Fairness through Optimization*

Violet (Xinying) Chen and J. N. Hooker

Carnegie Mellon University
xinyingc@andrew.cmu.edu, jh38@andrew.cmu.edu

January 2021

Abstract

We propose optimization as a general paradigm for formalizing fairness in AI-based decision models. We argue that optimization models allow formulation of a wide range of fairness criteria as social welfare functions, while enabling AI to take advantage of highly advanced solution technology. We show how optimization models can assist fairness-oriented decision making in the context of neural networks, support vector machines, and rule-based systems by maximizing a social welfare function subject to appropriate constraints. In particular, we state tractable optimization models for a variety of functions that measure fairness or a combination of fairness and efficiency. These include several inequality metrics, Rawlsian criteria, the McLoone and Hoover indices, alpha fairness, the Nash and Kalai-Smorodinsky bargaining solutions, combinations of Rawlsian and utilitarian criteria, and statistical bias measures. All of these models can be efficiently solved by linear programming, mixed integer/linear programming, or (in two cases) specialized convex programming methods.

Artificial intelligence is increasingly used not only to solve problems, but to recommend action decisions that range from awarding mortgage loans to granting parole. The prospect of making decisions immediately raises the question of ethics and fairness. If ethical norms are to be incorporated into artificial decision making, these norms must somehow be automated or formalized. The leading approaches to this challenge include

- *value alignment*, which strives to train or modify AI systems to reflect human ethical values automatically (Allen et al. [2005], Russell [2019], Gabriel [2020]);
- *logical formulations* of ethical and fairness principles that attempt to represent them precisely enough to govern a rule-based AI system (Bringsjord et al. [2006], Lindner et al. [2020], Hooker and Kim [2018]); and
- *statistical fairness metrics* that aim to ensure that benefits are allocated equitably (Dwork et al. [2012], Mehrabi et al. [2019], Chouldechova and Roth [2020]).

Each of these approaches can be useful in a suitable context. We wish to propose, however, a promising tool for formalizing ethics and fairness that has received less attention:

- *optimization*, which allows one to achieve equity or fairness by maximizing a *social welfare function*.

Welfare economics has long used social welfare functions to measure the desirability of a given distribution of benefits and costs. These functions can be designed to represent any of a wide range of conceptions of fairness and equity. They allow one to harness powerful optimization solvers, developed over a period of 80

*Relevant disciplines: fairness in AI, optimization methods, optimization modeling, ethics, welfare economics.

years or more, to achieve equity by maximizing social welfare—a task that can be carried out automatically by computer. Optimization methods are of course already employed in AI to train neural networks, calibrate support vector machines, solve clustering problems, and the like. Our proposal is to harness the power of optimization to identify ethical and fair decisions.

Achieving fairness by maximizing a social welfare function (SWF) offers at least two advantages in addition to the employment of optimization technology. One is that it allows one considerable flexibility to represent constraints on the problem. Decisions are normally made in the context of resource constraints or other limitations on possible options. These can be represented as constraints in the optimization problem, as nearly all state-of-the-art optimization methods are designed for constrained optimization. Also, a complex SWF can often be simplified by adding constraints to the optimization problem, resulting in a problem that is easier to solve.

Another advantage of maximizing a SWF is that equity and efficiency are naturally combined in an SWF. There is frequently an efficiency or utilitarian criterion along with an equity criterion in AI-based decision making, such as predictive accuracy or economic benefit. A properly designed SWF can combine utilitarian and fairness criteria in principled ways. For example, the proportional fairness SWF, already widely used in engineering, balances throughput (utility) and fairness in a fashion that can be given theoretical justification. One can of course maximize a pure efficiency objective subject to a constraint on inequality or some other measure of inequity, but this provides no principled way of balancing efficiency against equity. Formulation of a SWF encourages one to think explicitly about how they should be balanced, as well as allowing one to govern the equity/efficiency trade-off with one or more parameters. In any event, maximizing a SWF sacrifices no generality, because one can always implement a constraint on inequity by penalizing constraint violations in the SWF.

As a running example, we consider a bank's decision as to which applicants for mortgage loans should receive loans, based on individual credit profiles. We wish to identify decisions that maximize social welfare, subject to a constraint on the amount of funds available. Social welfare is a function of the utilities enjoyed by each of the stakeholders involved, including the applicants, the bank, and perhaps other parties such as stockholders and the community at large. The utilities are, in turn, a function of the funds allocated to each applicant. They can be measured as wealth, negative cost, or some broader type of benefit that is affected by the loan decisions. A neural network or support vector machine for the mortgage loan problem would normally be trained to maximize the accuracy of predicting loan defaults or their probability. This implies a concern with efficiency, since greater predictive accuracy results in a more efficient use of capital and perhaps greater welfare overall. The social welfare function can be designed to take into account the distribution of utilities as well as total net utility, thus balancing equity and efficiency. AI technology would then be designed to maximize social welfare rather than simply predictive accuracy. For example, machine learning (e.g. neural networks) could be used to predict the probability of default, and optimization then used to make loan decisions. Social welfare maximization can also be incorporated into a support vector machine or into purely rule-based AI.

Our contribution in this paper, aside from pointing out the potential of optimization as a general paradigm for achieving fairness, is to show how to formulate a number of fairness-related SWFs as tractable optimization problems. Much of the art of optimization is formulating the problem in a way that makes it suitable for existing solvers. We draw upon known modeling techniques for some of the SWFs and introduce new techniques for others. To our knowledge, most of these formulations do not appear in the AI literature or elsewhere.

The paper is organized as follows. In Section 1, we begin with introducing the optimization problem of maximizing a social welfare function in a general resource allocation context. We demonstrate concrete specifications of this social welfare optimization model on our running example of a bank's decision to grant mortgage loans, and state our assumptions on the linearity of constraints. In Section 2, we discuss related work on optimization methods for fair and ethical decision making that appears in the optimization, machine learning and AI literature, and distinguish this paper's contribution from previous research. We then study four types of fairness seeking schemes and present tractable optimization formulations for a variety of fairness measures. The first type we examine uses an inequality measure as the SWF and evaluates fairness based on the degree of equality in the utility distribution (Section 3). The second scheme defines fairness as giving a certain amount of priority to less advantaged stakeholders in the social welfare optimization problem (Section 4). The third type of methods handles the important challenge of balancing fairness and efficiency through optimizing SWFs that combine the two objectives (Section 5). Lastly, in Section 6, we discuss fairness in support vector machines to eliminate disparity in treatment received by different groups as well as to maximize social welfare in general.

1 The Basic Optimization Problem

The general problem of maximizing social welfare can be stated

$$\max_{\boldsymbol{x}} \big\{ W\big(\boldsymbol{U}(\boldsymbol{x})\big) \mid \boldsymbol{x} \in S_{\boldsymbol{x}} \big\} \tag{1}$$

where $\boldsymbol{x} = (x_1, \ldots, x_n)$ is a vector of resources distributed across stakeholders $1, \ldots, n$, and $\boldsymbol{U} = (U_1, \ldots, U_n)$ is a vector of utility functions corresponding to the stakeholders. Also $S_{\boldsymbol{x}}$ is the set of feasible values of $\boldsymbol{x}$, and W is a social welfare function. The problem maximizes social welfare over all feasible resource allocations.

It is convenient to model the utility functions $\boldsymbol{U}$ using constraints, because this results in problems better suited for optimization solvers. We therefore write (1) as

$$\max_{\boldsymbol{x},\boldsymbol{u}} \big\{ W'(\boldsymbol{u}) \mid (\boldsymbol{x}, \boldsymbol{u}) \in S_{\boldsymbol{xu}} \big\} \tag{2}$$

where $\boldsymbol{u}$ is a vector of utilities, and $S_{\boldsymbol{xu}}$ is defined so that $(\boldsymbol{x}, \boldsymbol{u}) \in S_{\boldsymbol{xu}}$ implies $\boldsymbol{x} \in S_{\boldsymbol{x}}$ and $\boldsymbol{u} \leq \boldsymbol{U}(\boldsymbol{x})$. The function W' is a possibly simplified version of W that yields an equivalent optimization problem due to constraints defining $S_{\boldsymbol{xu}}$. We illustrate this maneuver throughout the paper.

1.1 Example: Mortgage loans

In the mortgage problem described earlier, we can let I_A be the set of loan applicants and I_B the set of other stakeholders, such as the bank. Then x_i for $i \in I_A$ is the loan amount allocated to applicant i. The feasible set $S_{\boldsymbol{xu}}$ could be defined in part by the budget constraint $\sum_i x_i \leq H$, where H is the amount of available funds. We would also have constraints $x_i \leq h_i$, where $h_i = 0$ for $i \in I_B$, and h_i is the requested loan amount for $i \in I_A$.

