constraints parametrized by a tolerance parameter $\epsilon > 0$ and trace individuals' and groups' welfares as a function of ϵ. We assume that an individual benefits from receiving a positive classification, and thus we define group welfare as

$$W_k = \frac{1}{n_k} \sum_{i|z_i=k} \frac{h(\mathbf{x}_i) + 1}{2}, \qquad k \in \{0, 1\} \qquad (2)$$

where n_k give the number of individuals in group $z = k$. We note that W_k can be defined in ways other than (2), which assumes that positive classification are always and only welfare-enhancing. Other work has considered the possibility that positive classifications may in fact make individuals worse-off if they are false positives [22]. The definition of W_k can be generalized to account for these cases.

First, in Section 3.1, we present an instantiation of the ϵ-fair ERM problem with a fairness constraint proposed in prior work in algorithmic fairness. We work from the Soft-Margin SVM program and derive the various dual formulations that will be of use in the following analyses. In Section 3.2, we move on to show how $\Delta\epsilon$ perturbations to the fairness constraint in the ϵ-fair ERM problem yield changes in classification outcomes for individuals and by extension, how they impact a group's overall welfare. Our approach, which draws a connection between fairness perturbations and searches for an optimal SVM regularization parameter, tracks changes in an individual's classification by taking advantage of the codependence of variables in the dual of the SVM. By perturbing the fairness constraint, we observe changes in not its own corresponding dual variable but in the corresponding dual of the margin constraints, which relay the classification fates of data points.

Leveraging this technique, we plot the "solution paths" of the dual variable as a function of ϵ, which in turn allows us to compute group welfares as a function of ϵ and draw out substantive results on the dynamics of how classification outcomes change in response to ϵ-fair learning. We prove that stricter fairness standards do not necessarily support welfare-enhancing outcomes for the disadvantaged group. In many such cases, the learning goal of ensuring group-based fairness is incompatible with the Pareto Principle.

Definition 3.1 (Pareto Principle). Let x, y be two social alternatives. Let $\succeq_i$ be the preference ordering of individuals $i \in [n]$, and $\succeq_P$ be the preference ordering of a social planner. The planner abides by the *Pareto Principle* if $x \succeq_P y$ whenever $x \succeq_i y$ for all i.

In welfare economics, the Pareto Principle is a standard requirement of social welfare functionals—it would appear that the selection of an allocation that is Pareto dominated by an available alternative would be undesirable and even irresponsible! Nevertheless, we show that applying fairness criteria to loss minimization tasks in some cases do just that. We perform our analysis on the Soft-Margin SVM optimization problem and, for concreteness, work with a well-known fairness formulation in the literature. However, we note that our methods and results apply to fairness-constrained convex loss minimization programs more generally.

We also show that this method of analysis can form practical tools. In Section 3.3, we present a computationally efficient algorithmic implementation of our analyses, fitting full welfare solution paths for all $\epsilon \in [0, 1]$ values in a time complexity that is on the same order as that of a single SVM fit. We close this section by working from the shadow price of the fairness constraint to derive local and global sensitivities of the optimal solution to $\Delta\epsilon$ perturbations.

3.1 Setting up the ϵ-fair ERM program

The general fairness-constrained empirical loss minimization program can be written as

$$\begin{aligned} \underset{h \in \mathcal{H}}{\text{minimize}} \quad & \ell(h(\mathbf{x}), y) \\ \text{subject to} \quad & f_h(\mathbf{x}, y) \leq \epsilon \end{aligned} \qquad (3)$$

where $\ell(h(\mathbf{x}), y)$ gives the empirical loss of a classifier $h \in \mathcal{H}$ on the dataset $\mathcal{X}$. To maximize accuracy, the learner ought to minimize 0-1 loss; however because the loss function ℓ_{0-1} is non-convex, a convex surrogate loss such as hinge loss (ℓ_h) or log loss ($\ell_{\log}$) is frequently substituted in its place to ensure that globally optimal solutions may be efficiently found. $f_h(\mathbf{x}, y) \leq \epsilon$ gives a group-based fairness constraint of the type given in Definition 2.1, where $\epsilon > 0$ is the unfairness "tolerance parameter"—a greater ϵ permits a greater group disparity on a metric of interest; a smaller ϵ more tightly restricts the level of permissible disparity.

We examine the behavior of fairness-constrained linear SVM classifiers, though we note that our techniques generalize to nonlinear kernels SVMs, since interpretations of the dual of the SVM and the full SVM regularization path are the same with kernels [24]. Our learner minimizes hinge loss with L_1 regularization; equivalently, she seeks a Soft-Margin SVM that is "ϵ-fair." Both SVM models and "fair training" approaches are in broad circulation. The fair empirical risk minimization program is thus given as

$$\begin{aligned} \underset{\boldsymbol{\theta}, b}{\text{minimize}} \quad & \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C\sum_{i=1}^{n}\xi_i \\ \text{subject to} \quad & y_i(\boldsymbol{\theta}^\top \mathbf{x}_i + b) - 1 + \xi_i \geq 0, \qquad (\epsilon\text{-fair Soft-SVM}) \\ & \xi_i \geq 0, \\ & f_{\boldsymbol{\theta},b}(\mathbf{x}, y) \leq \epsilon \end{aligned}$$

where the learner seeks SVM parameters $\boldsymbol{\theta}, b$; ξ_i are non-negative slack variables that violate the margin constraint in the Hard-Margin SVM problem $y_i(\boldsymbol{\theta}^\top \mathbf{x}_i + b) - 1 \geq 0$, and $C > 0$ is a hyperparameter tunable by the learner to express the trade-off between preferring a larger margin and penalizing violations of the margin. $f_{\boldsymbol{\theta},b}(\mathbf{x}, y)$ is the group parity-based fairness constraint.

The abundant literature on algorithmic fairness presents a long menu of options for the various forms that $f_{\boldsymbol{\theta},b}$ could take, but generally speaking, the constraints are non-convex. As such, much work has enlisted methods that depart from directly pursuing efficient constraint-based convex programming techniques in order to solve them [5, 6, 9, 16, 18]. Researchers have also devised convex proxy alternatives, which have been shown to approximate the intended outcomes of original fairness constraints well [5, 17, 28]. In particular, in this paper, we work with the proxy constraint proposed by Zafar et al. [5], which constrains disparities in covariance between group membership and the (signed) distance between individuals' feature vectors and the hyperplane decision boundary:

$$f_{\boldsymbol{\theta},b}(\mathbf{x}, y) = \left|\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})(\boldsymbol{\theta}^\top \mathbf{x}_i + b)\right| \leq \epsilon \tag{4}$$

$\bar{z}$ reflects the bias in the demographic makeup of $\mathcal{X}$: $\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$. Let ($\epsilon$-fair-SVM1-P) be the Soft-Margin SVM program with this covariance constraint. The corresponding Lagrangian is

$$\begin{aligned} &\mathcal{L}_P(\boldsymbol{\theta}, b, \boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \gamma_1, \gamma_2) = \\ &\qquad \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\lambda_i - \sum_{i=1}^{n}\mu_i(y_i(\boldsymbol{\theta}^\top \mathbf{x}_i + b) - 1 + \xi_i) \\ &\qquad - \gamma_1\Big(\epsilon - \frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})(\boldsymbol{\theta}^\top \mathbf{x}_i + b)\Big) \qquad (\epsilon\text{-fair-SVM1-L}) \\ &\qquad - \gamma_2\Big(\epsilon - \frac{1}{n}\sum_{i=1}^{n}(\bar{z} - z_i)(\boldsymbol{\theta}^\top \mathbf{x}_i + b)\Big) \end{aligned}$$

where $\boldsymbol{\theta} \in \mathbb{R}^d, b \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^n$ are primal variables. The (non-negative) Lagrange multipliers $\boldsymbol{\lambda}, \boldsymbol{\mu} \in \mathbb{R}^n$ correspond to the n non-negativity constraints $\xi_i \geq 0$ and the margin-slack constraints $y_i(\boldsymbol{\theta}^\top \mathbf{x}_i + b) - 1 + \xi_i \geq 0$ respectively. The multipliers $\gamma_1, \gamma_2 \in \mathbb{R}$ correspond to the two linearized forms of the absolute value fairness constraint. By complementary slackness, dual variables reveal information about the satisfaction or violation of their corresponding constraints. The analyses in the subsequent two subsections will focus on these interpretations.

By the Karush-Kuhn-Tucker (KKT) conditions, at the solution of the convex program, the gradients of $\mathcal{L}$ with respect to $\boldsymbol{\theta}$, b, and ξ_i are zero. Plugging in these conditions, the dual Lagrangian is

$$\mathcal{L}_D(\boldsymbol{\mu}, \gamma) = -\frac{1}{2}\left\|\sum_{i=1}^{n}\mu_i y_i \mathbf{x}_i - \frac{\gamma}{n}\sum_{i=1}^{n}(z_i - \bar{z})\mathbf{x}_i\right\|^2 + \sum_{i=1}^{n}\mu_i - |\gamma|\epsilon \tag{5}$$

where $\gamma = \gamma_1 - \gamma_2$. The dual maximizes this objective subject to the constraints $\mu_i \in [0, C]$ for all $i \in [n]$ and $\sum_{i=1}\mu_i y_i = 0$. We thus arrive at the Wolfe dual problem

$$\begin{aligned} \underset{\boldsymbol{\mu}, \gamma, V}{\text{maximize}} \quad & -\frac{1}{2}\left\|\sum_{i=1}^{n}\mu_i y_i \mathbf{x}_i - \frac{\gamma}{n}\sum_{i=1}^{n}(z_i - \bar{z})\mathbf{x}_i\right\|^2 + \sum_{i=1}^{n}\mu_i - V\epsilon \\ \text{subject to} \quad & \mu_i \in [0, C], \qquad i = 1, \ldots, n, \qquad (\epsilon\text{-fair-SVM1-D}) \\ & \sum_{i=1}^{n}\mu_i y_i = 0, \\ & \gamma \in [-V, V] \end{aligned}$$

where we have introduced the variable V to eliminate the absolute value function $|\gamma|$ in the objective. Notice that when $\gamma = 0$ and neither of the constraints bind, we recover the standard dual SVM program. Since we are concerned with fair learning that does alter an optimal solution, we consider cases where V is strictly positive. We introduce additional dual variables β_- and β_+, corresponding to the $\gamma \in [-V, V]$ constraint and derive the Lagrangian

$$\begin{aligned} \mathcal{L}(\boldsymbol{\mu}, \gamma, V, \beta_-, \beta_+) = &-\frac{1}{2}\left\|\sum_{i=1}^{n}\mu_i y_i \mathbf{x}_i - \frac{\gamma}{n}\sum_{i=1}^{n}(z_i - \bar{z})\mathbf{x}_i\right\|^2 + \sum_{i=1}^{n}\mu_i \\ &- V\epsilon + \gamma(\beta_- - \beta_+) + V(\beta_- + \beta_+) \end{aligned}$$

Under KKT conditions, $\beta_- + \beta_+ = \epsilon$ and

$$\gamma^* = \frac{n(n(\beta_- - \beta_+) + \sum_{i=1}^{n}\mu_i y_i\langle \mathbf{x}_i, \mathbf{u}\rangle)}{\|\mathbf{u}\|^2} \tag{6}$$

where $\mathbf{u} = \sum_{i=1}^{n}(z_i - \bar{z})\mathbf{x}_i$ geometrically gives some group-sensitive "average" of $\mathbf{x} \in \mathcal{X}$. We can now rewrite ($\epsilon$-fair-SVM1-D) as

$$\begin{aligned} \underset{\boldsymbol{\mu}, \beta_-, \beta_+}{\text{maximize}} \quad & -\frac{1}{2}\left\|\sum_{i=1}^{n}\mu_i y_i(I - P_{\mathbf{u}})\mathbf{x}_i\right\|^2 + \sum_{i=1}^{n}\mu_i \\ & + \frac{2n\sum_i \mu_i y_i\langle \mathbf{x}_i, \mathbf{u}\rangle + n^2(\beta_- - \beta_+)}{2\|\mathbf{u}\|^2}(\beta_- - \beta_+) \\ \text{subject to} \quad & \mu_i \in [0, C], \qquad i = 1, \ldots, n, \\ & \sum_{i=1}^{n}\mu_i y_i = 0, \qquad (\epsilon\text{-fair SVM2-D}) \\ & \beta_-, \beta_+ \geq 0, \\ & \beta_- + \beta_+ = \epsilon \end{aligned}$$

where $I, P_{\mathbf{u}} \in \mathbb{R}^{d\times d}$. The former is the identity matrix, and the latter is the projection matrix onto the vector $\mathbf{u}$. As was also observed by Donini et al., the $\epsilon = 0$ version of (ϵ-fair SVM2-D) is equivalent to the standard formulation of the dual SVM program with Kernel $K(\mathbf{x}_i, \mathbf{x}_j) = \langle (I - P_{\mathbf{u}})\mathbf{x}_i, (I - P_{\mathbf{u}})\mathbf{x}_j \rangle$ [17].

Since we are interested in the welfare impacts of fair learning when fairness constraints *do* have an impact on optimal solutions, we will assume that the fairness constraint binds. For clarity of exposition, we assume that the positive covariance constraint binds, and thus that $\beta_- = 0$ and $\beta_+ = \epsilon$ in (ϵ-fair SVM2-D). This is without loss of generalization—the same analyses apply when the negative covariance constraint binds. The dual ϵ-fair SVM program becomes

$$\begin{aligned} \underset{\boldsymbol{\mu}}{\text{minimize}} \quad & \frac{1}{2}\left\|\sum_{i=1}^{n} \mu_i y_i (I - P_{\mathbf{u}})\mathbf{x}_i\right\|^2 - \sum_{i=1}^{n} \mu_i + \frac{n\epsilon(2\sum_i \mu_i y_i \langle \mathbf{x}_i, \mathbf{u}\rangle - n\epsilon)}{2\|\mathbf{u}\|^2} \\ \text{subject to} \quad & \mu_i \in [0, C], \qquad i = 1, \ldots, n, \qquad (\epsilon\text{-fair SVM-D}) \\ & \sum_{i=1}^{n} \mu_i y_i = 0 \end{aligned}$$

We will work from this formulation of the constrained optimization problem for the remainder of the paper.