In a very simple version of the mortgage model, we could define $U_i(\boldsymbol{x}) = a_i x_i$ for $i \in I_A$ and $U_i(\boldsymbol{x}) = \sum_i r_i p_i x_i$ for $i \in I_B$. Here a_i is a constant that roughly indicates the utility value of loan dollars to applicant i, and p_i is the probability that applicant i will repay the loan. Also r_i for $i \in B$ is the total rate of return to stakeholder i over the lifetime of the loans. The social welfare function W is nonlinear in general but can be converted to a linear W' in many cases if appropriate constraints are included. To take a simple example, a maximin criterion $W(\boldsymbol{u}) = \min_i\{u_i\}$ can be linearized by letting $W'(\boldsymbol{u}) = w$ and adding constraints $w \leq u_i$ for all i. Then the social welfare maximization problem is

$$\max_{\boldsymbol{x},\boldsymbol{u},w} \left\{ w \;\middle|\; \begin{array}{c} u_i \leq a_i x_i, \; i \in I_A \\ u_i \leq \sum_{j \in I_A} r_j p_j x_j, \; i \in I_B \\ w \leq u_i, \; 0 \leq x_i \leq h_i, \; \text{all } i \\ \sum_{i \in I_A} x_i \leq H \end{array} \right\}$$

We note that the constraints and objective function are linear and therefore define an easily solved linear programming problem.

The probabilities p_i of default for the set I_A of current mortgage applicants can be estimated by machine learning, and the optimization model can be applied to these applicants to arrive at fair loan decisions. An alternative approach is to solve the optimization problem for a pre-defined set I_A of hypothetical applicants, each associated with certain financial characteristics. The probability of repayment for each hypothetical applicant could again be estimated by machine learning, using the pre-defined characteristics as input to the neural network. Then each new applicant could be awarded or denied a loan based on the solution of the optimization problem and the hypothetical applicant i to which the new applicant is most similar.

If the optimal solution partially funds an applicant i (i.e., $0 < x_i < h_i$), the bank could make a judgment call as to whether to grant the loan, or else solve a variant of the optimization problem that requires $x_i \in \{0, h_i\}$. The latter is accomplished by introducing 0–1 variables δ_i and solving the problem

$$\max_{\boldsymbol{x},\boldsymbol{u},w,\boldsymbol{\delta}} \left\{ w \;\middle|\; \begin{array}{c} u_i \leq \beta_i x_i, \; i \in I_A \\ u_i \leq \sum_{j \in I_B} r_j p_j x_j, \; i \in I_B \\ x_i = h_i \delta_i, \; \delta_i \in \{0,1\}, \text{ all } i \\ w \leq u_i, \text{ all } i; \; \sum_{i \in I_A} x_i \leq H \end{array} \right\} \tag{3}$$

The 0–1 variables make the problem harder to solve, but it is a mixed integer/linear programming problem, for which solution technology is highly advanced. It is likely to be solved without difficulty, since it is not posed for all individuals in the training set. Rather, it is solved for applicants currently under consideration, or for hypothetical applicants as described above.

1.2 Linearity assumption for constraints

We will assume for the present discussion that the feasible set $S_{\boldsymbol{xu}}$ is defined by a linear system $A\boldsymbol{x} + B\boldsymbol{u} \leq \boldsymbol{b}$, aside from any integrality conditions on the variables. This simplifies the optimization problem while providing a great deal of modeling flexibility. For example, the assumption is satisfied when two conditions are met: (a) the constraints on feasible resource allocations $\boldsymbol{x}$ are linear, which is normally the case (as when there are one or more budget constraints), and (b) utilities u_i are a linear or concave function of resources (the latter indicating the typical situation of decreasing returns to scale) and can therefore be approximated by a concave piecewise linear function. These conditions are met by the mortgage problem and a wide variety of other decision problems.

We do *not* assume that the social welfare function W is linear, and it is in fact nonlinear in most interesting cases. Yet we can convert a nonlinear W to a linear W' in almost all of the optimization models described here. In fact, all of the models are of one of the following types:

- *Linear programming (LP).* This is an optimization problem with continuous variables, a linear objective function, and linear inequality and/or equality constraints. It is extremely well solved using the simplex method or an interior point method. Computation time is not an issue except for truly huge instances.

- *Mixed integer/linear programming (MILP).* This is an LP problem except that some variables are discrete (in our case, 0–1). It is a combinatorial problem but is often tractable for hundreds or even thousands of discrete variables using state-of-the-art software.

- *Convex nonlinear programming with linear constraints.* These is an LP problem except for a convex nonlinear (in one case, quadratic) objective function. Only two models have this form and can be efficiently solved by a quadratic programming, reduced gradient, or other specialized method.

2 Previous Work

We briefly survey two research streams on optimization and fairness, one from the optimization literature, and one from the AI literature. The former stream focuses on formulating optimization models to represent practical problems where fairness is an important concern, such as resource allocation, capacity planning, routing, scheduling and so forth. While some of these applications appear in AI research, we are concerned in this paper with understanding how optimization can be viewed more broadly as a general paradigm for formulating fairness by maximizing social welfare. A comprehensive survey of fairness in the optimization literature is provided by Karsu and Morton (2015).

Bandwidth allocation in telecommunication networks is a popular application studied in early works on fair resource allocation (Luss [1999], Ogryczak and Śliwiński [2002], Ogryczak et al. [2008]). For problems in this domain, a standard strategy is to define an objective function that is consistent with a Rawlsian criterion, and then to solve the corresponding model to obtain equitable allocations that optimize the worst performance among activities or services that compete for bandwidth. Iancu and Trichakis (2014) studied fairness in the context of portfolio optimization, and used an optimization model to determine a portfolio design that would attain desirable trade-offs between optimal trading performance and equitable cost-sharing among accounts. Project assignment is another application where fairness is often relevant, as the involved stakeholders may have different preferences over projects. For instance, Chiarandini et al. (2019) worked with a real-life decision to assign projects to university students. They formulated the allocation problem as a MILP model and compared the empirical performance of using SWFs that capture different fairness-efficiency balancing principles as objective functions. Fair optimization has also received attention in humanitarian operations. Eisenhandler and Tzur (2019) studied an important logistical challenge in food bank operations, food pickup and distribution. They designed a routing resource allocation model to seek both fair allocation of food to different agency locations and efficient delivery of as much food as possible. Mostajabdaveh et al. (2019) considered a disaster preparation task of selecting shelter locations and assigning neighborhoods to shelters. To make fair and efficient decisions while accounting for uncertainty, they used a stochastic programming model to optimize an objective function related to the Gini coefficient.

Recent AI research has developed efficient algorithms that take fairness into account. This effort differs from our proposal in that it develops algorithms to solve specific problems that have a fairness component, rather than formulating optimization models that can be submitted to state-of-the-art software. Algorithmic design tasks are often associated with fair matching decisions, such as kidney exchange (McElfresh and Dickerson [2018]), paper-reviewer assignment in peer review (Stelmakh et al. [2019]), or online decision procedures for a complex situation such as ridesharing (Nanda et al. [2020]).

Fair machine learning is a rapidly growing field in recent years. Fair ML methods in literature can be categorized as pre-, in-, or post-processing approaches, which respectively attain fairness by modifying standard ML methods before, during, or after the training phase. A main difference between our proposal and the majority of literature is how fairness is defined and measured. ML fairness requires the elimination of bias and discrimination and is measured in terms of predictions from ML models, while we adopt a utility- and welfare-based view of fairness. However, some recent research has discussed the usefulness of a social-welfare-maximization perspective in fair ML (Heidari et al. [2018], Hu and Chen [2020]). Perhaps the optimization models and techniques we present here can benefit future work in this area.

Since optimization is a core technique in ML regardless of whether fairness is taken into account, we focus here on research that uses optimization in the fairness-seeking component. Pre-processing methods aim to prepare training data to prevent bias and disparity, and optimization models can be used to find the best data modifications. For example, Zemel et al. (2013) and Calmon et al. (2017) proposed optimization models to learn fair representations of the original training data. Their models used objective functions that capture the trade-off among preserving prediction accuracy, limiting data distortion and eliminating potential discrimination associated with protected attributes.

Post-processing seeks fairness by adjusting the predictions generated from the trained model. Similar to the pre-processing case, we can determine the optimal tuning rules with optimization models. Examples of this strategy can be found in Hardt et al. (2016) and Alabdulmohsin (2020).