3.2 Impact of Fair Learning on Individuals' Welfares

We now move on to investigate the effects of perturbing a fixed ϵ-fair SVM by some $\Delta\epsilon$ on the classification outcomes that are issued. We ask, *"How are individuals and groups' classifications, and thus their welfares, impacted when a learner tightens or loosens a fairness constraint?"* The key insight that drives our methods and results is that rather than perform sensitivity analysis directly on the dual variable corresponding to the fairness constraint—which, as we will see in Section 3.4, only gives information about the change in the learner's objective value—we track changes in the classifier's behavior by analyzing the effect of $\Delta\epsilon$ perturbations on another set of dual variables: μ_i that correspond to the primal margin constraints. Each of these n dual variables indicate whether its corresponding vector $\mathbf{x}_i$ is correctly classified, lies in the margin, or incorrectly classified. Leveraging how these μ_i change as a function fo ϵ thereby allows us to track the solution paths of individual points and by extension, compute group welfare paths.

Define a function $p(\epsilon) : \mathbb{R} \to \mathbb{R}$ that gives the optimal value of the ϵ-fair loss minimizing program in (ϵ-fair SVM1-P), which by duality is also the optimal value of (ϵ-fair SVM-D). We begin at a solution $p(\epsilon)$ and consider changes in classifications at the solution $p(\epsilon + \Delta\epsilon)$, where $\Delta\epsilon$ are perturbations can be positive or negative, so long as $\epsilon + \Delta\epsilon > 0$. At an optimal solution, the classification fate of each data point $\mathbf{x}_i$ is encoded in the dual variable μ_i^*, which is a function of ϵ. $\mu_i(\epsilon)$ is the ϵ-parameterized solution path of μ_i such that at any particular solution $p(\epsilon)$, the optimal value of the dual variable $\mu_i^* = \mu_i(\epsilon)$. As a slight abuse of notation, we reserve notation $\mu_i(\epsilon)$ for the functional form of the solution path and write μ_i^ϵ; to refer to the value of the dual variable at a given ϵ.

Lemma 3.2. *The dual variable paths $\mu_i(\epsilon)$ for all $i \in [n]$ are piecewise linear in ϵ.*

Though this lemma seems merely of technical interest, it is a workhorse result for both our methodological contributions—our analytical results and our computationally efficient algorithm, which converts fairness constraints to welfare paths—as well as our substantive fairness results about how fairness perturbations impact individual and learner welfares. The algorithm we present in Section 3.3, performs full welfare analysis for all values of ϵ in a computationally efficient manner by taking advantage of the piecewise linear form of individual and group welfares. Piecewise linearity also sets the stage for the later substantive results about the tension between fairness improvements and the Pareto Principle. We thus walk through the longer proof of this key result in the main text of the paper as it provides important exposition, definitions, and derivations for subsequent results.

Proof. Let D^ϵ be the value of the objective function in (ϵ-fair SVM-D). By the dual formulation of the Soft-Margin SVM, we can use the value of $\frac{\partial D^\epsilon}{\partial \mu_j}$ to partition the set of indices $j \in [n]$ in a way that corresponds to the classification fates of individual vectors $\mathbf{x}_j$ at the optimal solution:

$$\frac{\partial D^\epsilon}{\partial \mu_j} > 0 \longrightarrow \mu_j^\epsilon = 0, \text{ and } j \in \mathcal{F}^\epsilon \tag{7}$$

$$\frac{\partial D^\epsilon}{\partial \mu_j} = 0 \longrightarrow \mu_j^\epsilon \in [0, C], \text{ and } j \in \mathcal{M}^\epsilon \tag{8}$$

$$\frac{\partial D^\epsilon}{\partial \mu_j} < 0 \longrightarrow \mu_j^\epsilon = C, \text{ and } j \in \mathcal{E}^\epsilon \tag{9}$$

Hence, $\mathbf{x}_j$ are either correctly classified free vectors (7), vectors in the margin (8), or error vectors (9). We track membership in these sets by letting $\{\mathcal{F}, \mathcal{M}, \mathcal{E}\}^\epsilon$ be the index set partition at the ϵ-fair solution. To analyze the impact that applying a fairness constraint has on individuals' or groups' welfares, we track the behavior of $\frac{\partial D^\epsilon}{\partial \mu_j}$ and observe how vector index membership in sets $\mathcal{F}^\epsilon$, $\mathcal{M}^\epsilon$, and $\mathcal{E}^\epsilon$ change under a perturbation to ϵ. This information will in turn reveal how classifications change or remain stable upon tightening or loosening the fairness constraint.

Fairness perturbations do not always shuffle data points across the different membership sets $\mathcal{F}^\epsilon$, $\mathcal{M}^\epsilon$, and $\mathcal{E}^\epsilon$. It is clear that for $j \in \{\mathcal{F}, \mathcal{E}\}^\epsilon$, so long as a perturbation of $\Delta\epsilon$ does not cause $\frac{\partial D^\epsilon}{\partial \mu_j}$ to flip signs or to vanish to 0, j will belong to the same set and $h^\epsilon(\mathbf{x}_j) = h^{\epsilon+\Delta\epsilon}(\mathbf{x}_j)$ where $h^\epsilon(\mathbf{x}_j)$ gives the ϵ-fair classification outcome for $\mathbf{x}_j$. In these cases, an individual's welfare is unaffected by the change in the fairness tolerance level from ϵ to $\epsilon + \Delta\epsilon$.

In contrast, vectors $\mathbf{x}_j$ with $j \in \mathcal{M}^\epsilon$ are subject to a different condition to ensure that they stay in the margin: $\frac{\partial D^\epsilon}{\partial \mu_j} = \frac{\partial D^{\epsilon+\Delta\epsilon}}{\partial \mu_j} = 0$, i.e., perturbing by $\Delta\epsilon$ does not lead to any changes in $\frac{\partial D^\epsilon}{\partial \mu_j}$:

$$\frac{\partial D^\epsilon}{\partial \mu_j} = \sum_{i=1}^{n} \mu_i y_i (I - P_{\mathbf{u}})\mathbf{x}_i y_j (I - P_{\mathbf{u}})\mathbf{x}_j + \frac{n\epsilon y_j \langle \mathbf{x}_j, \mathbf{u}\rangle}{\|\mathbf{u}\|^2} + b y_j - 1 = 0 \tag{10}$$

for all $j \in \mathcal{M}^\epsilon$. Let $r_j^{\epsilon,\Delta\epsilon}$ be the change in μ_j^ϵ upon perturbing ϵ by $\Delta\epsilon$, then we have

$$\mu_j^{\epsilon+\Delta\epsilon} = \mu_j^\epsilon + r_j^{\epsilon,\Delta\epsilon} \tag{11}$$

recalling that μ_j^ϵ is the value of μ_j at the optimal solution $p(\epsilon)$. Let $\boldsymbol{r}^{\epsilon,\Delta\epsilon} \in \mathbb{R}^{n+1}$ be the vector of μ_i^ϵ sensitivities to perturbations $\Delta\epsilon$ with $r_0^{\epsilon,\Delta\epsilon}$ as the change in the offset b. For all unshuffled $j \in \mathcal{M}^\epsilon$, we can compute $r_j^{\epsilon,\Delta\epsilon}$ by taking the finite difference of (10) with respect to a $\Delta\epsilon$ perturbation,

$$\sum_{i=1}^{n} r_i^{\epsilon,\Delta\epsilon} y_i y_j \langle (I-P_{\mathbf{u}})\mathbf{x}_i, (I-P_{\mathbf{u}})\mathbf{x}_j\rangle + r_0^{\epsilon,\Delta\epsilon} y_j = \frac{-n y_j \Delta\epsilon}{\|\mathbf{u}\|^2}\langle \mathbf{u}, \mathbf{x}_j\rangle$$

It is clear that $r_i^{\epsilon,\Delta\epsilon} = 0$ for all i that are left unshuffled in the partition $\{\mathcal{F},\mathcal{E}\}^\epsilon$. For these "stable ranges" where no i changes its index set membership, we can simplify the previous expression by summing over only those $r_i^{\epsilon,\Delta\epsilon}$ where $i \in \mathcal{M}^\epsilon$:

$$\sum_{i\in\mathcal{M}^\epsilon} r_i^{\epsilon,\Delta\epsilon} y_i y_j \langle (I-P_{\mathbf{u}})\mathbf{x}_i, (I-P_{\mathbf{u}})\mathbf{x}_j\rangle + r_0^{\epsilon,\Delta\epsilon} y_j = \frac{-n y_j \Delta\epsilon}{\|\mathbf{u}\|^2}\langle \mathbf{u}, \mathbf{x}_j\rangle$$

Thus we can compute $r_i^{\epsilon,\Delta\epsilon}$ by inverting the matrix

$$K^\epsilon = \left(\begin{array}{c|cccc} 0 & y_1 & y_2 & \cdots & y_{|\mathcal{M}^\epsilon|} \\ \hline y_1 & & & & \\ \vdots & & \multicolumn{3}{c}{y_i y_j \langle (I-P_{\mathbf{u}})\mathbf{x}_i, (I-P_{\mathbf{u}})\mathbf{x}_j\rangle} \\ y_2 & & & & \\ y_{|\mathcal{M}^\epsilon|} & & & & \end{array}\right) \in \mathbb{R}^{(|\mathcal{M}^\epsilon|+1)\times(|\mathcal{M}^\epsilon|+1)} \tag{12}$$

where indices are renumbered to only reflect $i, j \in \mathcal{M}^\epsilon$. This matrix is invertible so long as the margin is not empty and the Kernel $K(\mathbf{x}_i,\mathbf{x}_j) = \langle (I-P_{\mathbf{u}})\mathbf{x}_i, (I-P_{\mathbf{u}})\mathbf{x}_j\rangle$ forms a positive definite matrix. Since the objective function in (ϵ-fair SVM-D) is quadratic, a sufficient condition for K^ϵ to be invertible is that the objective is strictly convex—we assume this as a technical condition.[2] The sensitivities of μ_j^ϵ for $j \in \mathcal{M}^\epsilon$ to $\Delta\epsilon$ perturbations are given by

$$\boldsymbol{r}^{\epsilon,\Delta\epsilon} = \underbrace{(K^\epsilon)^{-1}\Big(\frac{-n}{\|\mathbf{u}\|^2}\mathbf{v}\Big)}_{\boldsymbol{r}^\epsilon}\Delta\epsilon, \quad \text{where } \mathbf{v} = \begin{bmatrix} 0 \\ \vdots \\ y_j\langle \mathbf{u}, \mathbf{x}_j\rangle \\ \vdots \end{bmatrix} \in \mathbb{R}^{|\mathcal{M}^\epsilon|+1} \tag{13}$$

Plugging this back into (11), we have

$$\mu_j^{\epsilon+\Delta\epsilon} = \mu_j^\epsilon + \underbrace{\bigg(\Big(K^\epsilon\Big)^{-1}\Big(\frac{-n}{\|\mathbf{u}\|^2}\mathbf{v}\Big)\bigg)_j}_{r_j^\epsilon}\Delta\epsilon \tag{14}$$

Hence, for all $j \in \mathcal{M}^\epsilon$ that stay in the margin, the solution path function $\mu_j(\epsilon)$ is linear in ϵ. For $j \in \{\mathcal{F},\mathcal{E}\}^\epsilon$ that stay in their partition sets, $\mu_j(\epsilon+\Delta\epsilon) = \mu_j(\epsilon)$, so the function is constant.

When $\Delta\epsilon$ perturbations do result in changes in the partition, there are four ways that indices could be shuffled across sets:

(1) $j \in \mathcal{E}^\epsilon$ moves into $\mathcal{M}^{\epsilon+\Delta\epsilon}$

(2) $j \in \mathcal{F}^\epsilon$ moves into $\mathcal{M}^{\epsilon+\Delta\epsilon}$

(3) $j \in \mathcal{M}^\epsilon$ moves into $\mathcal{F}^{\epsilon+\Delta\epsilon}$

(4) $j \in \mathcal{M}^\epsilon$ moves into $\mathcal{E}^{\epsilon+\Delta\epsilon}$

Since index transitions only occur by way of changes to the margin, we need now only confirm that each of these transitions maintains continuous $\mu_j(\epsilon)$ paths for all $j \in [n]$ in order to conclude the proof that the paths are piecewise-linear. □

The linearity of paths $\mu_j(\epsilon)$ for $j \in \mathcal{M}^\epsilon$ gives conditions on the ranges of ϵ wherein individuals' classification outcomes do not change. As such, for any given tolerance parameter ϵ, we can compute the $\Delta\epsilon$ perturbations that yield no changes to individuals' welfares. The following Proposition gives the analytical form of these stable regions, where although fairness appears to be "improving" or "worsening," the adjusted learning process has no material effects on the classificatory outcomes that individuals receive.