Fairness through optimization can fit naturally into in-processing methods. For standard ML algorithms that essentially solve optimization problems, such as support vector machines, logistic regression, etc., it is convenient to obtain fair alternatives by adding fairness constraints or including fairness components in objective function. A wide variety of fairness definitions have been studied in different ML frameworks. A series of papers have formulated constraints to denote well-known parity based fairness notions including demographic parity, equality of opportunity and predictive rate parity (Zafar et al. [2017, 2019], Olfat and Aswani [2018], Donini et al. [2018], Heidari et al. [2019]). A different direction explored in the literature is to define objective function to encode both fairness and the conventional training accuracy goals (Berk et al. [2017], Goel et al. [2018], Heidari et al. [2018]). These papers have demonstrated the empirical potentials of their optimization models for fair classification and regression. Another related work is that of Aghaei et al. (2019), which developed a mixed integer optimization based framework for learning fair decision trees, where a discrimination measure is used as penalty regularizer to eliminate disparity in the loss minimization model for training optimal decision trees.

3 Inequality Measures

One possible measure of fairness is the degree of equality in the distribution of utilities, for which several statistical metrics have been proposed (Cowell [2000], Jenkins and Van Kerm [2011]). Equality is not the same concept as fairness, but it is related and can be a useful criterion in some cases (Frankfurt [2015], Parfit [1997], Scanlon [2003]). We present optimization models for relative range, relative mean deviation, coefficient of variation, and the Gini coefficient. We begin with a brief review of linear-fractional programming, which is useful for converting these and some other models to easily solved LP problems.

3.1 Linear-fractional programming

Charnes and Cooper (1962) provides a mechanism for converting optimization problems whose objective function is a ratio of affine functions to linear programming (LP) problems, which are easy to solve. It applies to problems of the form

$$\max_{\boldsymbol{u}} \left\{ \frac{\boldsymbol{c}^\top \boldsymbol{u} + c_0}{\boldsymbol{d}^\top \boldsymbol{u} + d_0} \;\middle|\; A\boldsymbol{u} \leq \boldsymbol{b} \right\} \tag{4}$$

where the denominator $\boldsymbol{du} + d_0$ is positive in the feasible set $\{\boldsymbol{u} \mid A\boldsymbol{u} \leq \boldsymbol{b}\}$. We introduce a scalar variable t and use the change of variable $\boldsymbol{u} = \boldsymbol{u}'/t$ to write (4) as the LP problem

$$\max_{\boldsymbol{u}',t} \left\{ \boldsymbol{c}^\top \boldsymbol{u}' + c_0 t \;\middle|\; A\boldsymbol{u}' \leq \boldsymbol{b}t,\ \boldsymbol{d}^\top \boldsymbol{u}' + d_0 t = 1,\ t \geq 0 \right\} \tag{5}$$

Then if $(\hat{\boldsymbol{u}}', \hat{t})$ is an optimal solution of (5), $\boldsymbol{u} = \hat{\boldsymbol{u}}'/\hat{t}$ solves (4). This technique can be extended to nonlinear and MILP models, as we will see below.

3.2 Measures of relative dispersion

As a simple example, linear-fractional programming can be used when the measure of inequality is the *relative range* of utilities. The SWF is $W(\boldsymbol{u}) = -(u_{\max} - u_{\min})/\bar{u}$, where $u_{\max} = \max_i\{u_i\}$, $u_{\min} = \min_i\{u_i\}$, and $\bar{u} = (1/n)\sum_i u_i$. We assume with little loss of generality that $A\boldsymbol{x} + B\boldsymbol{u} \leq \boldsymbol{b}$ implies $\bar{\boldsymbol{u}} > 0$. The problem of maximizing $W(\boldsymbol{u})$ subject to linear constraints $A\boldsymbol{x} + B\boldsymbol{u} \leq \boldsymbol{b}$ can then be written as the LP problem

$$\min_{\substack{\boldsymbol{x}',\boldsymbol{u}',t \\ u'_{\min},u'_{\max}}} \left\{ u'_{\max} - u'_{\min} \;\middle|\; \begin{array}{c} u'_{\min} \leq u'_i \leq u'_{\max},\ \text{all } i \\ A\boldsymbol{x}' + B\boldsymbol{u}' \leq \boldsymbol{b}t,\ \bar{u}' = 1,\ t \geq 0 \end{array} \right\}$$

where $u'_{\min}, u'_{\max}$ are regarded as variables along with $\boldsymbol{x}'$, $\boldsymbol{u}'$, and t. If $(\hat{\boldsymbol{x}}', \hat{\boldsymbol{u}}', \hat{u}'_{\min}, \hat{u}'_{\max}, \hat{t})$ solves this problem, then $\boldsymbol{u} = \hat{\boldsymbol{u}}'/\hat{t}$ is a distribution that minimizes the relative range.

Another dispersion metric is the *relative mean deviation*, for which the SWF is $W(\boldsymbol{u}) = -(1/\bar{u})\sum_i |u_i - \bar{u}|$. This, too, can be optimized by linear-fractional programming:

$$\min_{\boldsymbol{x}',\boldsymbol{u}',\boldsymbol{v},t} \left\{ \sum_i v_i \;\middle|\; \begin{array}{c} -v_i \leq u'_i - \bar{u}' \leq v_i,\ \text{all } i \\ A\boldsymbol{x}' + B\boldsymbol{u}' \leq \boldsymbol{b}t,\ \bar{u}' = 1,\ t \geq 0 \end{array} \right\}$$

where $v_1, \ldots, v_n$ are new variables.

The *coefficient of variation* is the standard deviation with normalized mean. The SWF is

$$W(\boldsymbol{u}) = -\frac{1}{\bar{u}} \left[\frac{1}{n} \sum_i (u_i - \bar{u})^2 \right]^{\frac{1}{2}}$$

Although the numerator is nonlinear, we can use the same change of variable to formulate the optimization problem as

$$\min_{\boldsymbol{x}',\boldsymbol{u}',\boldsymbol{v},t} \left\{ \Big[\frac{1}{n}\sum_i (u'_i - \bar{u}')^2\Big]^{\frac{1}{2}} \;\middle|\; \begin{matrix} A\boldsymbol{x}' + b\boldsymbol{u}' \le \boldsymbol{b}t \\ \bar{u}' = 1,\ t \ge 0 \end{matrix} \right\}$$

This is not an LP problem, but we can obtain the same optimal solution by solving it without the exponent $\frac{1}{2}$. This yields a convex quadratic programming problem with linear constraints, for which there are efficient algorithms in state-of-the-art optimization packages.

3.3 Gini coefficient

The *Gini coefficient* is by far the best known measure of inequality, as it is routinely used to measure income and wealth inequality. It is proportional to the area between the Lorenz curve and a diagonal line representing perfect equality and therefore vanishes under perfect equality. The SWF is $W(\boldsymbol{u}) = -(1/2\bar{u}n^2)\sum_{i,j}|u_i - u_j|$. Again applying linear-fractional programming, the problem of minimizing the Gini coefficient subject to linear constraints is equivalent to the LP problem

$$\min_{\boldsymbol{x}',\boldsymbol{u}',V,t} \left\{ \frac{1}{2n^2}\sum_{i,j} v_{ij} \;\middle|\; \begin{matrix} -v_{ij} \le u'_i - u'_j \le v_{ij},\ \text{all } i,j \\ A\boldsymbol{x}' + B\boldsymbol{u}' \le \boldsymbol{b}t,\ \bar{u}' = 1,\ t \ge 0 \end{matrix} \right\}$$

where v_{ij} is a new variable for all i, j.

4 Fairness for the Disadvantaged

Rather than focus solely on inequality, fairness measures can enhance equality while giving preference to those who are less advantaged. Far and away the most famous of such measures is the difference principle of John Rawls (1999), a maximin criterion that is based on careful philosophical argument and debated in a vast literature (Freeman [2003], Richardson and Weithman [1999]). The difference principle can be plausibly extended to a lexicographic maximum principle. There are also the Hoover and McLoone indices, which are statistical measures that emphasize the lot of the less advantaged.