Proposition 3.3. *Denote the optimal μ_j^* values at an ϵ-fair SVM solution as μ_j^ϵ for $j \in [n]$. Let*

$$\begin{aligned} r_j &= \bigg((K^\epsilon)^{-1}\Big(\frac{-n}{\|\mathbf{u}\|^2}\mathbf{v}\Big)\bigg)_j \quad \textit{with } K^\epsilon \textit{ and } \mathbf{v} \textit{ as defined in (12) and (13),} \\ d_j &= \sum_{i\in\mathcal{M}^\epsilon} r_i y_i y_j \langle (I-P_{\mathbf{u}})\mathbf{x}_i, (I-P_{\mathbf{u}})\mathbf{x}_j\rangle + r_0 y_j \\ g_j &= 1 - \bigg(\sum_{i=1}^{n} \mu_i^\epsilon y_i (I-P_{\mathbf{u}})\mathbf{x}_i y_j (I-P_{\mathbf{u}})\mathbf{x}_j + \frac{n\epsilon y_j\langle \mathbf{x}_j, \mathbf{u}\rangle}{\|\mathbf{u}\|^2} + b y_j\bigg) \end{aligned} \tag{15}$$

All perturbations of ϵ in the range $\Delta\epsilon \in \Big(\max_j m_j, \min_j M_j\Big)$ where

$$m_j = \begin{cases} \begin{cases} \frac{g_j}{d_j}, & j \in \mathcal{F}^\epsilon, d_j > 0 \\ -\infty, & j \in \mathcal{F}^\epsilon, d_j < 0 \end{cases} \\ \min\{\frac{C-\mu_j^\epsilon}{r_j}, \frac{-\mu_j^\epsilon}{r_j}\}, \ j \in \mathcal{M}^\epsilon \\ \begin{cases} -\infty, & j \in \mathcal{E}^\epsilon, d_j > 0 \\ \frac{g_j}{d_j}, & j \in \mathcal{E}^\epsilon, d_j < 0 \end{cases} \end{cases} \quad M_j = \begin{cases} \begin{cases} \infty, & j \in \mathcal{F}^\epsilon, d_j > 0 \\ \frac{g_j}{d_j}, & j \in \mathcal{F}^\epsilon, d_j < 0 \end{cases} \\ \min\{\frac{C-\mu_j^\epsilon}{r_j}, \frac{-\mu_j^\epsilon}{r_j}\}, \ j \in \mathcal{M}^\epsilon \\ \begin{cases} \frac{g_j}{d_j}, & j \in \mathcal{E}^\epsilon, d_j > 0 \\ \infty, & j \in \mathcal{E}^\epsilon, d_j < 0 \end{cases} \end{cases} \tag{16}$$

yield no changes to index memberships in the partition $\{\mathcal{F},\mathcal{M},\mathcal{E}\}^\epsilon$.

We defer the interested reader to the Appendix for the full proof of this Proposition, though we provide a sketch here. The result follows from observing that the sensitivities $r_i^\epsilon \neq 0$ for $i \in \mathcal{M}^\epsilon$ defined in (13) affect the values $\frac{\partial D^\epsilon}{\partial \mu_j}$ for all $j \in [n]$, and additional conditions must hold to ensure that the vectors that are not on the

[2] We mention the case in which the margin is empty in Section 3.3, though we refer the interested reader to the Appendix for a full exposition of how μ_j^ϵ are updated when the margin is empty and as a result, we cannot compute how i move across index sets via the sensitivities $\boldsymbol{r}$.

margin are also unshuffled by the fairness perturbation. Define

$$g_j^\epsilon = 1 - \Big(\sum_{i=1}^{n} \mu_i^\epsilon y_i (I - P_{\mathbf{u}})\mathbf{x}_i y_j (I - P_{\mathbf{u}})\mathbf{x}_j + \frac{n\epsilon y_j \langle \mathbf{x}_j, \mathbf{u} \rangle}{\|\mathbf{u}\|^2} + b y_j \Big) \tag{17}$$

$$d_j^\epsilon = \frac{\partial D^\epsilon}{\partial \mu_j \partial \epsilon} = \sum_{i \in \mathcal{M}^\epsilon} r_i^\epsilon y_i y_j \langle (I - P_{\mathbf{u}})\mathbf{x}_i, (I - P_{\mathbf{u}})\mathbf{x}_j \rangle + r_0^\epsilon y_j \tag{18}$$

The $\Delta\epsilon$ condition for stability of vectors $\mathbf{x}_j$ for $j \notin \mathcal{M}^\epsilon$ is given by

$$\frac{g_j^\epsilon}{d_j^\epsilon} \tag{19}$$

Recall the conditions of membership in sets $\mathcal{F}$ and $\mathcal{E}$ as given in (7) and (9) respectively. The following observations are critical to computing the bounds of the stable region:

For $j \in \mathcal{F}^\epsilon$, perturbations $\Delta\epsilon$ that increase g_j^ϵ do not threaten j's exiting the set; if $\Delta\epsilon$ decreases g_j^ϵ, then j can enter $\mathcal{M}^{\epsilon+\Delta\epsilon}$.

Inversely, for $j \in \mathcal{E}^\epsilon$, perturbations $\Delta\epsilon$ that decrease g_j^ϵ ensure that j stays in the same partition, i.e., $j \in \mathcal{E}^{\epsilon+\Delta\epsilon}$. Perturbations that increase g_j^ϵ can cause j to shuffle into $\mathcal{M}^{\epsilon+\Delta\epsilon}$.

For $j \in \mathcal{M}^\epsilon$ to stay in the margin, we need $\mu_j^{\epsilon+\Delta\epsilon} \in [0, C]$. Once μ_j^ϵ hits either endpoint of the interval, j risks shuffling across to $\mathcal{F}^{\epsilon+\Delta\epsilon}$ or $\mathcal{E}^{\epsilon+\Delta\epsilon}$.

Computing these transition inequalities results in a set of conditions that ensure that a partition is stable. Since $\Delta\epsilon$ can be either positive or negative, we take the maximum of the lower bounds (m_j) and the minimum of the upper bounds (M_j) to arrive at the range of stable perturbations given in (16). We call the bounds of this interval the "breakpoints" of the solution paths.

This Proposition reveals a mismatch between the ostensible changes to the fairness level of an ϵ-fair Soft-Margin SVM learning process and the actual felt changes in outcomes by the individuals who are subject to the system. This results from the simple fact that the optimization problem captures changes in the learner's optimal solution but does not offer such fine-grained information on how individuals' outcomes vary as a result of $\Delta\epsilon$ perturbations. So long as the fairness constraint is binding and its associated dual variable $\gamma > 0$, then tightening or loosening a fairness constraint *does* alter the loss of the optimal learner classifier—the actual SVM solution changes—yet analyzed from the perspective of the individual agents $\mathbf{x}_i$, so long as the $\Delta\epsilon$ perturbation occurs within the range given by (16), classifications issued under this $\epsilon + \Delta\epsilon$-fair SVM solution are identical to those under the ϵ-fair solution. Thus despite the apparent more "fair" signal that a classifier abiding by $\epsilon + \Delta\epsilon < \epsilon$ sends, agents are made no better off in terms of welfare. This result is summarized in the following Corollary.

Corollary 3.4. *Let $\{p(\epsilon), W_0(\epsilon), W_1(\epsilon)\}$ be a triple expressing the welfares of the learner, group $z = 0$, and group $z = 1$ under the ϵ-fair SVM solution. Then for any $\Delta\epsilon \in (\max_j m_j, 0)$ where m_j is defined in (16), $\{p(\epsilon), W_0(\epsilon), W_1(\epsilon)\} \succsim \{p(\epsilon + \Delta\epsilon), W_0(\epsilon + \Delta\epsilon), W_1(\epsilon + \Delta\epsilon)\}$.*

Once we have demarcated the limits of $\Delta\epsilon$ perturbations that yield no changes to the partition, i.e., $\{\mathcal{F}, \mathcal{M}, \mathcal{E}\}^\epsilon = \{\mathcal{F}, \mathcal{M}, \mathcal{E}\}^{\epsilon+\Delta\epsilon}$, we can move on to consider the welfare effects of $\Delta\epsilon$ perturbations that exceed the stable region outlined in Proposition 3.3. At each such breakpoint when $\Delta\epsilon$ reaches $\max_j m_j$ or $\min_j M_j$ as defined in (16), the margin set changes: $\mathcal{M}^\epsilon \neq \mathcal{M}^{\epsilon+\Delta\epsilon}$. As such, $r_j^{\epsilon+\Delta\epsilon}$ for $j \in \mathcal{M}^{\epsilon+\Delta\epsilon}$ must be recomputed via (13). These sensitivities hold until the next breakpoint when the set $\mathcal{M}$ updates again.

We can associate a group welfare with the classification scheme at each of the breakpoints. As already illustrated, index partitions are static in the stable regions around each breakpoint, so group welfares will also be unchanged in these regions. As such, we need only compute welfares at breakpoints to characterize the paths for $\epsilon \in [0, 1]$. This method of analysis allows practitioners to straightforwardly determine whether the next ϵ breakpoint actually translates into better or worse outcomes for the group as a whole.

Of the four possible events that occur at a breakpoint, index transitions between the partitions $\mathcal{M}$ and $\mathcal{E}$ correspond to changed classifications that affect group utilities. The following Proposition characterizes those breakpoint transitions that effect welfares triples $\{p(\epsilon), W_0(\epsilon), W_1(\epsilon)\}$ for the learner, group $z = 0$, and group $z = 1$, that are *strictly Pareto dominated* by the welfare triple at a neighboring ϵ breakpoint. The full proof is left to the Appendix.

Proposition 3.5. *Consider the welfare triple at the optimal ϵ-fair SVM solution given by $\{p(\epsilon), W_0(\epsilon), W_1(\epsilon)\}$. Let $b_L = \max_j m_j < 0$ be the neighboring lower breakpoint where index $\ell = \arg\max_j m_j$; let $b_U = \min_j M_j > 0$ be the neighboring upper breakpoint where index $u = \arg\min_j M_j$, assuming uniqueness in the* arg max *and* arg min. *If $\ell \in \mathcal{E}^\epsilon$ and $y_\ell = -1$, or if $\ell \in \mathcal{M}^\epsilon$ and $y_\ell = +1$, then*

$$\{p(\epsilon + b_L), W_0(\epsilon + b_L), W_1(\epsilon + b_L)\}\} \prec \{p(\epsilon), W_0(\epsilon), W_1(\epsilon)\}$$

If $u \in \mathcal{E}^\epsilon$ and $y_u = +1$, or if $u \in \mathcal{M}^\epsilon$ and $y_u = -1$, then

$$\{p(\epsilon), W_0(\epsilon), W_1(\epsilon)\} \prec \{p(\epsilon + b_U), W_0(\epsilon + b_U), W_1(\epsilon + b_U)\}$$

Thus minimizing loss in the presence of stricter fairness constraints need not correspond to monotonic gains or losses in the welfare levels of social groups. Fairness perturbations do not have a straightforward effect on classifications. Further, these results do not only arise as an unfortunate outcome of using the particular proxy fairness constraint suggested by Zafar et al [5]. So long as the ϵ parameter appears in the linear part of the dual Soft-Margin SVM objective function, the $\mu_j(\epsilon)$ paths exhibit a piecewise linear form characterized by stable regions and breakpoints. Hence, these results apply to many proxy fairness criteria that have so far been proposed in the literature [5, 17, 28]. Even when the dual variable paths are not piecewise linear, so long as they are non-monotonic, fairer classification outcomes do not necessarily confer welfare benefits to the disadvantaged group. Monotonicity in welfare space is mathematically distinct from monotonicity in fairness space.

The preceding analyses show that although fairness constraints are often intended to improve classification outcomes for some disadvantaged group, they in general do not abide by the Pareto Principle, a common welfare economic axiom for deciding among social alternatives. That is, asking that an algorithmic procedure abide by a more stringent fairness criteria can lead to enacting classification schemes that actually make every stakeholder group worse-off. Here, the supposed "improved fairness" achieved by decreasing the unfairness tolerance parameter ϵ fails to translate into any meaningful improvements in the number of desirable outcomes issued to members of either group.

Theorem 3.6. *Consider two fairness-constrained ERM programs parameterized by ϵ_1 and ϵ_2 where $\epsilon_1 < \epsilon_2$. Then a decision-maker who always prefers the classification outcomes issued under the "more fair" ϵ_1-fair solution to those under the "less fair" ϵ_2-fair solution does not abide by the Pareto Principle.*

3.3 Algorithm and Complexity

We build upon the previous section of translating fairness constraints into individual welfare outcomes by considering the operationalization of our analysis and its practicality. The algorithmic procedure presented in this section computes ϵ breakpoints and tracks the solution paths of the $\mu_j(\epsilon)$ for all individuals. Hence, the procedure enables the comparison of different social groups' welfares—where welfare is determined by the machine's allocative outcome—by aggregating the classification outcomes of all individuals j in a group z. Algorithm 1 outputs two useful fairness-relevant constructs that have as yet not been explored in the literature: 1) solution paths $\mu_j(\epsilon)$ for $j \in [n]$ tracking individuals' welfares, and 2) full ϵ parameterized curves tracking groups' welfares.

The analysis of the previous section forms the backbone of the main update rules that construct the $\mu_j(\epsilon)$ paths in Algorithm 1. In particular the values r_j^ϵ, g_j^ϵ, and d_j^ϵ as defined in (13), (17), and (18) respectively are key to computing the ϵ breakpoints, which in turn fully determine the piecewise linear form of $\mu_j(\epsilon)$. There is, however, one corner case that the procedure must check that was not discussed in the preceding section. We had previously required that the matrix K^ϵ be invertible, which is the case whenever our objective function is strictly convex. But if the margin is empty, the standard update procedure, which computes sensitivities r_j^ϵ and K^ϵ, will not suffice. The KKT optimality condition $\sum_{i=1}^n \mu_i y_i = 0$ requires that the multiple indices moving in the margin at once must be positive and negative examples. For this reason we must refer to a different procedure to compute the ϵ breakpoint at which this transition occurs. For continuity of the main text of this paper, the full exposition of this analysis is given in the Appendix.

The following complexity result highlights the practicality of implementing the fairness-to-welfare mapping in Algorithm 1 to track the full solution paths of an ϵ-fair SVM program. We note that standard SVM algorithms such as `LibSVM` run in $O(n^3)$, and thus once the algorithm has been initialized with the unconstrained SVM solution, the complexity of computing both the full individual solution paths $\mu_j(\epsilon)$ and the full group welfare curves $\{W_0(\epsilon), W_1(\epsilon)\}$ is on the same order as that of computing a single SVM solution.

Theorem 3.7. *Each iteration of Algorithm 1 runs in $O(n^2 + |\mathcal{M}|^2)$. For breakpoints on the order of n, the full run time complexity is $O(n^3 + n|\mathcal{M}|^2)$.*

Proof. Each iteration of the fairness-to-welfare algorithm requires the inversion of matrix $K^\epsilon \in \mathbb{R}^{|\mathcal{M}^\epsilon|+1}$ and the computations of $r_j^\epsilon \in \mathbb{R}^{|\mathcal{M}^\epsilon|}$ for $j \in \mathcal{M}^\epsilon$, and g_j^ϵ and d_j^ϵ for $j \in \{\mathcal{F}, \mathcal{E}\}^\epsilon$.

The standard Gauss-Jordan matrix inversion technique runs in $O(|\mathcal{M}|^3)$, but we take advantage of partition update rules to lower the number of computations: Since at each new breakpoint, the partition tends to change because of additions or eliminations of a single index j from the set $\mathcal{M}$, we can use the Cholesky decomposition rank-one update or downdate to ease the need to recompute the full matrix inverse at every iteration, thereby reducing the complexity of the operation to $O(|\mathcal{M}|^2)$. Computing the stability region conditions for $j \in \{\mathcal{F}, \mathcal{E}\}$ requires $O\Big((n - |\mathcal{M}|)|\mathcal{M}|\Big)$ steps. As such, at each breakpoint, the total computational cost is $O(|\mathcal{M}|^2 + n^2)$.

The number of breakpoints for each full run of the algorithm depends on the data distribution and how sensitive the solution is to the constraint. As a heuristic, datasets whose fairness constraints bind for smaller ϵ have fewer breakpoints. Previous empirical results on the full SVM path for L1 and L2 regularization have found that the number of breakpoints tends to be on the order of n [24–27]. Thus after initialization with 0-fair SVM solution, the final complexity for the algorithm is $O(n^3 + n|\mathcal{M}|^2)$. □

ALGORITHM 1: Fairness-to-welfare solution paths as a function of ϵ

```
Input: set $\mathcal{X}$ of $n$ data points $\{\mathbf{x}_i, z_i, y_i\}$
Output: solutions paths $\boldsymbol{\mu}(\epsilon)$ and group welfare curves $\{W_0(\epsilon), W_1(\epsilon)\}$
$\boldsymbol{\mu}^0 = \arg\min_{\boldsymbol{\mu}} D(\boldsymbol{\mu})$ of (0-fair SVM-D);
$\epsilon = 0, \Delta\epsilon = 0$;
$|n_0| = \sum_{i=1}^n \mathbb{1}[z_i = 0], |n_1| = \sum_{i=1}^n \mathbb{1}[z_i = 1]$;
while $\epsilon < 1$ do
    $W_0 = 0, W_1 = 0$;
    for each $\mu_i^\epsilon$ do
        update $\{\mathcal{F}, \mathcal{M}, \mathcal{E}\}^\epsilon$ according to (7), (8), (9);
        if $(\mu_i < C \,\&\, y_i = 1) \,||\, (\mu_i = C \,\&\, y_i = 0)$ then
            $W_{z_i} = W_{z_i} + 1$;
        end
    end
    $W_0(\epsilon) = \frac{W_0}{n_0}; W_1(\epsilon) = \frac{W_1}{n_1}$;
    if $|\mathcal{M}^\epsilon| = 0$ then
        $\Delta\epsilon = \min_i M_i$ as given in (26);
        update $\{\mathcal{F}, \mathcal{M}, \mathcal{E}\}^\epsilon$ according to (28) and (29);
        $\epsilon = \epsilon + \Delta\epsilon$;
    end
    compute $\boldsymbol{r}^\epsilon, \boldsymbol{d}^\epsilon$ according to (13), (18);
    $\Delta\epsilon = \min_i M_i$ as given in (16);
    $\mu_i^{\epsilon+\Delta\epsilon} = \mu_i^\epsilon + r_i^\epsilon \Delta\epsilon$ for $i \in \mathcal{M}^\epsilon$, $\mu_i^\epsilon = \mu_i^{\epsilon+\Delta\epsilon}$ for $i \in \{\mathcal{F}, \mathcal{E}\}^\epsilon$;
    $\epsilon = \epsilon + \Delta\epsilon$;
end
return $(\boldsymbol{\mu}(\epsilon), W_0(\epsilon), W_1(\epsilon))$
```

3.4 Impact of Fair Learning on Learner's Welfare

Having proven the main welfare-relevant sensitivity result for groups, we return to more standard analysis of the effect of $\Delta\epsilon$ perturbations on the learner's loss. In this case, we directly solve for the dual variable of the fairness constraint. Recall γ^* from (23):

$$\gamma^* = \gamma_1^* - \gamma_2^* = \frac{n(n(\beta_- - \beta_+) + \sum_{i=1}^n \mu_i y_i \langle \mathbf{x}_i, \mathbf{u} \rangle)}{||\mathbf{u}||^2} \tag{20}$$

By complementary slackness, one of β_- and β_+ is zero, and the other is ϵ. In particular, if $\beta_- = 0$, then $\beta_+ = \epsilon$, then we know that $\gamma > 0$. Thus the original fairness constraint that binds is the upper bound on covariance, suggesting that the optimal classifier must be constrained to limit its positive covariance with group $z = 1$. If $\beta_+ = 0$, then $\beta_- = \epsilon$ and $\gamma < 0$, and the classifier must be constrained to limit its positive covariance with group $z = 0$.

We can interpret the value of the dual variable Lagrange multiplier as the shadow price of the fairness constraint. It gives the additional loss in the objective value that the learner would achieve if the fairness constraint were infinitesimally loosened. Whenever a fairness constraint binds, its shadow price is readily computable and is given by $|\gamma^*|$. It bears noting that because (ϵ-fair Soft-SVM) is not a linear program, $|\gamma^*|$ can only be interpreted as a measure of *local* sensitivity, valid only in a small neighborhood around an optimal solution. But through an alternative lens of sensitivity analysis, we can derive a lower bound on global sensitivity due to changes in the fairness tolerance parameter ϵ. By writing ϵ as a perturbation variable, we can perform sensitivity analysis on the same ϵ-constrained problem. Returning to the perturbation function $p(\epsilon)$, we have

$$p(\epsilon) \geq \sup_{\mu, \gamma} \{\mathcal{L}(\boldsymbol{\mu}^*, \gamma^*) - \epsilon|\gamma^*|\} \tag{21}$$

where $\mathcal{L}(\boldsymbol{\mu}^*, \gamma^*)$ gives the solution to the 0-fair SVM problem.

$$\mathcal{L}(\boldsymbol{\mu}^*, \gamma^*) = \max_{\boldsymbol{\mu} \in [0,C]^n, \gamma} -\frac{1}{2}\left\|\sum_{i=1}^{n} \mu_i y_i (I - P_u)\mathbf{x}_i\right\|^2 + \sum_{i=1} \mu_i \tag{22}$$

The perturbation formulation given in (21) is identical in form to the original program (ϵ-fair-SVM1-P) but gives a global bound on $p(\epsilon)$ for all $\epsilon \in [0, 1]$. Since (21) gives a lower bound, the global sensitivity bound yields an asymmetric interpretation.

Proposition 3.8. *If* $\Delta\epsilon < 0$ *and* $|\gamma^*| \gg 0$, *then* $p(\epsilon+\Delta\epsilon) - p(\epsilon) \gg 0$. *If* $\Delta\epsilon > 0$ *and* $|\gamma^*| < \delta$ *for small* δ, *then* $p(\epsilon+\Delta\epsilon) - p(\epsilon) \in [-\delta\Delta\epsilon, 0]$, *and is thus also small in magnitude.*

Proposition 3.8 shows that tightening the fairness constraint when its shadow price is high leads to a great increase in learner loss, but loosening the fairness constraint when its shadow price is small leads only to a small decrease in loss.

4 EXPERIMENTS

To demonstrate the efficacy of our approach, we track the impact of ϵ-fairness constrained SVM programs on the classification outcomes of individuals in the Adult dataset. The target variable in the dataset is a binary value indicating whether the individual has an annual income of more or less than \$50,000. If such a dataset were used to train a tool to be deployed in consequential resource allocation—say, for the purpose of determining access to credit—then classification decisions directly impact individuals' welfares.

Individual solution paths and relative group welfare changes are given in Figure 1. As ϵ increases from left to right, the fairness constraint is loosened, and outcomes become "less fair." In the case of the ϵ-fair SVM solution to the Adult dataset, the fairness constraint ceases to bind at the optimal solution when $\epsilon \approx 0.175$. The top panel shows example individual piecewise linear paths of dual variables $\mu_i(\epsilon)$, providing a visual depiction of how individual points can transition across index sets: from $\mu_i = 0, i \in \mathcal{F}$ and being correctly labeled, to $\mu_i \in (0, 1), i \in \mathcal{M}$, being correctly labeled but in margin; to $\mu_i = 1, i \in \mathcal{E}$ and being incorrectly labeled. Solid paths indicate individuals coded female; dashed paths indicate those coded males. As the top panel of Figure 1 shows, the actual "journey" of these paths are varied as ϵ changes.

As expected, tightening the fairness constraint in the ϵ-fair program does tend to lead to improved welfare outcomes for females as a group (more female individuals receive a positive classification), while males experience a relative decline in group welfare (receiving fewer positive classifications). However, as suggested by our results in Section 3.2, these welfare changes are not monotonic for either group. Tightening the fairness constraint could lead to declines in both groups' welfares, demonstrating that preferring more fair solutions in this predictive model does not abide by the Pareto Principle. We highlight an instance of this result in the bottom panel of Figure 1, where orange dashed lines to the left of black ones mark off solutions where "more fair" outcomes (orange) are Pareto-dominated by "less fair" (black) ones. A practitioner working in a domain in which welfare considerations might override parity-based fairness ones may prefer the outcomes of a fair learning procedure with $\epsilon \approx 0.045$ to one with $\epsilon \approx 0.015$. Additional plots showing absolute changes in group welfare and optimal learner value are given in the Appendix.

5 DISCUSSION

The question that leads off this paper—*How do leading notions of fairness as defined by computer scientists map onto longer-standing notions of social welfare?*—sets an important agenda to come for the field of algorithmic fairness. It asks that the community look to disciplines that have long considered the problem of allocating goods in accordance with ideals of justice and fairness. For example, the notion of welfare in this paper draws from work in welfare and public economics. The outcomes issued by an optimal classifier can, thus, be interpreted using welfare economic tools developed for considerations of social efficiency and equity. In an effort to situate computer scientists' notions of fairness within a broader understanding of distributive justice, we also show that loss minimization problems can indeed be mapped onto welfare maximization ones and vice versa. For reasons of continuity, analyses of this correspondence do not appear in the main text—we defer the interested reader to the Appendix—though we present an abbreviated overview here. We encourage readers to consider the main results of this paper, which construct welfare paths out of fair learning algorithms, as a part of this larger project of bridging the two approaches.

5.1 Bridging Fair Machine Learning and Social Welfare Maximization

To highlight the correspondence between the machine learning and welfare economic approaches to allocation, we show that loss minimizing solutions can be understood as welfare maximizing ones under a particular social welfare function. In the Planner's Problem, a planner maximizes social welfare represented as the weighted sum of utility functions. Inverting the Planner's Problem gives a question concerning social equity: *"Given a particular allocation, what is the presumptive social weight function that would yield it as optimal?"* We show that the set of predictions issued by the optimal classifier of any loss minimization task can be given as the set of allocations in the Planner's Problem over the same individuals endowed with a set of welfare weights. Analyzing the distribution of

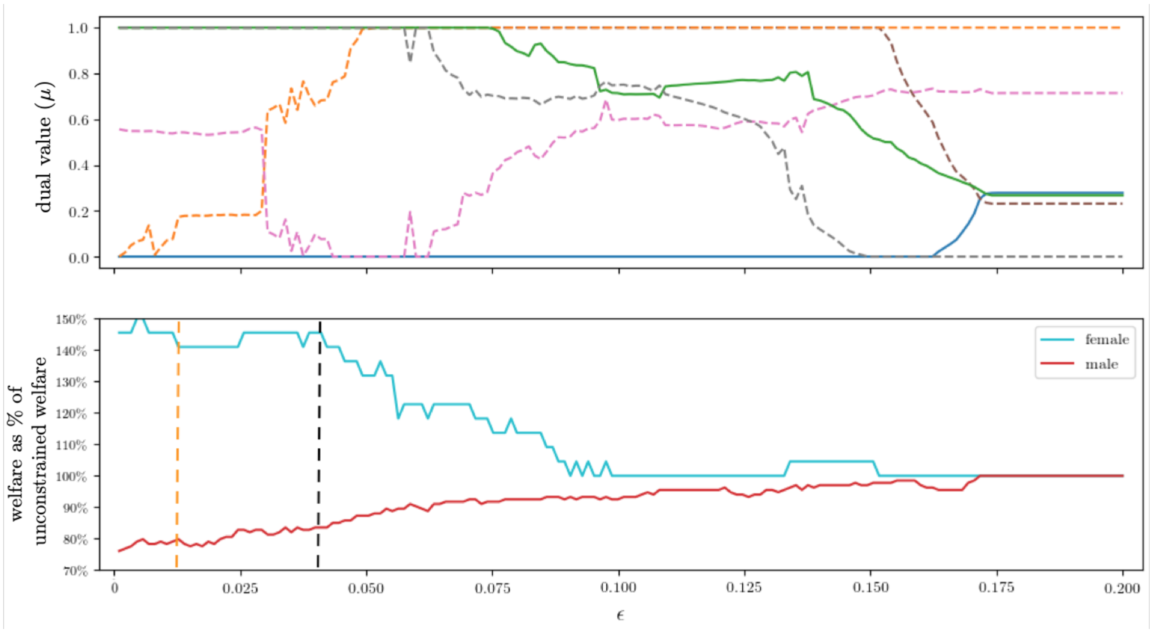

Figure 1: Fairness-to-welfare solution paths for individuals (top panel) and groups (bottom panel) on the Adult dataset.

implied weights of individuals and groups offers a welfare economic way of considering the "fairness" of classifications. We also derive a converse result: *"Given a social welfare maximizing allocation, what model that can achieve an equivalent classification?"* Our solution's approach records the set of welfares, defined by the number of positively labeled individuals, achievable for each social group.