4.1 Rawlsian criteria

The Rawlsian *difference principle* states that inequality should exist only to the extent that it is necessary to improve the lot of the worst-off. It is defended with a social contract argument that, in its simplest form, maintains that the structure of society must be negotiated in an "original position" in which people do not yet know their station in society. But one can rationally assent to the possibility of ending up on the bottom only if that person would have been even worse off in any other social structure, whence an imperative to maximize the lot of the worst-off. The principle is intended to apply only to the design of social institutions, and only to the distribution of "primary goods," which are goods that any rational person would want. Yet it can be adopted as a general criterion for distributing utility, namely a *maximin* criterion that maximizes the simple SWF $W(\boldsymbol{u}) = \min_i\{u_i\}$. This is readily formulated as the LP problem

$$\max_{\boldsymbol{x},\boldsymbol{u},w} \left\{ w \;\middle|\; w \le u_i,\ \text{all } i;\ A\boldsymbol{x} + B\boldsymbol{u} \le \boldsymbol{b} \right\}$$

The maximin criterion can be plausibly extended to *lexicographic maximization* (leximax) by first maximizing the smallest utility, then holding this utility fixed while maximizing the smallest among those that remain, and so forth. This is known as *pre-emptive goal programming* in the optimization literature and is achieved by solving a sequence of optimization problems

$$\max_{\boldsymbol{x},\boldsymbol{u},w} \left\{ w \;\middle|\; \begin{matrix} w \le u_i,\ u_i \ge \hat{u}_{i_{k-1}},\ i \in I_k \\ A\boldsymbol{x} + B\boldsymbol{u} \le \boldsymbol{b} \end{matrix} \right\} \tag{6}$$

for $k = 1, \ldots, n$, where $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{u}})$ is an optimal solution of problem k, $\hat{u}_{i_0} = -\infty$, and

$$\hat{u}_{i_k} = \min_{i \in I_k}\{\hat{u}_i\}, \text{ with } I_k = \{1, \ldots, n\} \setminus \{i_1, \ldots, i_{k-1}\}$$

Ogryczak and Śliwiński (2006) showed how to obtain a leximax solution with a single optimization model, but it is impractical for most purposes due to the very large coefficients required in the objective function.

4.2 Hoover and McLoone Indices

The *Hoover index* is related to the Gini coefficient, as it is proportional to the maximum vertical distance between the Lorenz curve and a diagonal line representing perfect equality. It is also proportional to the relative mean deviation. It can be interpreted as the fraction of total utility that would have to be transferred from the richer half of the population to the poorer half to achieve perfect equality. The SWF is $W(\boldsymbol{u}) = -(1/2n\bar{u})\sum_i |u_i - \bar{u}|$. The Hoover index can be minimized by solving the same LP problem as for the relative mean deviation.

The *McLoone index* compares the total utility of individuals at or below the median utility to the utility they would enjoy if all were brought up to the median utility. The index is 1 if nobody's utility is strictly below the median, and it approaches 0 if nearly everyone below the median has utility much smaller than the median (on the assumption that all utilities are positive.) The McLoone index benefits the disadvantaged by rewarding equality in the lower half of the distribution, but it is unconcerned by the existence of very rich individuals in the upper half. The SWF is

$$W(\boldsymbol{u}) = \frac{1}{|I(\boldsymbol{u})|u^{\text{med}}} \sum_{i \in I(\boldsymbol{u})} u_i$$

where u^{med} is the median of utilities in $\boldsymbol{u}$ and $I(\boldsymbol{u})$ is the set of indices of utilities at or below the median, so that $I(\boldsymbol{u}) = \{i \mid u_i \leq u^{\text{med}}\}$.

We can formulate the maximization problem as an MILP problem, but with a fractional objective function, by using standard "big-M" modeling techniques from integer programming. The model uses 0–1 variables δ_i, where $\delta_i = 1$ when $i \in I(\boldsymbol{u})$. The constant M is a large number chosen so that $u_i < M$ for all i. The model is

$$\max_{\substack{\boldsymbol{x},\boldsymbol{u},m \\ \boldsymbol{y},\boldsymbol{z},\boldsymbol{\delta}}} \left\{ \frac{\sum_i y_i}{\sum_i z_i} \;\middle|\; \begin{array}{c} m - M\delta_i \leq u_i \leq m + M(1-\delta_i), \text{ all } i \\ y_i \leq u_i, \; y_i \leq M\delta_i, \; \delta_i \in \{0,1\}, \text{ all } i \\ z_i \geq 0, \; z_i \geq m - M(1-\delta_i), \text{ all } i \\ A\boldsymbol{x} + B\boldsymbol{u} \leq \boldsymbol{b}, \; \sum_i \delta_i \leq n/2 \end{array} \right\}$$

where the new variable m represents the median, variable y_i is u_i if $\delta_i = 1$ and 0 otherwise, and variable z_i is m if $\delta_i = 1$ and 0 otherwise in the optimal solution. The fractional objective function can be removed, resulting in an MILP problem, by using the same change of variable as in linear-fractional programming:

$$\max_{\substack{\boldsymbol{x}',\boldsymbol{u}',m' \\ \boldsymbol{y}',\boldsymbol{z}',t,\boldsymbol{\delta}}} \left\{ \sum_i y'_i \;\middle|\; \begin{array}{c} u'_i \geq m' - M\delta_i, \text{all } i \\ u'_i \leq m' + M(1-\delta_i), \text{ all } i \\ y'_i \leq u'_i, \; y'_i \leq M\delta_i, \; \delta_i \in \{0,1\}, \text{ all } i \\ z'_i \geq 0, \; z'_i \geq m' - M(1-\delta_i), \text{ all } i \\ A\boldsymbol{x}' + B\boldsymbol{u}' \leq \boldsymbol{b}t, \; \sum_i z'_i = 1, \; t \geq 0 \\ \sum_i \delta_i \leq n/2 \end{array} \right\}$$

5 Combining Fairness and Efficiency

In many practical applications involving fairness, efficiency is desired as well. A standard efficiency measure is the *utilitarian* SWF, $W(\boldsymbol{u}) = \sum_{i=1}^{n} u_i$, which is indifferent to the inequalities among individual utilities.

One obvious strategy for combining equity and efficiency is to define a SWF that is a convex combination of utility and a fairness criterion, such as one of those described in previous sections. Although this strategy is convenient, it poses the difficult challenge of selecting and interpreting the weight parameters of the convex combination. A popular alternative is alpha fairness, of which proportional fairness (the Nash bargaining solution) is a special case. The Kalai-Smorodinsky bargaining solution is another option. There are also schemes that combine Rawlsian and utilitarian criteria based on justice principles proposed in Williams and Cookson (2000).

5.1 Alpha fairness and proportional fairness

Alpha fairness regulates the relative importance of equity and efficiency with a parameter α in the SWF

$$W_\alpha(\boldsymbol{u}) = \begin{cases} \dfrac{1}{1-\alpha}\displaystyle\sum_i u_i^{1-\alpha} & \text{for } \alpha \geq 0,\ \alpha \neq 1 \\ \displaystyle\sum_i \log(u_i) & \text{for } \alpha = 1 \end{cases}$$

The SWF is purely utilitarian when $\alpha = 0$ and becomes purely maximin as $\alpha \to \infty$. If one person's utility u_i is less than another's utility u_j, then u_j must be reduced by $(u_j/u_i)^\alpha$ units to compensate for a unit increase in u_i while maintaining constant social welfare. Thus larger values of α imply greater sacrifice from the person j who is better off and can therefore be interpreted as giving more emphasis to fairness. Lan et al. (2010) give an axiomatic justification of alpha fairness in the context of network resource allocation.

The problem of maximizing $W_\alpha(\boldsymbol{u})$ can be solved directly in the form

$$\max_{\boldsymbol{x},\boldsymbol{u}} \left\{ W_\alpha(\boldsymbol{u}) \;\middle|\; A\boldsymbol{x} + B\boldsymbol{u} \leq \boldsymbol{b} \right\}$$

without reformulation. The objective function is nonlinear, but since it is concave for all $\alpha \geq 0$, any local optimum is a global optimum. The problem can therefore be solved to optimality by such efficient algorithms as the reduced gradient method, which is a straightforward generalization of the simplex method for LP, particularly since $W_\alpha(\boldsymbol{u})$ has a simple closed-form gradient. Maximizing alpha fairness can therefore be regarded as tractable for reasonably large instances.

A well-known special case of α-fairness is *proportional fairness*, which corresponds to setting $\alpha = 1$. Maximizing proportional fairness is equivalent to solving the Nash bargaining problem (Nash [1950]). Nash gave an axiomatic argument for the model, and it has also been justified as the result of certain rational bargaining procedures (Harsanyi [1977], Rubinstein [1982], Binmore et al. [1986]). Proportional fairness is widely used in engineering to maximize throughput while maintaining some degree of fairness, as for example in telecommunication networks and traffic signal timing.