5.2 Interpreting Welfare Alongside Fairness

Welfare economics can lend particular insights into formalizing notions of distributional fairness and general insights into building a "technical" field and methodology that grapples with normative questions. The field is concerned with what public policies ought to be, how to improve individuals' well-beings, and what distribution of outcomes are preferable. Answers to these questions appeal to values and judgments that refer to more than just descriptive or predictive facts about the world. The success of fair machine will largely hang on how well it can adapt to a similar ambitious task.

However, welfare economics is not the only—nor should it even serve as the main—academic resource for thinking through how goods ought to be provisioned in a just society. In this moment of broad appeal to the prowess of algorithmic systems, researchers in computing are called on to advise on matters beyond their specialized expertise and training. Many of these matters require explicit normative, political, and social-scientific reasoning. Insights and methods from across the arts, humanities, social sciences, and natural sciences bear fruit in answering these questions.

This paper does not look to contribute a new fair learning algorithm or a new fairness definition. We take a popular classification algorithm, the Soft Margin SVM, append a parity-based fairness constraint, and analyze its implications on welfare. The constraint that we center in the paper is just one concretization of a large menu of fairness notions that have been offered up to now. The method of analysis developed in the paper applies generally to any convex formulations of these constraints, including versions of balance for false positives, balance for false negatives, and equality of opportunity that have circulated in the literature [17, 18, 28]. It is important future work to investigate the welfare implications of state-of-the-art fair classification algorithms that the community continues to develop, which can deal with a wider range of models and constraints, including non-convex ones.

This paper asks that researchers in fair machine learning reevaluate not only their lodestars of optimality and efficiency but also their latest metrics of fairness. By viewing classification outcomes as allocations of a good, we incorporate considerations of individual and group utility in our analysis of classification regimes. The concept of "utility" in evaluations of social policy remains controversial, but in many cases of social distribution, utility considerations provide a partial but still important perspective on what is at stake within an allocative task. Utility-based notions of welfare can capture the relative benefit that a particular good can have on a particular individual. If machine learning systems are in effect serving as resource distribution mechanisms, then questions about fairness should align with questions of "Who benefits?" Our results show that many parity-based formulations of fairness do not ensure that disadvantaged groups benefit. Preferring a classifier that better accords with a fairness measure can lead to selecting allocations that lower the welfare for every group. Nevertheless, there remain reasons in favor of limiting levels of inequality not reflected in utilitarian calculus. In some cases, the gap between groups is itself objectionable, and minimizing this difference overrides maximizing the absolute utility level of disadvantaged groups. But without acknowledging and accounting for these reasons, well-intentioned optimization tasks that seek to be "fairer" can further disadvantage social groups for no reason but to satisfy a given fairness metric.

REFERENCES

[1] Amartya Sen. *Equality of What?* Cambridge University Press, Cambridge, 1980. Reprinted in John Rawls et al., Liberty, Equality and Law (Cambridge: Cambridge University Press, 1987).

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

[3] Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016.

[4] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226. ACM, 2012.

[5] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1171–1180. International World Wide Web Conferences Steering Committee, 2017.

[6] Yahav Bechavod and Katrina Ligett. Learning fair classifiers: A regularization-inspired approach. *arXiv preprint arXiv:1707.00044*, 2017.

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

[8] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *Proceedings of the 8th Innovations in Theoretical Computer Science Conference*, pages 43:1–43:23. ACM, 2017.

[9] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. Fairness-aware learning through regularization approach. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, pages 643–650. IEEE, 2011.

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

[11] Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*, pages 325–333, 2016.

[12] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems*, pages 5680–5689, 2017.

[13] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pages 3992–4001, 2017.

[14] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076, 2017.

[15] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, pages 656–666, 2017.

[16] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2569–2577, 2018.

[17] Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, pages 2791–2801, 2018.

[18] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69, 2018.

[19] Sendhil Mullainathan. Algorithmic fairness and the social welfare function. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 1–1. ACM, 2018.

[20] Hoda Heidari, Claudio Ferrari, Krishna Gummadi, and Andreas Krause. Fairness behind a veil of ignorance: A welfare analysis for automated decision making. In *Advances in Neural Information Processing Systems*, pages 1265–1276, 2018.

[21] Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

[22] Lydia Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pages 3156–3164, 2018.

[23] Christopher P Diehl and Gert Cauwenberghs. Svm incremental learning, adaptation and optimization. In *Neural Networks, 2003. Proceedings of the International Joint Conference on*, volume 4, pages 2685–2690. IEEE, 2003.

[24] Trevor Hastie, Saharon Rosset, Robert Tibshirani, and Ji Zhu. The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, 5(Oct):1391–1415, 2004.

[25] Li Wang, Ji Zhu, and Hui Zou. The doubly regularized support vector machine. *Statistica Sinica*, 16(2):589, 2006.

[26] Gang Wang, Dit-Yan Yeung, and Frederick H Lochovsky. A kernel path algorithm for support vector machines. In *Proceedings of the 24th international conference on Machine learning*, pages 951–958. ACM, 2007.

[27] Masayuki Karasuyama, Naoyuki Harada, Masashi Sugiyama, and Ichiro Takeuchi. Multi-parametric solution-path algorithm for instance-weighted support vector machines. *Machine learning*, 88(3):297–330, 2012.

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

[29] Marc Fleurbaey and François Maniquet. *A theory of fairness and social welfare*, volume 48. Cambridge University Press, 2011.

[30] Emmanuel Saez and Stefanie Stantcheva. Generalized social marginal welfare weights for optimal tax theory. *American Economic Review*, 106(1):24–45, 2016.

[31] Lucy F Ackert, Jorge Martinez-Vazquez, and Mark Rider. Social preferences and tax policy design: some experimental evidence. *Economic Inquiry*, 45(3):487–501, 2007.

[32] Floris T Zoutman, Bas Jacobs, and Egbert LW Jongen. Optimal redistributive taxes and redistributive preferences in the netherlands. *Erasmus University Rotterdam*, 2013.

[33] Vidar Christiansen and Eilev S Jansen. Implicit social preferences in the norwegian system of indirect taxation. *Journal of Public Economics*, 10(2):217–245, 1978.

[34] Matthew Adler. *Well-being and fair distribution: beyond cost-benefit analysis*. Oxford University Press, 2012.

[35] Marc Fleurbaey, François Maniquet, et al. Optimal taxation theory and principles of fairness. Technical report, Université catholique de Louvain, Center for Operations Research and ?, 2015.

[36] Ilyana Kuziemko, Michael I Norton, Emmanuel Saez, and Stefanie Stantcheva. How elastic are preferences for redistribution? evidence from randomized survey experiments. *American Economic Review*, 105(4):1478–1508, 2015.

[37] Herbert Edelsbrunner and Ernst Peter Mücke. Simulation of simplicity: a technique to cope with degenerate cases in geometric algorithms. *ACM Transactions on Graphics (tog)*, 9(1):66–104, 1990.

6 APPENDIX

6.1 Dual derivations of the ϵ-fair SVM program

In this Appendix section, we walk through the preliminary setup of the ϵ-fair SVM program given in Section 5.1 and present intermediate derivations omitted from the main text.

Recall that the fair empirical risk minimization program of central focus is

$$\begin{aligned} \underset{\boldsymbol{\theta}, b}{\text{minimize}} \quad & \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C\sum_{i=1}^{n} \xi_i \\ \text{subject to} \quad & y_i(\boldsymbol{\theta}^\top \mathbf{x}_i + b) - 1 + \xi_i \geq 0, \qquad (\epsilon\text{-fair Soft-SVM}) \\ & \xi_i \geq 0, \\ & f_{\boldsymbol{\theta},b}(\mathbf{x}, y) \leq \epsilon \end{aligned}$$

The hyperplane parameters are $\boldsymbol{\theta} \in \mathbb{R}^d$ and $b \in \mathbb{R}$. The nonnegative ξ_i allow the margin constraints to have some slack—this is why these variables are commonly called "slack variables." In the Soft-Margin (as opposed to the Hard-Margin) SVM, the margin is permitted to be less than 1. A slack variable $\xi_i > 0$ corresponds to a point $\mathbf{x}_i$ having a functional margin of less than 1. There is a cost associated with this margin violation, even though it need not correspond to a classification error. $C > 0$ is a hyperparameter tunable by the learner to optimize this trade-off between preferring a larger margin and penalizing violations of the margin.

When we combine the general Soft-Margin SVM with the covariance parity constraint in (4) proposed by Zafar et al. [5], we have the program

$$\begin{aligned} \underset{\boldsymbol{\theta}, b}{\text{minimize}} \quad & \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C\sum_{i=1}^{n} \xi_i \\ \text{subject to} \quad & y_i(\boldsymbol{\theta}^\top \mathbf{x}_i + b) - 1 + \xi \geq 0, \qquad (\epsilon\text{-fair-SVM1-P}) \\ & \left|\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})(\boldsymbol{\theta}^\top \mathbf{x}_i + b)\right| \leq \epsilon \end{aligned}$$

where $\bar{z}$ reflects the bias in the demographic makeup of $\mathcal{X}$: $\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$. The corresponding Lagrangian is

$$\begin{aligned} & kearns2018preventing\mathcal{L}_P(\boldsymbol{\theta}, b, \boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \gamma_1, \gamma_2) \\ & = \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C\sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \lambda_i - \sum_{i=1}^{n} \mu_i(y_i(\boldsymbol{\theta}^\top \mathbf{x}_i + b) - 1 + \xi_i) \\ & - \gamma_1\Big(\epsilon - \frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})(\boldsymbol{\theta}^\top \mathbf{x}_i + b)\Big) \\ & \qquad (\epsilon\text{-fair-SVM1-L}) \\ & - \gamma_2\Big(\epsilon - \frac{1}{n}\sum_{i=1}^{n}(\bar{z} - z_i)(\boldsymbol{\theta}^\top \mathbf{x}_i + b)\Big) \end{aligned}$$

where $\boldsymbol{\theta} \in \mathbb{R}^d, b \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^n$ are Primal variables. The (nonnegative) Lagrange multipliers $\boldsymbol{\lambda}, \boldsymbol{\mu} \in \mathbb{R}^n$ correspond to the n nonnegativity constraints $\xi_i \geq 0$ and the margin-slack constraints $y_i(\boldsymbol{\theta}^\top \mathbf{x}_i + b) - 1 + \xi_i \geq 0$ respectively. The multiplier μ_i relays information about the functional margin of its corresponding point $\mathbf{x}_i$. If the margin is greater than 1 in the Primal, *i.e.*, there is slack in the constraint), then by complementary slackness, $\mu_i = 0$. Otherwise, if the constraint holds with equality, $\mu_i \in (0, C]$. When the classifier commits an error on $\mathbf{x}_i$, $y_i(\boldsymbol{\theta}^\top \mathbf{x}_i + b) \leq$, and then by the KKT conditions, $\mu_i = C$.

The multipliers $\gamma_1, \gamma_2 \in \mathbb{R}$ correspond to the two linearized forms of the absolute value fairness constraint. Notice that these two constraints cannot simultaneously hold with equality for $\epsilon > 0$. Thus, by complementary slackness again, we know that at least one of γ_1, γ_2 is zero, and the other is strictly positive.

By the Karush-Kuhn-Tucker conditions, at the solution of the convex program, the gradients of $\mathcal{L}$ with respect to $\boldsymbol{\theta}$, b, and ξ_i are zero:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}} &:= 0 \Rightarrow \boldsymbol{\theta} = \sum_{i=1}^{n} \mu_i y_i \mathbf{x}_i - \frac{\gamma}{n}\Big(\sum_{i=1}^{n}(z_i - \bar{z})\mathbf{x}_i\Big) \\ \frac{\partial \mathcal{L}}{\partial b} &:= 0 \Rightarrow \sum_{i=1}^{n} \mu_i y_i = \frac{\gamma}{n}\sum_{i=1}^{n}(z_i - \bar{z}) = 0 \\ \frac{\partial \mathcal{L}}{\partial \xi_i} &:= 0 \Rightarrow \lambda_i + \mu_i = C, \qquad i = 1, \ldots, n \end{aligned}$$

Plugging in these optimality conditions, the dual Lagrangian is

$$\mathcal{L}_D(\boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \gamma_1, \gamma_2) = -\frac{1}{2}\left\|\sum_{i=1}^{n} \mu_i y_i \mathbf{x}_i - \frac{\gamma}{n}\sum_{i=1}^{n}(z_i - \bar{z})\mathbf{x}_i\right\|^2 + \sum_{i=1}^{n} \mu_i - |\gamma|\epsilon$$

where we have $\gamma = \gamma_1 - \gamma_2$, since at most one side of the fairness constraint binds, thereby ensuring that at least one of γ_1 or γ_2 is 0. The dual maximizes this objective subject to the constraints $\mu_i \in [0, C]$ for all i and $\sum_{i=1} \mu_i y_i = 0$. Hence, we derive the full dual problem

$$\begin{aligned} \underset{\boldsymbol{\mu}, \gamma, V}{\text{maximize}} \quad & -\frac{1}{2}\left\|\sum_{i=1}^{n} \mu_i y_i \mathbf{x}_i - \frac{\gamma}{n}\sum_{i=1}^{n}(z_i - \bar{z})\mathbf{x}_i\right\|^2 + \sum_{i=1}^{n} \mu_i - V\epsilon \\ \text{subject to} \quad & \mu_i \in [0, C], \qquad i = 1, \ldots, n, \\ & \qquad (\epsilon\text{-fair-SVM1-D}) \\ & \sum_{i=1}^{n} \mu_i y_i = 0, \\ & \gamma \in [-V, V] \end{aligned}$$

where we have introduced the variable V to eliminate the absolute value function $|\gamma|$ in the objective. Notice that when $\gamma = 0$ and neither of the constraints bind, we recover the standard dual SVM program. Since we are concerned with fair learning that does in fact alter an optimal solution, we consider cases in which V is strictly positive. From this program, we introduce additional dual variables β_- and β_+, corresponding to the $\gamma \in [-V, V]$ constraint and derive the Lagrangian

$$\begin{aligned} \mathcal{L}(\boldsymbol{\mu}, \gamma, V, \beta_-, \beta_+) = &-\frac{1}{2}\left\|\sum_{i=1}^{n} \mu_i y_i \mathbf{x}_i - \frac{\gamma}{n}\sum_{i=1}^{n}(z_i - \bar{z})\mathbf{x}_i\right\|^2 + \sum_{i=1}^{n} \mu_i \\ &- V\epsilon + \gamma(\beta_- - \beta_+) + V(\beta_- + \beta_+) \end{aligned}$$

Under KKT conditions, $\beta_- + \beta_+ = \epsilon$ and

$$\gamma^* = \frac{n(n(\beta_- - \beta_+) + \sum_{i=1}^{n} \mu_i y_i \langle \mathbf{x}_i, \mathbf{u} \rangle)}{\|\mathbf{u}\|^2} \tag{23}$$

where $\mathbf{u} = \sum_{i=1}^{n}(z_i - \bar{z})\mathbf{x}_i$ gives some group-sensitive geometric "average" of $\mathbf{x} \in \mathcal{X}$. We can subsequently rewrite (ϵ-fair-SVM1-D) as

$$\begin{aligned}
\underset{\boldsymbol{\mu},\beta_-,\beta_+}{\text{maximize}} \quad & -\frac{1}{2}\left\|\sum_{i=1}^{n}\mu_i y_i (I-P_{\mathbf{u}})\mathbf{x}_i\right\|^2 + \sum_{i=1}^{n}\mu_i \\
& + \frac{2n\sum_i \mu_i y_i\langle \mathbf{x}_i,\mathbf{u}\rangle + n^2(\beta_- - \beta_+)}{2\|\mathbf{u}\|^2}(\beta_- - \beta_+) \\
\text{subject to} \quad & \mu_i \in [0,C], \qquad i=1,\ldots,n, \\
& \sum_{i=1}^{n}\mu_i y_i = 0, \qquad\qquad (\epsilon\text{-fair SVM2-D}) \\
& \beta_-,\beta_+ \geq 0, \\
& \beta_- + \beta_+ = \epsilon
\end{aligned}$$

where $I, P_{\mathbf{u}} \in \mathbb{R}^{d\times d}$. The former is the identity matrix, and the latter is the projection matrix onto the vector $\mathbf{u}$. As was also observed by Donini et al., the $\epsilon = 0$ version of (ϵ-fair SVM2-D) is equivalent to the standard formulation of the dual SVM program with Kernel $K(\mathbf{x}_i,\mathbf{x}_j) = \langle (I-P_{\mathbf{u}})\mathbf{x}_i, (I-P_{\mathbf{u}})\mathbf{x}_j\rangle$ [17].

Since we are interested in the welfare impacts of fair learning when fairness constraints *do* have an impact on optimal solutions, we will assume that the fairness constraint binds. For clarity of exposition, we assume that the positive covariance constraint binds, and thus that $\beta_- = 0$ and $\beta_+ = \epsilon$ in (ϵ-fair SVM2-D). This is without loss of generalization—the same analyses apply when the negative covariance constraint binds. The dual ϵ-fair SVM program becomes

$$\begin{aligned}
\underset{\boldsymbol{\mu}}{\text{minimize}} \quad & \frac{1}{2}\left\|\sum_{i=1}^{n}\mu_i y_i (I-P_{\mathbf{u}})\mathbf{x}_i\right\|^2 - \sum_{i=1}^{n}\mu_i + \frac{n\epsilon(2\sum_i \mu_i y_i\langle \mathbf{x}_i,\mathbf{u}\rangle - n\epsilon)}{2\|\mathbf{u}\|^2} \\
\text{subject to} \quad & \mu_i \in [0,C], \qquad i=1,\ldots,n, \qquad (\epsilon\text{-fair SVM-D}) \\
& \sum_{i=1}^{n}\mu_i y_i = 0
\end{aligned}$$

6.2 Algorithms

6.2.1 *Finding the next breakpoint when $|\mathcal{M}^\epsilon| = 0$.*

When $|\mathcal{M}^\epsilon| = 0$, the standard procedure that finds the next breakpoint by computing sensitivities to μ_i in the margin ($i \in \mathcal{M}^\epsilon$) by inverting the matrix K in (12) fails. Without r_i^ϵ, we also cannot compute changes to d_i for i not in the margin ($i \in \{\mathcal{F},\mathcal{E}\}^\epsilon$) as defined in (18) to track when points enter the margin. As a result, we need a special procedure to find the next breakpoint when the margin becomes empty.

If the solution is to remain optimal, it must continue to abide by KKT conditions; in particular $\sum_{i=1}^{n}\mu_i y_i = 0$. Notice then that if the margin is empty, we have that $\sum_{i\in\mathcal{E}^\epsilon}\mu_i y_i = 0 = C\sum_{i\in\mathcal{E}^\epsilon} y_i$, which means that there are equal numbers of +1 and −1 vectors that are misclassified. Thus at the next breakpoint, both +1 and −1 vectors will enter the margin at the same time, offsetting each other exactly to retain the optimality of the solution.

Tracking how vectors enter the margin at the solution $p(\epsilon)$ requires tracking sign changes of $\frac{\partial D^\epsilon}{\partial \mu}$:

$$\sum_{i=1}^{n}\mu_i y_i (I-P_{\mathbf{u}})\mathbf{x}_i y_j (I-P_{\mathbf{u}})\mathbf{x}_j + \frac{n\epsilon y_j\langle \mathbf{x}_j,\mathbf{u}\rangle}{\|\mathbf{u}\|^2} + b y_j - 1 \underset{\mathcal{E}^\epsilon}{\overset{\mathcal{F}^\epsilon}{\gtrless}} 0$$

We can perturb ϵ by $\Delta\epsilon$ and narrow the range of eligible optimal b. Consider how the SVM boundary splits the dataset. On the positive side of the boundary, we have

$$b > y_i\left(1 - \sum_{i=1}^{n}\mu_i y_i (I-P_{\mathbf{u}})\mathbf{x}_i y_j (I-P_{\mathbf{u}})\mathbf{x}_j - \frac{n\epsilon y_j\langle \mathbf{x}_j,\mathbf{u}\rangle}{\|\mathbf{u}\|^2}\right)$$

for i with $y_i = +1$ and $y_i \in \mathcal{F}^\epsilon$, as well as $y_i = -1$ and $y_i \in \mathcal{E}^\epsilon$. Call this set of indices R. On the other hand,

$$b < y_i\left(1 - \sum_{i=1}^{n}\mu_i y_i (I-P_{\mathbf{u}})\mathbf{x}_i y_j (I-P_{\mathbf{u}})\mathbf{x}_j - \frac{n\epsilon y_j\langle \mathbf{x}_j,\mathbf{u}\rangle}{\|\mathbf{u}\|^2}\right)$$

for i with $y_i = -1$ and $y_i \in \mathcal{F}^\epsilon$, as well as $y_i = +1$ and $y_i \in \mathcal{E}^\epsilon$. Call this set of indices L. Let

$$s(\epsilon) = 1 - \sum_{i=1}^{n}\mu_i y_i (I-P_{\mathbf{u}})\mathbf{x}_i y_j (I-P_{\mathbf{u}})\mathbf{x}_j - \frac{n\epsilon y_j\langle \mathbf{x}_j,\mathbf{u}\rangle}{\|\mathbf{u}\|^2}$$

.Then we have the range

$$b \in \mathcal{B}_\epsilon = [\max_{i\in R} y_i s(\epsilon), \min_{i\in L} y_i s(\epsilon)] \tag{24}$$

Perturbations of $\Delta\epsilon$ result in changes of

$$t(\Delta\epsilon) = -y_i\frac{n\Delta\epsilon\langle \mathbf{x}_j,\mathbf{u}\rangle}{\|\mathbf{u}\|^2}$$

so we can write

$$\mathcal{B}_\epsilon(\Delta\epsilon) = [\max_{i\in R} y_i s(\epsilon) - t(\Delta\epsilon), \min_{i\in L} y_i s(\epsilon) - t(\Delta\epsilon)] \tag{25}$$

In increasing the magnitude of $\Delta\epsilon$, the interval $\mathcal{B}_\epsilon(\Delta\epsilon)$ shrinks until it collapses onto a single value of b. The $\Delta\epsilon$ be the perturbation when

$$\max_{i\in R} y_i s(\epsilon) - t(\Delta\epsilon) = \min_{i\in L} y_i s(\epsilon) - t(\Delta\epsilon) \tag{26}$$

determines the next breakpoint. The indices

$$k = \arg\max_{i\in R} y_i s(\epsilon) - t(\Delta\epsilon), \qquad \ell = \arg\min_{i\in L} y_i s(\epsilon) - t(\Delta\epsilon) \tag{27}$$

leave their respective sets and enter the margin. The partition is updated as:

$$\mathcal{M}^{\epsilon+\Delta\epsilon} = \{k,\ell\} \tag{28}$$

$$\{\mathcal{F},\mathcal{E}\}^{\epsilon+\Delta\epsilon} = \{\mathcal{F},\mathcal{E}\}^{\epsilon} - \{k,\ell\} \tag{29}$$

6.3 Additional Figures

Figure 2 gives more information on the welfare impacts of ϵ-fair SVM-solutions on the Adult dataset. Increasing ϵ from left to right loosens fairness constraint, and classification outcomes become "less fair." Paths level off at $\epsilon \approx 0.175$ when constraint ceases to bind at the optimal solution. The top panel shows that the learner objective value monotonically decreases as the fairness constraint loosens. The bottom panel gives the group-specific welfare change at an ϵ-fair SVM solution given as an absolute change in the number of positively labeled examples compared to the unconstrained solution baseline.

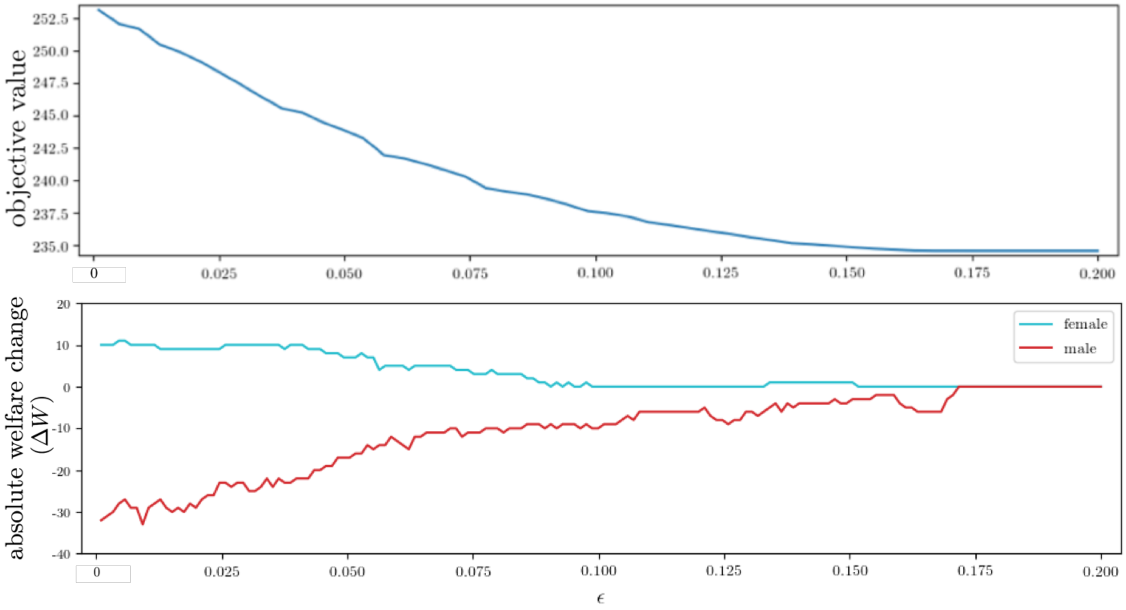

Figure 2: Impact of fair SVM learning on learner objective value (top panel) and group welfare given as absolute welfare changes for female and male groups (bottom panel) on the Adult dataset.

6.4 Results on the Correspondence between Loss Minimization and Social Welfare Maximization

In the Planner's Problem, a planner maximizes a social welfare functional (SWF) given as a weighted sum of individual utilities, $W = \sum_{i=1}^{n} w_i u_i$. An individual i's contribution to society's total welfare is a product of her utility u_i and her social weight $w_i \in [0, 1]$ normalized so that $\sum_i^n w_i = 1$. Utility functions $u_i : \mathcal{X} \to \mathbb{R}_+$ assign positive utilities to a set of attributes or goods $\mathbf{x}_i$. We suppose a utility function is everywhere continuous and differentiable with respect to its inputs.

Since a planner who allocates a resource h impacts her recipients' utilities, she solves $h^{SWF}(\mathbf{x}; \boldsymbol{w}) := \arg\max_{\boldsymbol{h}} \sum_{i=1}^{n} w_i u(\mathbf{x}_i, h_i)$ under a budget constraint: $\sum_{i=1}^{n} h_i \leq B$. Since we consider cases of social planning in which a desirable good is being allocated, it is natural to suppose that u is strictly monotone with respect to h. As is common in welfare economics, we take u to be concave in h, so that receiving the good exhibits diminishing marginal returns. Further, we require that the social welfare functional W be symmetric: $W(\boldsymbol{h}; \mathbf{x}, \boldsymbol{w}) = W(\sigma(\boldsymbol{h}); \sigma(\mathbf{x}), \sigma(\boldsymbol{w}))$ for all possible permutations of $\sigma(\cdot)$. This property implies that the utility functions in the Planner's Problem are not individualized. In the case of binary classification, the planner decides whether to allocate the discrete good to individual i or not ($h_i \in \{0, 1\}$).