5.2 Kalai-Smorodinsky bargaining solution

The *Kalai-Smorodinsky* bargaining solution, proposed as an alternative to the Nash bargaining solution, minimizes each person's relative concession (Kalai and Smorodinsky [1975]). That is, it provides the largest possible utility relative to the maximum one could obtain if other players are disregarded, subject to the condition that all persons obtain the same fraction β of their maximum. It has been defended by Thompson (1994) and is consistent with the "contractarian" ethical philosophy of Gautier (1983). The SWF is

$$W(\boldsymbol{u}) = \begin{cases} \sum_i u_i, & \text{if } \boldsymbol{u} = \beta \boldsymbol{u}^{\max} \\ & \quad \text{for some } \beta \text{ with } 0 \leq \beta \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

where $u_i^{\max} = \max_{\boldsymbol{x},\boldsymbol{u}}\{u_i \mid A\boldsymbol{x} + B\boldsymbol{u} \leq \boldsymbol{b}\}$ for each i. The optimization problem is a straightforward LP:

$$\max_{\beta,\boldsymbol{x},\boldsymbol{u}} \left\{ \beta \;\middle|\; \boldsymbol{u} = \beta \boldsymbol{u}^{\max},\ A\boldsymbol{x} + B\boldsymbol{u} \leq \boldsymbol{b},\ \beta \leq 1 \right\}$$

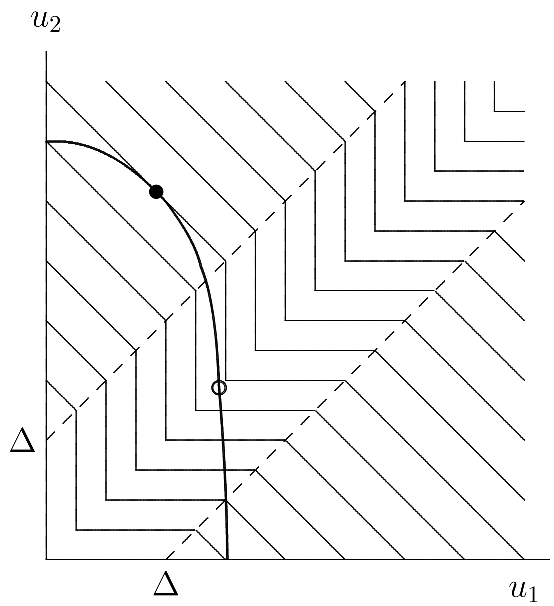

Figure 1: Contours for the equity-based Williams-Cookson SWF.

5.3 Combining maximin and utilitarian criteria

We now show how to formulate optimization models that directly combine Rawlsian and utilitarian criteria. These formulations are useful when either criterion in isolation would be too extreme for policy making. Williams and Cookson (2000) suggest two ideas for combining maximin and utilitarian objectives in the case of two persons, and these can be generalized to n persons and formulated as MILP models. They correspond to opposite approaches to combining equity and efficiency: an equity-based approach that begins by maximizing equity but switches to a utilitarian criterion to avoid extreme solutions, and a utility-based approach that does the opposite. We also show how to generalize the equity-based approach so as to combine leximax and and utilitarian objectives in a sequence of MILP models (Chen and Hooker [2020b]).

5.3.1 Equity-based Williams-Cookson SWF.

The 2-person equity-based model of Williams and Cookson (2000) pursues fairness until the efficiency cost becomes too high, whereupon it switches to a utilitarian objective. It uses a maximin criterion when the two utilities are sufficiently close to each other, specifically $|u_1 - u_2| < \Delta$, and otherwise it uses a utilitarian criterion. This is illustrated in Fig. 1, where the feasible set is the area under the curve. The maximin solution (open circle) requires a substantial sacrifice from person 2. As a result, the utilitarian solution (black dot) earns slightly more social welfare and is the preferred choice. The SWF can be written

$$W_1(u_1, u_2) = \begin{cases} u_1 + u_2, & \text{if } |u_1 - u_2| \geq \Delta \\ 2\min\{u_1, u_2\} + \Delta, & \text{otherwise} \end{cases}$$

The maximin criterion is modified from the standard formula $\min\{u_1, u_2\}$ to ensure continuity of the SWF as one shifts between the utilitarian and the maximin objective.

Hooker and Williams (2012) generalize W_1 to n persons. The utility u_i of person i belongs to the *fair region* if $u_i - u_{min} \leq \Delta$ and otherwise to the *utilitarian region*, where $u_{\min} = \min_i\{u_i\}$. A person whose utility is in the fair region is considered sufficiently disadvantaged to deserve priority. The generalized SWF $W_1(\boldsymbol{u})$ counts all utilities in the fair region as equal to $u_{\min}$, so that they are treated in solidarity with the worst-off, and all other utilities as themselves. Similar to the 2-person case, copies of Δ are added to the maximin criterion to ensure continuity of W_1.

$$W_1(\boldsymbol{u}) = (n-1)\Delta + \sum_{i=1}^{n} \max\{u_i - \Delta, u_{min}\} \tag{7}$$

The parameter Δ regulates the equity/efficiency trade-off, with $\Delta = 0$ corresponding to a purely utilitarian objective and $\Delta = \infty$ to a purely maximin objective.

In addition, Hooker and Williams (2012) extended $W_1(\boldsymbol{u})$ to represent the social welfare associated with the utility distribution to groups of recipients. Suppose there are n groups of possibly different sizes, and let s_i and u_i respectively denote the number of individuals in group i and the utility of each individual in the group. The function $W_1^g(\boldsymbol{u})$ considers a group i to be in the fair region when its per capita u_i is within Δ of u_{min}, and it prioritizes only the groups in the fair region.

$$W_1^g(\boldsymbol{u}) = \Big(\sum_{i=1}^{n} s_i - 1\Big)\Delta + \sum_{i=1}^{n} s_i \max\{u_i - \Delta, u_{min}\} \tag{8}$$

Hooker and Williams (2012) provided tractable MILP models to maximize $W_1(\boldsymbol{u})$ and $W_1^g(\boldsymbol{u})$ subject to auxiliary constraints $u_i - u_j \le M$ required for MILP representability. The model for maximizing W_1 is

$$\max_{\boldsymbol{x},\boldsymbol{u},\boldsymbol{\delta},\boldsymbol{v},w,z} \left\{ z \;\middle|\; \begin{array}{c} z \le (n-1)\Delta + \sum_i v_i \\ u_i - \Delta \le v_i \le u_i - \Delta\delta_i, \text{ all } i \\ w \le v_i \le w + (M-\Delta)\delta_i, \text{ all } i \\ u_i - u_i \le M, \text{ all } i, j \\ A\boldsymbol{x} + B\boldsymbol{u} \le \boldsymbol{b} \\ u_i \ge 0,\ \delta_i \in \{0,1\}, \text{all } i \end{array} \right\} \tag{9}$$

and the model for W_1^g is similar. The practicality of these models was verified with experiments on a healthcare resource allocation instance of realistic size.

5.3.2 Utility-based Williams-Cookson SWF.

Alternatively, when efficiency is the initial objective, fairness is not considered until inequality in the utility distribution becomes intolerable. For the 2-person case, Williams and Cookson (2000) define the SWF to be utilitarian when $|u_1 - u_2| < \Delta$, which corresponds to the case where enforcing fairness is unnecessary. The SWF is maximin (again with a modification for continuity) otherwise. In Fig. 2, the utilitarian solution (open dot) is unfair to person 1, and the welfare-maximizing solution is more egalitarian (black dot). The SWF is

$$\overline{W}_1(u_1, u_2) = \begin{cases} 2\min\{u_1, u_2\} + \Delta, & \text{if } |u_1 - u_2| \ge \Delta \\ u_1 + u_2, & \text{otherwise} \end{cases}$$

We generalize this view to define SWFs to capture the combined objective for n persons or groups with techniques similar to those used by Hooker and Williams. The main difference is that we now say a utility u_i belongs to the fair region if $u_i - u_{min} \ge \Delta$, otherwise it is in the utilitarian region. In the SWFs $\overline{W}_1$ and $\overline{W}_1^g$, we still count the fair region utilities as equivalent to u_{min}, the utilitarian region utilities as their exact values, and add the needed multiples of Δ to obtain continuous SWFs.

$$\overline{W}_1(\boldsymbol{u}) = (n-1)\Delta + \sum_{i=1}^{n} \min\{u_i - \Delta, u_{min}\}$$

$$\overline{W}_1^g(\boldsymbol{u}) = \Big(\sum_{i=1}^{n} s_i - 1\Big)\Delta + \sum_{i=1}^{n} s_i \min\{u_i - \Delta, u_{min}\}$$

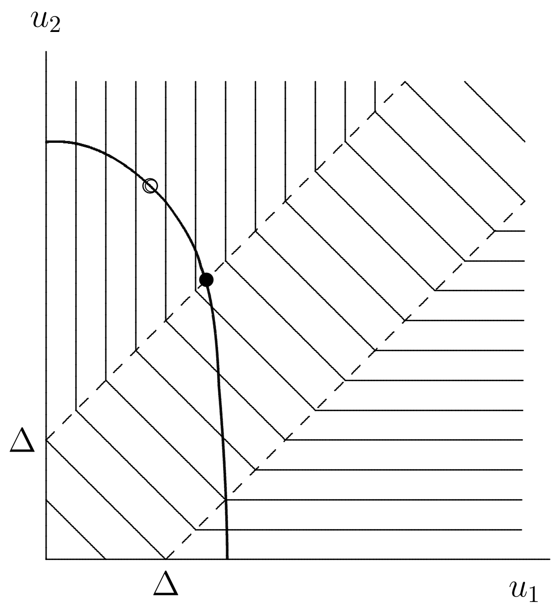

Figure 2: Contours for the utility-based Williams-Cookson SWF.

Because $\overline{W}_1(\boldsymbol{u})$ and $\overline{W}_2^g(\boldsymbol{u})$ are continuous and concave functions, the corresponding maximization problems have simple LP formulations that are convenient to solve. We first derive the LP formulation for maximizing $\overline{W}_1(\boldsymbol{u})$. The maximization problem is

$$\max_{\boldsymbol{x},\boldsymbol{u},\boldsymbol{v}} \left\{ (n-1)\Delta + \sum_{i=1}^{n} v_i \;\middle|\; \begin{array}{c} v_i = \min\{u_i - \Delta, u_{min}\}, \\ \text{all } i \\ A\boldsymbol{x} + B\boldsymbol{u} \leq \boldsymbol{b} \end{array} \right\}$$

It is equivalent to the following LP because the constraints on v_i hold if and only if $v_i \leq u_i - \Delta$ and $v_i \leq w$, where $w = u_{min}$. We require $w \leq u_i$ for all i to ensure that w is set to u_{min} in the optimal solution.

$$\max_{\boldsymbol{x},\boldsymbol{u},\boldsymbol{v},w,z} \left\{ z \;\middle|\; \begin{array}{c} z \leq (n-1)\Delta + \sum_{i=1}^{n} v_i \\ v_i \leq w \leq u_i, \text{ all } i \\ v_i \leq u_i - \Delta, \text{ all } i \\ w \geq 0, v_i \geq 0, \text{ all } i \\ A\boldsymbol{x} + B\boldsymbol{u} \leq \boldsymbol{b} \end{array} \right\}$$

The formulation for maximizing $\overline{W}_1^g$ is similar.

$$\max_{\boldsymbol{x},\boldsymbol{u},\boldsymbol{v},w,z} \left\{ z \;\middle|\; \begin{array}{c} z \leq (\sum_{i=1}^{n} s_i - 1)\Delta + \sum_{i=1}^{n} s_i v_i \\ v_i \leq w \leq u_i, \text{ all } i \\ v_i \leq u_i - \Delta, \text{ all } i \\ w \geq 0, v_i \geq 0, \text{ all } i \\ A\boldsymbol{x} + B\boldsymbol{u} \leq \boldsymbol{b} \end{array} \right\}$$

5.4 Combining leximax and utilitarian criteria

A leximax criterion offers broader sensitivity to equity than a maximin criterion, which is concerned only with the worst-off. Chen and Hooker (2020b,2020a) offer a series of SWFs that can be maximized sequentially to combine leximax and utilitarian criteria in a principled way:

$$\begin{aligned} W_k(\boldsymbol{u}) = & \textstyle\sum_{i=1}^{k-1}(n-i+1)u_{\langle i\rangle} \\ & +(n-k+1)\min\left\{u_{\langle 1\rangle}+\Delta, u_{\langle k\rangle}\right\} \\ & +\textstyle\sum_{i=k}^{n}\left(u_{\langle i\rangle}-u_{\langle 1\rangle}-\Delta\right)^+,\ k=2,\ldots,n \end{aligned}$$

where $\gamma^+ = \max\{0,\gamma\}$, and where $u_{\langle 1\rangle},\ldots,u_{\langle n\rangle}$ are $u_1,\ldots,u_n$ in nondecreasing order. The initial function W_1 is given by (7). The parameter Δ again regulates the efficiency/equity trade-off by giving preference to individuals whose utility is within Δ of the lowest, with greater weight to the more disadvantaged.

The MILP model for maximizing W_1 is (9). Using notation similar to that for the goal programming model (6), the MILP formulation for maximizing W_k, $k \geq 2$, is

$$\max_{\substack{\boldsymbol{x},\boldsymbol{u},\boldsymbol{\delta},\boldsymbol{\epsilon}\\ \boldsymbol{v},w,\tau,z}} \left\{ z \;\middle|\; \begin{array}{c} z \leq (n-k+1)\tau + \sum_{i\in I_k} v_i \\ 0 \leq v_i \leq M\delta_i,\ i\in I_k \\ v_i \leq u_i - \hat{u}_{i_1} - \Delta + M(1-\delta_i),\ i \in I_k \\ \tau \leq \hat{u}_{i_1}+\Delta,\ \tau\leq w,\ w \geq \hat{u}_{i_1} \\ w \leq u_i \leq w + M(1-\epsilon_i),\ i\in I_k \\ u_i - \hat{u}_{i_1} \leq M,\ i \in I_k \\ A\boldsymbol{x} + B\boldsymbol{u} \leq \boldsymbol{b} \\ \sum_{i\in I_k}\epsilon_i = 1;\ \delta_i,\epsilon_i \in\{0,1\}, i\in I_k \end{array} \right\}$$

The model is demonstrated in Chen and Hooker (2020b) and found to solve rapidly on healthcare resource and earthquake shelter location problems.

6 Fairness in Support Vector Machines

Optimization-based fairness is readily implemented in support vector machines (SVMs), because the problem of finding a separating hyperplane is already a constrained optimization problem. In principle, one need only replace the usual objective of maximizing the margin with that of maximizing a social welfare function.

However, fairness metrics can be incorporated into the SVM problem only if 0–1 variables are introduced to indicate how individuals are classified. Since the optimization problem is solved over all observations in the training set, this can result in a very large number of discrete variables and associated constraints. If one relies on an off-the-shelf MILP solver, this could limit the size of the training set to a few hundred observations.

Scaling up therefore remains a research issue, but possible strategies suggest themselves. An obvious one is to solve the problem using a subset of representative observations, perhaps selected with clustering techniques. Another is to draw on specialized techniques in the MILP literature for solving problems with many 0–1 variables. *Branch-and-price* methods, for example, routinely solve practical problems with millions of 0–1 variables by using column generation techniques to introduce variables only as they are needed to improve the solution. A related technique, "shrinking," is already used in sequential minimal optimization algorithms for SVMs.

We begin with showing how the classical SVM problem can be reformulated into an LP problem by normalizing in a slightly different way. This clears the way for fairness problems containing 0–1 variables to be formulated as MILP problems. We then show how statistical fairness metrics can be incorporated into the SVM problem. This is followed by a more general MILP model that allows the use of any SWF that is linearized in previous sections of this paper.

6.1 Linearizing the maximum margin problem

Adopting notation from the SVM literature, let each observation i in the training set consists of a real-valued vector $\boldsymbol{x}^i$ of features and an indication y_i as to whether the individual should receive a true classification

($y_i = 1$ for true, $y_i = -1$ for false). A *separating hyperplane* $\{\boldsymbol{x} \mid \boldsymbol{\theta}^\intercal \boldsymbol{x} + b = 0\}$ is one such that $\boldsymbol{\theta}^\intercal \boldsymbol{x}^i + b \geq 0$ for all and only true observations i. Note that $\boldsymbol{x}^i$ and y_i are problem data, not variables. The SVM problem is to find a separating hyperplane that maximizes the *margin* σ, which is the minimum distance from the hyperplane to point $\boldsymbol{x}^i$ over all i. The standard SVM model interprets distance as the L_2-norm:

$$\max_{\boldsymbol{\theta},b,\sigma} \left\{ \sigma \;\middle|\; \tfrac{1}{2}\sigma \leq y_i(\boldsymbol{\theta}^\intercal \boldsymbol{x}^i + b)/\|\boldsymbol{\theta}\|_2, \text{ all } i \right\} \tag{10}$$

This is transformed to a problem with a convex nonlinear objective function and linear constraints:

$$\min_{\boldsymbol{\theta},b} \left\{ \tfrac{1}{2}\|\boldsymbol{\theta}\|_2^2 \;\middle|\; y_i(\boldsymbol{\theta}^\intercal \boldsymbol{x}^i + b) \geq 1, \text{ all } i \right\}$$

The "soft margin" version of the problem adds to the objective function the sum of the errors ξ_i that result when observation i falls on the wrong side of the hyperplane.

$$\min_{\boldsymbol{\theta},b,\boldsymbol{\xi}} \left\{ \tfrac{1}{2}\|\boldsymbol{\theta}\|_2^2 + C\sum_i \xi_i \;\middle|\; y_i(\boldsymbol{\theta}^\intercal \boldsymbol{x}^i + b) \geq 1 - \xi_i, \text{ all } i \right\}$$

This problem is solved by solving its Wolfe dual, which is much smaller because it no longer contains a constraint for each observation. The Wolfe dual is an equivalent problem because it satisfies Slater's condition, and there is no duality gap as a result.