To highlight the correspondence between the machine learning and welfare economic approaches to social allocation, we first show that we can understand loss minimizing solutions to also be welfare maximizing ones, albeit under a particular instantiation of the social welfare function. Since social welfare is given as the weighted sum of individuals' utilities, it is clear that manipulating weights $\boldsymbol{w}$ significantly alters the planner's solution. Thus just as we can compute optimal allocations under a fixed set of welfare weights, we can also begin with an optimal allocation and find welfare weights that would support them. In welfare economics, the form of $\boldsymbol{w}$ corresponds to societal preferences about what constitutes a fair distribution. For example, the commonly-called "Rawlsian" social welfare function named after political philosopher John Rawls, can be written as $W_{Rawls} = \min_i u_i$ where u_i gives the utility of individual i. This function is equivalent to the general form $\sum_{i=1}^{n} w_i u_i$ where the individual i with the lowest utility u_i has welfare weight $w_i = 1$ and all individuals $k \neq i$ have weight $w_k = 0$. On the other hand, the commonly-called "Benthamite" social welfare function named after the founder of utilitarianism Jeremy Bentham, aggregates social welfare such that an extra unit of utility contributes equally to the social welfare regardless of who receives it. Benthamite weights are equal across all individuals: $w_i = \frac{1}{n}$ for all $i \in [n]$.

Thus associating an optimal (possibly fairness constrained) loss minimizing allocation with a set of welfare weights that would make it socially optimal lends insight into how socially "fair" a classification is from a welfare economic perspective. The following Proposition formally states this correspondence between loss minimization and social welfare maximization.

Proposition 6.1. *For any vector of classifications $h^{ML}(\mathbf{x}_i)$ that solves a loss minimization task, there exists a set of welfare weights $\boldsymbol{w}$ with $\sum_{i=1}^{n} w_i = 1$ such that the planner who maximizes social welfare W with a budget B selects an optimal allocation $h^{SWF}(\mathbf{x}_i) = h^{ML}(\mathbf{x}_i)$ for all $i \in [n]$.*

Proof. First, we know that since $W(\mathbf{x}, \boldsymbol{w})$ is a weighted sum of functions u, which are concave in h, the planner can indeed find a social welfare maximizing allocation $\boldsymbol{h}^{SWF}$. Let $h^{ML}(\mathbf{x})$ be the empirical loss-minimizing classifier for $\{\mathbf{x}_i, z_i, y_i\}_{i=1}^n$. With these allocations given, we can invert the social welfare maximization problem to find the weights that $\boldsymbol{w}$ support them.

For a given utility function u, we evaluate $\frac{\partial u(\mathbf{x},h)}{\partial h}\Big|_{\{\mathbf{x}_i, h^{ML}(\mathbf{x}_i)\}} = m_i \;\forall i \in [n]$, which gives the marginal gain in utility for individual i from having received an infinitesimal additional allocation of h. Notice that at a welfare maximizing allocation $\boldsymbol{h}$, we must have that

$$w_i \frac{\partial u(\mathbf{x},h)}{\partial h}\Big|_{\{\mathbf{x}_i,h_i\}} = w_j \frac{\partial u(\mathbf{x},h)}{\partial h}\Big|_{\{\mathbf{x}_j,h_j\}} \quad \text{for all } i,j \in [n] \tag{30}$$

When the allocation $h^{ML}(\mathbf{x})$ has been fixed, we must have that $w_i m_i = w_j m_j = k$, where the constant k is set by the planner's budget B, for all i, j along with $\sum_{i=1}^n w_i = 1$. Since u is strictly monotone with respect to h, $m_i > 0$ for all i. We thus have a non-degenerate system of n equations with n variables, and there exists a unique solution of welfare weights $\boldsymbol{w}$ that support the allocation. □

Note that in the case of binary classification $h^{ML}(\mathbf{x}) \in \{-1, +1, \}$, so allocations are not awarded at a fractional level. Thus rather than the partial $\frac{\partial u(\mathbf{x},h)}{\partial h}$, the planner must consider the margin gain of receiving a positive classification. Nevertheless, Proposition 1 still holds, and the proof carries through with $\Delta u(\mathbf{x}, h(\mathbf{x})) = u(\mathbf{x}, 1) - u(\mathbf{x}, 0)$ in place of partial derivatives $\frac{\partial u(\mathbf{x},h)}{\partial h}$.

The equations given in (30) set an optimality condition for the planner. Its structure, though simple, reveals that welfare weights must be inversely proportional to an individuals' marginal utility gain from receiving an allocation. This result is formalized in the Proposition below.

Proposition 6.2. *For any set of optimal allocations* $\boldsymbol{h} = \arg\max_{\boldsymbol{h}} \sum_{i=1}^n \bar{w}_i u(\mathbf{x}_i, h_i)$ *with strictly monotonic utility function u concave in h, the supporting welfare weights have the form* $\bar{w}_i = \frac{k}{m_i}$ *where* $m_i = \frac{\partial u(\mathbf{x}_i)}{\partial h}|_{\{\mathbf{x}_i,h_i\}}$ *and $k > 0$ is a constant set by the planner's budget* $B = \sum_{i=1}^n h_i$.

By associating a set of classification outcomes with a set of implied welfare weights, one can inquire about the social fairness of the allocation scheme by investigating the distribution of welfare weights across individuals or across groups. While there may not be a single distribution of welfare weights that can be said to be "most fair," theoretical and empirical work in economics has been conducted on the range of fair distributions of societal weights [29, 30]. This research has considered weights as implied by current social policies [31–33], philosophical notions of justice [34, 35], and individuals' preferences in surveys and experiments [30, 31, 36]. They thus offer substantive notions of fairness currently uncaptured by many current algorithmic fairness approaches.

6.4.1 *An Algorithm that Records All Possible Labelings.*

In the previous section, we showed that for any vector of classifications, one can compute the implied societal welfare weights of the generic SWF that would yield the same allocations in the Planner's Problem. In this section, we work in the converse direction: Beginning with a planner's social welfare maximization problem, does there exist a classifier $h^{ML} \in \mathcal{H}$ that generates the same classification as the planner's optimal allocation such that for all $i \in [n]$, $h^{ML}(\mathbf{x}_i) = h^{SWF}(\mathbf{x}_i)$?

ALGORITHM 2: Record all possible labelings on a dataset $\mathcal{X}$ by linear separators

```
Input: Set $\mathcal{X}$ of $n$ data points $\mathbf{x} \in \mathbb{R}^d$
Output: All possible partitions $A$, $B$ attainable via linear separators;
        supporting hyperplane $h$
for all $V \subset \mathcal{X}$ with $|V| = d$ do
    Construct $d-1$-dimensional hyperplane $h_V$ defined by $\mathbf{v} \in V$;
    for each point $\mathbf{v} \in V$ do
        $P = V \setminus \mathbf{v}$;
        $h = pivot(h_V, P, \mathbf{v})$ ;          // $h_V$ pivots around the
          $d-2$-dimensional plane $P$ away from $\mathbf{v}$
        $h = translate(h, \mathbf{v})$ ;          // $h$ translates toward $\mathbf{v}$
        Record $A = \{\mathbf{x}|\mathbf{x} \in h^+\}$, $B = \{\mathbf{x}|\mathbf{x} \in h^-\}$, $h$;
    end
end
```

We answer this question for the hypothesis class of linear decision boundary-based classifiers by providing an algorithm that accomplishes a much more general task: Given a set $\mathcal{X}$, containing n d-dimensional nondegenerate data points $\mathbf{x} \in \mathbb{R}^d$, our algorithm enumerates all linearly separable labelings and can output a hyperplane parameterized by $\boldsymbol{\theta} \in \mathbb{R}^d$ and $b \in \mathbb{R}$ that achieves that set of labels. In order to build intuition for its construction, we first consider a hyperplane separation technique that applies to a very specific case: a case in which a hyperplane separates sets A and B, intersecting A at a single point and intersecting B at $d-1$ points.

Lemma 6.3. *Consider linearly separable sets A and B of points* $\mathbf{x} \in \mathbb{R}^d$. *For any $d-1$-dimensional hyperplane h_V with* $h_V \cap A = \mathbf{v}$ *and* $h_V \cap B = P$ *where* $|P| = d-1$ *that separates A and B into closed halfspaces* $\bar{h}_V^+$ *and* $\bar{h}_V^-$, *one can construct a $d-1$-dimensional hyperplane h that separates A and B into open halfspaces h^+ and h^-.*

Because its techniques are not of primary relevance for this Section, we defer the full proof of this Lemma to the Appendix but provide a brief exposition. The construction on which the Lemma relies is a "pivot-and-translate" maneuver. A hyperplane as described can separate points in open halfspaces by first pivoting (infinitesimally) on a $d-2$-dimensional facet P of a convex hull $C(B)$ away from $\mathbf{v} \in C(A)$ and then translating (infinitesimally) back toward $\mathbf{v}$ and away from $C(B)$. We show that all separable convex sets can be separated by such a hyperplane and procedure.

Note that since we seek enumerations of all labelings achievable by a linear separator on a given dataset, we are not *a priori* given convex hulls to separate. That is, we want to know which points *can* be made into distinct convex hulls and which cannot. Thus we take the preceding procedure and invert it—the central idea is to begin with the separators and from there, search for all possible convex hulls: Beginning with an arbitrary $d-1$-dimensional hyperplane h defined by d data points, we construct convex hulls out of the points in each halfspace created by h. Then we can use the pivot-and-translate procedure to construct a separation of the two sets into two open halfspaces. We must show that such a procedure is indeed exhaustive.

Theorem 6.4. *Given a dataset $\mathcal{X}$ consisting of n nondegenerate points $\mathbf{x} \in \mathbb{R}^d$, Algorithm 2 enumerates all possible labelings achievable by a $d-1$-dimensional hyperplane in $O(n^d d)$ time and outputs hyperplane parameters $(\boldsymbol{\theta}, b)$ that achieve each one.*

Proof. We have already shown that the pivot-and-translate construction is sufficient to linearly separate two sets A and B in the very specific case given in the preceding Lemma. But we must prove that all linearly separable sets can be constructed via Algorithm 2. We prove it is exhaustive by contradiction.

Suppose there exists a separation of $\mathcal{X}$ that is not captured by Algorithm 2. Then there exists disjoint sets A and B such that their convex hulls $C(A)$ and $C(B)$ do not intersect. By the hyperplane separating theorem, there exists a $d-1$-dimensional hyperplane h_{V_1} that separates A and B, defined by a set V_1 of d vertices $\mathbf{v}$, at least one of which is on the boundary of each convex hull. Without loss of generality, we assume that for all $\mathbf{x} \in A$, $\mathbf{x} \in h^+_{V_1}$ and for all $\mathbf{x} \in B$, $\mathbf{x} \in h^-_{V_1}$. Notice that this hyperplane is indeed "checked" by the Algorithm, and this hyperplane h_{V_1} correctly separates $\mathbf{x} \in \mathcal{X} \setminus V_1$ into the two sets A and B. Thus if the separation is not disclosed via the procedure, the omission must occur due to the pivot-and-translate procedure's being incomplete.

In Algorithm 2, the set V_1 is partitioned so that $V_1 = \mathbf{v}_{f,1} \cup P_1$ where $\mathbf{v}_{f,1}$ is the "free vertex" and P_1 is the pivot set consisting of $d-1$ vertices. This partition occurs d times so that each vertex $\mathbf{v} \in V_1$ has its turn as the "free vertex." Thus we can view the pivot-and-translate procedure as constituting a second partition—a partition of the d vertices that define the initial separating hyperplane. By contradiction, we claim that there exists a partition $D_1, E_1 \subset V_1$ such that $D_1 \coprod E_1 = V_1$ where $D_1 \subset A$ and $E_1 \subset B$ that is unaccounted for in the d pivot-and-translate operations applied to h_{V_1}. Thus $|D_1|, |E_1| \geq 2$. We use a "gift-wrapping" argument, a technique common in algorithms that construct convex hulls, to show that the partition A and B is indeed covered by Algorithm 2.

Select $\mathbf{v} \in D_1$ to be the free vertex $\mathbf{v}_{f,1}$, and let the pivot set $P_1 = V_1 \setminus \mathbf{v}_{f,1}$. We pivot around P_1 and away from $\mathbf{v}_{f,1}$ so that $\mathbf{v}_{f,1} \in h^+_{V_1}$. Rotations in d-dimensions are precisely defined as being around $d-2$-dimensional planes. Thus pivoting around the ridge P_1 away from $\mathbf{v}_{f,1}$ is a well-defined rotation in $\mathbb{R}^d$. Since h_{V_1} is a supporting hyperplane to $C(B)$, E_1 constitutes a $|E_1|-1$-dimensional facet of $C(B)$. There exists a vertex $\mathbf{v}_E \in C(B)$ such that $E_1 \cup \mathbf{v}_E$ gives a $|E_1|$-dimensional facet of $C(B)$. Let h_{V_2} be defined by the set $V_2 = P_1 \cup \mathbf{v}_E$. h_{V_2} continues to correctly separate all $\mathbf{x} \in \mathcal{X} \setminus V_2$.

We once again partition V_2 into sets D_2 and E_2 whose members must be ultimately classified in sets A and B respectively. Notice that $|D_2| = |D_1| - 1$, since h_{V_2} correctly classifies $\mathbf{v}_{f,1}$ as belonging to set A. Thus with each iteration of the pivot procedure, the separating classifier unhinges from a vertex in $C(A)$ and "wraps" around $C(B)$ just as in the gift wrapping algorithm to attach onto another vertex in $C(B)$. At each step, the hyperplane defined by d vertices continues to support and separate $C(A)$ and $C(B)$. Thus process iterates until in the $|D_1|-1$-th round, the hyperplane $h_{V_{|D_1|-1}}$ has partition $D_{|D_1|-1}$ and $E_{|D_1|-1}$ with $\left|D_{|D_1|-1}\right| = 1$. Applying the full pivot-and-translate procedure ensures the desired separation of sets A and B into open halfspaces.