To obtain an LP model, the SVM problem can be normalized with the L_1-norm instead of the L_2-norm, in which case problem (10) becomes

$$\max_{\boldsymbol{\theta},b,\sigma} \left\{ \sigma \;\middle|\; \tfrac{1}{2}\sigma \leq y_i(\boldsymbol{\theta}^\intercal \boldsymbol{x}^i + b)/\|\boldsymbol{\theta}\|_1, \text{ all } i \right\}$$

Early literature has studied this L_1-norm based SVM and discussed its potential advantages over the conventional L_2-norm based formulation (e.g. Bradley and Mangasarian [1998], Zhu et al. [2003]). Using a similar transformation as before, we have the soft margin problem

$$\min_{\boldsymbol{\theta},b,\boldsymbol{\xi}} \left\{ \|\boldsymbol{\theta}\|_1 + C\sum_i \xi_i \;\middle|\; y_i(\boldsymbol{\theta}^\intercal \boldsymbol{x}^i + b) \geq 1 - \xi_i, \text{ all } i \right\}$$

Since $\|\boldsymbol{\theta}\|_1 = \sum_i |\theta_i|$, we can now linearize the model to obtain an LP problem.

$$\min_{\boldsymbol{\theta},b,\boldsymbol{\xi},\boldsymbol{t}} \left\{ \sum_i (t_i + C\xi_i) \;\middle|\; \begin{array}{l} -t_i \leq \theta_j \leq t_j, \; t_j \geq 0, \text{ all } j \\ y_i(\boldsymbol{\theta}^\intercal \boldsymbol{x}^i + b) \geq 1 - \xi_i, \text{ all } i \end{array} \right\} \tag{11}$$

6.2 Modeling classifications

Social welfare and bias measures can be formulated only if we introduce 0–1 variables δ_i to indicate how an individual i is classified. The resulting problem can no longer be solved using the classical strategy of solving the Wolfe dual, because there is a duality gap when integer variables are present. We therefore solve the original (primal) model directly by formulating it as an MILP problem.

We wish to set $\delta_i = 1$ when individual i falls on the true side of the hyperplane; that is when $\boldsymbol{\theta}^\intercal \boldsymbol{x}^i + b \geq 0$. Due to the large number of variables δ_i, it is important to formulate this condition in a way that allows efficient solution. We therefore write a *sharp* MILP formulation of a disjunctive model for each δ_i (a sharp formulation is one whose continuous relaxation is the convex hull of the feasible set). The disjunctive model is

$$\begin{pmatrix} \boldsymbol{\theta}^\intercal \boldsymbol{x}^i + b \leq 0 \\ \delta_i = 0 \end{pmatrix} \vee \begin{pmatrix} \boldsymbol{\theta}^\intercal \boldsymbol{x}^i + b \geq 0 \\ \delta_i = 1 \end{pmatrix} \tag{12}$$

This model has an MILP representation if and only if the polyhedra described by the two disjuncts have the same recession cone. To ensure this, we impose in each disjunct the constraint $-M \leq \boldsymbol{\theta}^\intercal \boldsymbol{x}^i + b \leq M$, which is valid for sufficiently large M. The disjunction (12) now has a sharp MILP representation that simplifies to

$$-M(1-\delta_i) \leq \boldsymbol{\theta}^\intercal \boldsymbol{x}^i + b \leq M\delta_i, \;\; \delta_i \in \{0,1\}$$

This constraint can be added to the SVM model for each i to define δ_i and enable the objective function to reflect bias or fairness.

6.3 Incorporating fairness

We first indicate how statistical bias measures can be incorporated into the SVM problem. These metrics typically compare statistics for a protected group with those outside the protected group. For each observation i, let $z_i = 1$ when individual i belongs to the protected group. The simplest bias metric is *demographic parity*, which is based on the difference between the probability of true classification for and protected and unprotected individuals. This yields the bias metric $|\Delta(\boldsymbol{\delta})|$, where

$$\Delta(\boldsymbol{\delta}) = \frac{\sum_i z_i \delta_i}{\sum_i z_i} - \frac{\sum_i (1-z_i)\delta_i}{\sum_i (1-z_i)}$$

This metric can be incorporated into the SVM problem by minimizing a convex combination of the margin-plus-error and bias, $\lambda(\sigma + C\sum_i \xi_i) + (1-\lambda)|\Delta(\boldsymbol{\delta})|$, where $0 \leq \lambda \leq 1$. Since $\Delta(\boldsymbol{\delta})$ is a linear function of $\boldsymbol{\delta}$, the problem is easily linearized by extending the LP model (11) to obtain the following MILP model:

$$\min_{\substack{\boldsymbol{\theta},b,\boldsymbol{\xi}\\ \boldsymbol{t},\boldsymbol{\delta},w}} \left\{ \begin{array}{c|c} \begin{array}{c} \lambda \sum_i (t_i + C\xi_i) \\ +(1-\lambda)w \end{array} & \begin{array}{c} -t_i \leq \theta_j \leq t_j, \; t_j \geq 0, \text{ all } j \\ y_i(\boldsymbol{\theta}^\intercal \boldsymbol{x}^i + b) \geq 1 - \xi_i, \text{ all } i \\ -w \leq \Delta(\boldsymbol{\delta}) \leq w \\ -M(1-\delta_i) \leq \boldsymbol{\theta}^\intercal \boldsymbol{x}^i + b, \text{ all } i \\ \boldsymbol{\theta}^\intercal \boldsymbol{x}^i + b \leq M\delta_i, \text{ all } i \\ \delta_i \in \{0,1\}, \text{ all } i \end{array} \end{array} \right\}$$

A similar MILP model can be used for *equalized odds*, since the corresponding function $\Delta(\boldsymbol{\delta})$ is likewise a linear function of $\boldsymbol{\delta}$. *Predictive rate parity* requires a SWF that contains ratios of affine functions, but it can be formulated as an MILP problem using the same change of variable as in linear-fractional programming.

We now formulate a more general model that maximizes social welfare. For illustrative purposes, we maximize a convex combination of (negated) margin-plus-error and a social welfare function $W(\boldsymbol{u})$, so as to allow the accuracy of prediction to contribute to welfare. The function $W(\boldsymbol{u})$ could reflect fairness or a combination of fairness and utility, as discussed in previous sections. This general model requires that we define the utility u_i enjoyed by each individual i, which will depend on whether the individual receives the true classification. If we let $u_i = c_0 + c_i\delta_i$, then we can set $c_i > 0$ for an individual with $y_i = 1$ to indicate that the individual is better off being classified true, while we set $c_i < 0$ when $y_i = -1$ to indicate that being classified false is better. We now obtain the problem

$$\max_{\substack{\boldsymbol{\theta},b,\boldsymbol{\xi}\\ \boldsymbol{t},\boldsymbol{\delta},w}} \left\{ \begin{array}{c|c} \begin{array}{c} -\lambda \sum_i (t_i + C\xi_i) \\ +(1-\lambda)W(\boldsymbol{u}) \end{array} & \begin{array}{c} -t_i \leq \theta_j \leq t_j, \; t_j \geq 0, \text{ all } j \\ y_i(\boldsymbol{\theta}^\intercal \boldsymbol{x}^i + b) \geq 1 - \xi_i, \text{ all } i \\ u_i = c_0 + c_i\delta_i, \text{ all } i \\ -M(1-\delta_i) \leq \boldsymbol{\theta}^\intercal \boldsymbol{x}^i + b, \text{ all } i \\ \boldsymbol{\theta}^\intercal \boldsymbol{x}^i + b \leq M\delta_i, \text{ all } i \\ \delta_i \in \{0,1\}, \text{ all } i \end{array} \end{array} \right\}$$

Since the term $\sum_i (t_i + C\xi_i)$ in the objective function is linear, the objective function can be linearized whenever $W(\boldsymbol{u})$ can be linearized. In particular, there is an MILP model whenever previous sections describe an LP or MILP model for the social welfare function W.