Thus starting from a separable hyperplane defined by d vertices on the convex hulls $C(A)$ and $C(B)$, which must exist in virtue of the separability of sets A and B, we were able to use the pivot procedure in order to "gift-wrap" around one convex hull until we arrived at a d-dimensional separating hyperplane with only one vertex $\mathbf{v}_f \in C(A)$. This hyperplane is obviously checked by the first for-loop of Algorithm 2. The subsequent for-loop that performs the second partition of the d vertices into the free vector $\mathbf{v}_f$ and the pivot set P then directly applies and performs the pivot-and-translate procedure given in Algorithm 2 to achieve the desired separation. □

Degeneracies in the dataset can be handled by combining Algorithm 2 with standard solutions to degeneracy problems in geometric algorithms, which perform slight perturbations to degenerate data points to transform them into nondegenerate ones [37]. In concert with these solutions, Algorithm 2 automatically reveals which social welfare maximization solutions are attainable on a given dataset $\mathcal{X}$ via hyperplane-based classification and the $0-1$ accuracy loss each entails.

6.5 Proofs

6.5.1 *Proof of Proposition 3.3.*

Proof. For all $j \in \mathcal{F}^\epsilon$, remaining in $\mathcal{F}^{\epsilon+\Delta\epsilon}$ after the perturbation requires that $\frac{\partial D}{\partial \mu_j} > 0$ after the perturbation. Let μ_i^ϵ be the optimal μ_i solution at $p(\epsilon)$. Then following (10), we rewrite the quantity $\frac{\partial D}{\partial \mu_j}$ as

$$g_j = 1 - \left(\sum_{i=1}^{n} \mu_i^\epsilon y_i (I - P_{\mathbf{u}}) \mathbf{x}_i y_j (I - P_{\mathbf{u}}) \mathbf{x}_j + \frac{n\epsilon y_j \langle \mathbf{x}_j, \mathbf{u} \rangle}{\|\mathbf{u}\|^2} + b y_j \right) < 0$$

If $d_j \Delta\epsilon > 0$, then $j \in \mathcal{F}^{\epsilon+\Delta\epsilon}$. Otherwise, for $d_j \Delta\epsilon < 0$, if $\Delta\epsilon < \frac{g_j}{d_j}$, then $\frac{\partial D}{\partial \mu_j^{\epsilon+\Delta\epsilon}} > 0$, and $j \in \mathcal{F}^{\epsilon+\Delta\epsilon}$ after the perturbation. ✓
The same reasoning follows for $j \in \mathcal{E}^\epsilon$, except we have that $g_j > 0$. Thus if $d_j \Delta\epsilon < 0$, then $j \in \mathcal{E}^{\epsilon+\Delta\epsilon}$. Otherwise, for $d_j \Delta\epsilon > 0$, if $\Delta\epsilon < \frac{g_j}{d_j}$, then $\frac{\partial D}{\partial \mu_j^{\epsilon+\Delta\epsilon}} > 0$, and $j \in \mathcal{E}^{\epsilon+\Delta\epsilon}$ after the perturbation. ✓

To ensure that margin vectors do not escape the margin, we can directly look to $r_j = \frac{\partial \mu_j}{\partial \epsilon}$. Since for all $j \in \mathcal{M}^\epsilon$, $\mu_j^\epsilon \in [0, C]$, then staying in the margin and set $\mathcal{M}^{\epsilon+\Delta\epsilon}$ depends on the sign of r_j and requires that

$$r_j < 0 \longrightarrow \frac{C - \mu_j^\epsilon}{r_j} < \Delta\epsilon < \frac{-\mu_j^\epsilon}{r_j} \tag{31}$$

$$r_j > 0 \longrightarrow \frac{-\mu_j^\epsilon}{r_j} < \Delta\epsilon < \frac{C - \mu_j^\epsilon}{r_j} \tag{32}$$

Thus taking the minimum of the positive quantities gives an upper bound, while taking the maximum of the negative quantities gives a lower bound on $\Delta\epsilon$ perturbations, such that $\{\mathcal{F}, \mathcal{M}, \mathcal{E}\}^\epsilon =$

$\{\mathcal{F}, \mathcal{M}, \mathcal{E}\}^{\epsilon+\Delta\epsilon}$. Let

$$m_j = \begin{cases} \begin{cases} \frac{g_j}{d_j}, & j \in \mathcal{F}, d_j > 0 \\ -\infty, & j \in \mathcal{F}, d_j < 0 \end{cases} \\ \min\{\frac{C-\mu_j^\epsilon}{r_j}, \frac{-\mu_j^\epsilon}{r_j}\}, j \in \mathcal{M} \\ \begin{cases} -\infty, & j \in \mathcal{E}, d_j > 0 \\ \frac{g_j}{d_j}, & j \in \mathcal{E}, d_j < 0 \end{cases} \end{cases} , \quad M_j = \begin{cases} \begin{cases} \infty, & j \in \mathcal{F}, d_j > 0 \\ \frac{g_j}{d_j}, & j \in \mathcal{F}, d_j < 0 \end{cases} \\ \min\{\frac{C-\mu_j^\epsilon}{r_j}, \frac{-\mu_j^\epsilon}{r_j}\}, j \in \mathcal{M} \\ \begin{cases} \frac{g_j}{d_j}, & j \in \mathcal{E}, d_j > 0 \\ \infty, & j \in \mathcal{E}, d_j < 0 \end{cases} \end{cases}$$

Thus all perturbations of ϵ within the range

$$\Delta\epsilon \in \Big(\max_j m_j, \min_j M_j \Big)$$

satisfy the necessary conditions to ensure stable sets $\{\mathcal{F}, \mathcal{M}, \mathcal{E}\}$. Stable classifications $\hat{y}_i$ follow. □

6.5.2 *Proof of Corollary 3.4.*

Proof. For all $\Delta\epsilon$ in the stable region given in (16), $W_i(\epsilon) = W_i(\epsilon + \Delta\epsilon)$ where i gives group membership $z = i$. Thus the groups are welfare-wise indifferent between classifications at ϵ and $\Delta\epsilon$. For all $\Delta\epsilon < 0$, where the fairness constraint is tightened,$p(\epsilon) \leq p(\epsilon + \Delta\epsilon)$. Since the learner prefers lower loss, we have that $p(\epsilon) \geq p(\epsilon + \Delta\epsilon)$. Comparing the triples at each ϵ value, we thus have

$$\{p(\epsilon), W_0(\epsilon), W_1(\epsilon)\} \geq \{p(\epsilon + \Delta\epsilon), W_0(\epsilon + \Delta\epsilon), W_1(\epsilon + \Delta\epsilon)\}$$

as desired. □

6.5.3 *Proof of Proposition 3.8.*

Proof. Following much of the exposition in the main text, recall we have that the perturbation function in (21) is given as

$$p(\epsilon) \geq \sup_{\boldsymbol{\mu}, \gamma} \{\mathcal{L}(\boldsymbol{\mu}^*, \gamma^*) - \epsilon|\gamma^*|\}$$

which gives a global lower bound. Thus when a perturbation $\Delta\epsilon < 0$ causes $\mathcal{L}(\boldsymbol{\mu}^*, \gamma^*) - \epsilon|\gamma^*|$ to increase, then $p(\epsilon + \Delta\epsilon)$ is guaranteed to increase by at least $\Delta\epsilon|\gamma^*|$. Thus when $|\gamma^*| \gg 0, p(\epsilon+\Delta\epsilon) - p(\epsilon) \gg 0$. The learner experience a significant increase in her optimal value $p(\epsilon)$ (which she wishes to minimize).

On the other hand, when $\Delta\epsilon > 0$, then $\mathcal{L}(\boldsymbol{\mu}^*, \gamma^*) - \epsilon|\gamma^*|$ decreases. But the decrease gives only the lower bound, and thus when $|\gamma^*|$ is small, her optimal value $p(\epsilon)$ decreases but it is guaranteed not to decrease by much. □

6.5.4 *Proof of Proposition 3.5.*

Proof. Fix $\epsilon \in (0, 1)$ and consider the stable region of $\Delta\epsilon$ perturbations given by (b_L, b_U). Suppose $b_L = \frac{g_j}{d_j}$ with $j \in \mathcal{E}$, then if $y_j = -1$, $\hat{y_j} = +1$. Thus at the breakpoint $\Delta\epsilon = b_L$, j moves into $\mathcal{M}^{\epsilon+b_L}$ and $\hat{y_j} = +1$ and $u_{z_j}(\epsilon + b_L) < u_{z_j}(\epsilon)$ where z_j gives the group membership of $\mathbf{x}_j$. Since no other points transition, $u_{\bar{z}}(\epsilon + b_L) = u_{\bar{z}}(\epsilon)$ for all $\bar{z} \neq z_j$. Since $b_L < 0$, the fairness constraint is tightened and associated with a shadow price given by $\gamma > 0$ such that $p(\epsilon + b_L) < p(\epsilon)$. ✓

Suppose $b_L = \frac{C-\mu_j^\epsilon}{r_j}$ and $j \in \mathcal{M}^\epsilon$ with $y_j = +1$, then j moves into $j \in \mathcal{E}^{\epsilon+b_L}$ such that $\hat{y_j} = -1$. Thus $u_{z_j}(\epsilon + b_L) < u_{z_j}(\epsilon)$ and $u_{\bar{z}}(\epsilon + b_L) = u_{\bar{z}}(\epsilon)$ where z_j is the group membership of $\mathbf{x}_j$ and $\bar{z} \neq z_j$, and $p(\epsilon + b_L) \leq p(\epsilon)$. ✓

Suppose $b_U = \frac{g_j}{d_j} > 0$ where $j \in \mathcal{E}^\epsilon$, $y_j = +1$, and $\hat{y_j} = -1$. At the breakpoint, j moves into $\mathcal{M}^{\epsilon+b_U}$ such that $y_j = -1$. Then $u_{z_j}(\epsilon + b_U) > u_{z_j}(\epsilon)$ where z_j is the group membership of $\mathbf{x}_j$. For $\bar{z} \neq z_j$, $u_{\bar{z}}(\epsilon+b_U) = u_{\bar{z}}(\epsilon)$, and since $b_U > 0$, the fairness constraint is loosened and $p(\epsilon + b_U) > p(\epsilon)$.

Suppose $b_U = \frac{C-\mu_j^\epsilon}{r_j} > 0$ where $j \in \mathcal{M}^\epsilon$ and $y_j = -1$. At the breakpoint, j moves into $\mathcal{E}^{\epsilon+b_U}$ such that $\hat{y_j} = +1$. Then $u_{z_j}(\epsilon + b_U) > u_{z_j}(\epsilon)$ where z_j gives the group membership of $\mathbf{x}_j$. For $\bar{z} \neq z_j$, $u_{\bar{z}}(\epsilon + b_U) = u_{\bar{z}}(\epsilon)$, and since $b_U > 0$, the fairness constraint is loosened and $p(\epsilon + b_U) \geq p(\epsilon)$. ✓ □

6.5.5 *Proof of Theorem 3.6.*

Proof. Theorem 3.6 follows from Lemma 3.2, Proposition 3.3, Corollary 3.4, and Proposition 3.5. □

6.5.6 *Proof of Lemma 6.3 from Appendix Section 6.4.*

Proof. Let A and B be a pair of disjoint non-empty convex sets that partition $\mathcal{X} \subset \mathbb{R}^d$: $A \coprod B = \mathcal{X}$. Then by the hyperplane separation theorem, there exists a pair $(\boldsymbol{\theta}, b)$ such that for all $\mathbf{x} \in A$, $\boldsymbol{\theta}^\intercal \mathbf{x} \geq b$—call this closed halfspace $\bar{h}^+$—and for all $\mathbf{x} \in B$, $\boldsymbol{\theta}^\intercal \mathbf{x} \leq b$—call this closed halfspace $\bar{h}^-$. One such hyperplane can be constructed to separate the convex hulls of A and B

$$C(A) = \Big\{ \sum_{i=1}^{|A|} \alpha_i \mathbf{x}_i | \mathbf{x}_i \in A, \alpha_i \geq 0, \sum_{i=1}^{|A|} \alpha_i = 1 \Big\}$$

$$C(B) = \Big\{ \sum_{i=1}^{|B|} \alpha_i \mathbf{x}_i | \mathbf{x}_i \in B, \alpha_i \geq 0, \sum_{i=1}^{|B|} \alpha_i = 1 \Big\}$$

Let h_V be the $d-1$-dimensional hyperplane defined by the set V with $|V| = d$ such that $V \cap C(A) \neq \emptyset$ and $V \cap C(B) \neq \emptyset$. In order for the hyperplane to separate $C(A)$ and $C(B)$, h_V must also support each hull—we know that such a hyperplane always exists. In order to separate $C(A)$ and $C(B)$ so they are contained within open halfspaces h_V^+ and h_V^-, we wiggle the hyperplane so that it no longer passes through vertices $\mathbf{v} \in V$ but still maintains convex hull separation. This "wiggle" step is the final step of separating A and B.

Suppose V can be partitioned into a single vertex $\mathbf{v}_A$ in $C(A)$ and a set $P = \{\mathbf{v} | \mathbf{v} \in C(B)\}$ with $|P| = d - 1$. The set P defines a ridge on $C(B)$, since it is a $d-2$-dimensional facet of $C(B)$. Rotations in d-dimensions are precisely defined as being around $d-2$-dimensional planes. Thus pivoting h_V around the ridge P away from $\mathbf{v}_A$ is a well-defined rotation in $\mathbb{R}^d$. Selecting any infinitesimally small rotation angle ρ will be enough to have $C(A) \in h_V^+$. After the pivot, we translate h_V away from the ridge P back toward $\mathbf{v}_A$. An infinitesimal translation is sufficient, since we simply wish to dislodge h_V from the ridge P, so that $C(B) \in h_V^-$. □