7 Conclusion

We have shown how optimization can provide a general paradigm for incorporating fairness into AI and machine learning applications. In particular, we have illustrated how it can be used in rule-based systems, in conjunction with neural networks, or as part of the optimization problem in support vector machines. By expanding the fairness problem to one of maximizing a social welfare function, one can combine fairness with prediction accuracy and other efficiency goals in a principled way. Optimization models also provide the flexibility of adding constraints on resources and other problem elements while harnessing the power of highly advanced optimization solvers that have been developed over several decades.

Specifically, we have exhibited practical optimization models for 16 useful social welfare functions. Most of these models do not, to our knowledge, appear in previous literature. They can be efficiently solved by state-of-the-art software: six by linear programming solvers, eight by mixed integer/linear programming solvers, and two by convex nonlinear or quadratic programming solvers that accommodate linear constraints.

References

S. Aghaei, M. J. Azizi, and P. Vayanos. Learning optimal and fair decision trees for non-discriminative decision-making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1418–1426, 2019.

I. Alabdulmohsin. Fair classification via unconstrained optimization. *arXiv preprint*, 2005.14621, 2020.

C. Allen, I. Smit, and W. Wallach. Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology*, 7:149–155, 2005.

R. Berk, H. Heidari, S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, S. Neel, and A. Roth. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*, 2017.

K. Binmore, A. Rubinstein, and A. Wolinsky. The Nash bargaining solution in economic modeling. *RAND Journal of Economics*, 17:176–188, 1986.

P. S. Bradley and O. L. Mangasarian. Feature selection via concave minimization and support vector machines. In *ICML*, volume 98, pages 82–90, 1998.

S. Bringsjord, K. Arkoudas, and P. Bello. Toward a general logicist methodology for engineering ethically correct robots. *IEEE Intelligent Systems*, 21:38–44, 2006.

F. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pages 3992–4001, 2017.

A. Charnes and W. W. Cooper. Programming with linear fractional functionals. *Naval Research Logistics Quarterly*, 9:181–186, 1962.

V. Chen and J. N. Hooker. A just approach balancing Rawlsian leximax fairness and utilitarianism. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pages 221–227, 2020a.

V. Chen and J. N. Hooker. Balancing fairness and efficiency in an optimization model. *arXiv preprint*, 2006.05963, 2020b.

M. Chiarandini, R. Fagerberg, and S. Gualandi. Handling preferences in student-project allocation. *Annals of Operations Research*, 275(1):39–78, 2019.

A. Chouldechova and A. Roth. A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, 63(5):82–89, 2020.

F. A. Cowell. Measurement of inequality. In A. B. Atkinson and F. Bourguignon, editors, *Handbook of Income Distribution*, volume 1, pages 89–166. Elsevier, 2000.

M. Donini, L. Oneto, S. Ben-David, J. S. Shawe-Taylor, and M. Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, pages 2791–2801, 2018.

C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. S. Zemel. Fairness through awareness. In *Symposium on Innovations in Theoretical Computer Science (ITCS)*, pages 214–226, 2012.

O. Eisenhandler and M. Tzur. The humanitarian pickup and distribution problem. *Operations Research*, 67: 10–32, 2019.

H. G. Frankfurt. *On Inequality.* Princeton University Press, 2015.

S. Freeman, editor. *The Cambridge Companion to Rawls.* Cambridge University Press, 2003.

I. Gabriel. Artificial intelligence, values, and alignment. *Minds and Machines*, 30:411–437, 2020.

D. Gautier. *Morals by Agreement.* Oxford University Press, 1983.

N. Goel, M. Yaghini, and B. Faltings. Non-discriminatory machine learning through convex fairness criteria. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

J. C. Harsanyi. *Rational Behavior and Bargaining Equilibrium in Games and Social Situations.* Cambridge University Press, 1977.

H. Heidari, C. Ferrari, K. Gummadi, and A. 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.

H. Heidari, M. Loi, K. P. Gummadi, and A. Krause. A moral framework for understanding fair ml through economic models of equality of opportunity. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 181–190, 2019.

J. N. Hooker and T. W. Kim. Toward non-intuition-based machine and artificial intelligence ethics: A deontological approach based on modal logic. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pages 130–136, 2018.

J. N. Hooker and H. P. Williams. Combining equity and utilitarianism in a mathematical programming model. *Management Science*, 58:1682–1693, 2012.

L. Hu and Y. Chen. Fair classification and social welfare. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 535–545, 2020.

D. A. Iancu and N. Trichakis. Fairness and efficiency in multiportfolio optimization. *Operations Research*, 62(6):1285–1301, 2014.

S. P. Jenkins and P. Van Kerm. The measurement of economic inequality. In B. Nolan, W. Salverda, and T. M. Smeeding, editors, *The Oxford Handbook of Economic Inequality.* Oxford University Press, 2011.

E. Kalai and M. Smorodinsky. Other solutions to Nash's bargaining problem. *Econometrica*, 43:513–518, 1975.

O. Karsu and A. Morton. Inequity-averse optimization in operational research. *European Journal of Operational Research*, 245:343–359, 2015.

T. Lan, D. Kao, M. Chiang, and A. Sabharwal. An axiomatic theory of fairness in network resource allocation. In *Conference on Information Communications (INFOCOM 2010)*, pages 1343–1351. IEEE, 2010.

F. Lindner, R. Mattmüller, and B. Nebel. Evaluation of the moral permissibility of action plans. *Artificial Intelligence*, 287, 2020.

H. Luss. On equitable resource allocation problems: A lexicographic minimax approach. *Operations Research*, 47(3):361–378, 1999.

C. McElfresh and J. Dickerson. Balancing lexicographic fairness and a utilitarian objective with application to kidney exchange. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI 2018)*, pages 1161–1168, 2018.

N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint*, 1908.09635, 2019.

M. Mostajabdaveh, W. J. Gutjahr, and S. Salman. Inequity-averse shelter location for disaster preparedness. *IISE Transactions*, 51(8):809–829, 2019.

V. Nanda, P. Xu, K. A. Sankararaman, J. Dickerson, and A. Srinivasan. Balancing the tradeoff between profit and fairness in rideshare platforms during high-demand hours. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2210–2217, 2020.

J. Nash. The bargaining problem. *Econometrica*, 18:155–162, 1950.

W. Ogryczak and T. Śliwiński. On equitable approaches to resource allocation problems: The conditional minimax solutions. *Journal of Telecommunications and Information Technology*, pages 40–48, 2002.

W. Ogryczak and T. Śliwiński. On direct methods for lexicographic min-max optimization. In M. Gavrilova, O. Gervasi, V. Kumar, C. J. K. Tan, D. Taniar, A. Laganá, Y. Mun, and H. Choo, editors, *Proceedings of International Conference on Computational Science and Its Applications (ICCSA 2006)*, volume 3982 of *LNCS*, pages 802–811, 2006.

W. Ogryczak, A. Wierzbicki, and M. Milewski. A multi-criteria approach to fair and efficient bandwidth allocation. *Omega*, 36(3):451–463, 2008.

M. Olfat and A. Aswani. Spectral algorithms for computing fair support vector machines. In *International Conference on Artificial Intelligence and Statistics*, pages 1933–1942, 2018.

D. Parfit. Equality and priority. *Ratio*, pages 201–221, 1997.

J. Rawls. *A Theory of Justice* (revised). Harvard University Press (original edition 1971), 1999.

H. S. Richardson and P. J. Weithman, editors. *The Philosophy of Rawls* (5 volumes). Garland, 1999.

A. Rubinstein. Perfect equilibrium in a bargaining model. *Econometrica*, 50:97–109, 1982.

S. Russell. *Human Compatible: AI and the Problem of Control*. Bristol, UK: Allen Lane, 2019.

T. M. Scanlon. The diversity of objections to inequality. In T. M. Scanlon, editor, *The Difficulty of Tolerance: Essays in Political Philosophy*, pages 202–218. Cambridge University Press, 2003.

I. Stelmakh, N. B. Shah, and A. Singh. Peerreview4all: Fair and accurate reviewer assignment in peer review. *Proceedings of Machine Learning Research*, 98:1–29, 2019.

W. Thompson. Cooperative models of bargaining. In R. J. Aumann and S. Hart, editors, *Handbook of Game Theory*, volume 2, pages 1237–1284. North-Holland, 1994.

A. Williams and R. Cookson. Equity in health. In A. Culyer and J. Newhouse, editors, *Handbook of Health Economics*. Elsevier, 2000.

M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. 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, 2017.

M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20(75):1–42, 2019.

R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013.

J. Zhu, S. Rosset, R. Tibshirani, and T. Hastie. 1-norm support vector machines. *Advances in neural information processing systems*, 16:49–56, 2003.