A Guide to Formulating Equity and Fairness in an Optimization Model

Violet (Xinying) Chen[a], J.N. Hooker[a]

[a] *Carnegie Mellon University,*

Abstract

Optimization models typically seek to maximize overall benefit or minimize total cost. Yet equity and fairness are important elements of many practical decisions, and it is much less obvious how to express them mathematically. We provide a critical survey of various schemes that have been proposed for formulating ethics-related criteria, including those that integrate efficiency and equity concerns. The survey covers inequality measures, Rawlsian maximin and leximax criteria, alpha fairness and proportional fairness (also known as the Nash bargaining solution), Kalai-Smorodinsky bargaining, and recently proposed threshold-based schemes for combining utilitarian with maximin and leximax criteria. It also examines statistical fairness metrics that are popular in machine learning, including demographic parity, equalized odds, accuracy parity, and predictive rate parity. We present what appears to be the best practical approach to formulating each criterion in a linear, nonlinear, or mixed integer programming model. We analyze the mathematical properties of the various formulations, presenting new results in several cases, and indicate some of the strengths and weaknesses of each. We also cite relevant philosophical and ethical literature where appropriate.

Keywords: equity modeling, fairness, distributive justice

1. Introduction

There is growing interest in incorporating equity-related criteria into optimization models. Practical applications in health care, disaster management, telecommunications, facility location, and other areas increasingly raise issues related to the fair allocation of resources. Yet it is far from

Email addresses: xinyingc@andrew.cmu.edu (Violet (Xinying) Chen), jh38@andrew.cmu.edu (J.N. Hooker)

obvious how to formulate such ethical concerns mathematically. While it is normally straightforward to formulate an objective function that reflects efficiency or cost, fairness can be understood in multiple ways, with no generally accepted method for representing any of them in a mathematical idiom. While methods for formulating equity concerns frequently appear in research papers, they are often discussed and selected in an ad hoc manner.

We therefore undertake to provide a survey and assessment of a broad range of equity criteria that can be incorporated into an optimization model. We cover inequality metrics, Rawlsian maximin and leximax criteria, various convex combinations of these, alpha fairness and proportional fairness (the latter also known as the Nash bargaining solution), the Kalai-Smorodinsky bargaining solution, and recently proposed threshold-based criteria for combining utilitarianism with maximin and leximax criteria. We also examine statistical fairness metrics that are popular in machine learning, including demographic parity, equalized odds, accuracy parity, and predictive rate parity. We present what appears to be the best practical approach to formulating each criterion in a linear, nonlinear, or mixed integer programming model. We analyze some mathematical properties of the various formulations, present new results in several cases, and indicate some of the strengths and weaknesses of each. We place particular emphasis on models that combine efficiency and equity criteria, because both are important in most practical applications.

To our knowledge, there is no existing survey of this kind. Karsu and Morton (2015) discuss several models in their excellent survey of inequality-averse optimization, along with applications and some underlying mathematical theory. Ogryczak et al. (2014) survey fairness criteria that have been used in communication networks and location models, with a discussion of their properties and relationship with leximax criteria. Our contribution differs from these in that it aims for broad coverage of equity concepts while providing a practical guide for the analyst who wishes to incorporate equity concerns into an optimization model of a given application. It accordingly includes a focus on how to formulate the various criteria for efficient solution by mathematical programming software. It also covers equity formulations developed since the earlier surveys, as well as fairness measures from machine learning. Finally, it provides some references to relevant philosophical and ethical literature, although we make no attempt here to resolve underlying philosophical issues.

We begin below by stating a generic optimization problem that provides a framework for the discussion to follow. In particular, we suppose that each equity criterion we consider is encapsulated in a social welfare function

(SWF) that serves as the objective function of the optimization model. We next describe two properties that are possessed by some of the SWFs and that are helpful for understanding the nature of the equity concepts they represent. These are the well-known Pigou–Dalton condition and a lesser known, slightly weaker Chateauneuf–Moyes condition that is arguably better suited to assess equity criteria. We determine which SWFs satisfy one or both of these conditions, resulting in several new theorems that are proved in the Appendix. The equity criteria we study are summarized in Tables 1–3, which indicate the section of the paper that deals with each. The concluding section of the paper draws on the foregoing discussion to suggest some general guidelines for selecting an equity criterion for a given application.

In the interest of brevity, we omit discussion of several fairness metrics that are similar to those covered here, designed for specific applications, and/or difficult to optimize. These include some of the fairness measures developed for communications networks, such as Jain's index (Jain et al. 1984), QoE fairness (Georgopoulos et al. 2013, Hoßfeld et al. 2018), TCP fairness (Pokhrel et al. 2016), G's fairness index, and Bossaert's fairness index (Mehta 2020). Additional metrics from the economics literature include the entropy-based Theil index (Theil 1967, Cowell and Kuga 1981) and the related Atkinson index (Atkinson 1975). We also omit some statistical bias measures used in machine learning, as described in Section 10.

2. Generic Optimization Problem

The task before us is to incorporate equity into an existing optimization model of the form $\max_{\boldsymbol{x}}\{f(\boldsymbol{x}) \mid \boldsymbol{x} \in S_{\boldsymbol{x}}\}$. A generic optimization problem that accomplishes this can be stated

$$\max_{\boldsymbol{u},\boldsymbol{x}} \big\{ W(\boldsymbol{u}) \;\big|\; \boldsymbol{u} = U(\boldsymbol{x}),\; \boldsymbol{x} \in S_{\boldsymbol{x}} \big\} \tag{1}$$

where $\boldsymbol{u} = (u_1, \ldots, u_n)$ is a vector of utilities distributed across parties $1, \ldots, n$. The utilities can be profit, negative cost, or some other benefit that is appropriate to the application. It is this distribution of utilities that we wish to be equitable as well as, perhaps, efficient. We replace the original objective function $f(\boldsymbol{x})$, if any, with a *social welfare function* $W(\boldsymbol{u})$ that measures the desirability of a given utility distribution $\boldsymbol{u}$. We want $W(\boldsymbol{u})$ to incorporate equity, as well as perhaps efficiency elements measured by $f(\boldsymbol{x})$. The intent of this paper is to survey and evaluate functions $W(\boldsymbol{u})$.

Table 1: Summary of fairness criteria, part 1. The columns labeled P–G and C–M indicate whether the Pigou–Dalton and Chateauneuf–Moyes conditions are satisfied. The model type assumes that the original problem constraints are linear with continuous variables.

Criterion	P–G?	C–M?	Model	Comments
Inequality measures				
Relative range (Section 4.1)	yes	yes	LP	The spread between min and max utilities, normalized by the mean. Inequality metrics may be appropriate when there is a particular interest in equality rather than broader concepts of fairness.
Relative mean deviation (Section 4.1)	yes	yes	LP	The normalized average deviation from the mean. Takes in account all utilities rather than only the two extremes.
Coefficient of variation (Section 4.1)	yes	yes	NLP[1]	The normalized standard deviation. Can be used when large deviations from the mean are disproportionately significant.
Gini coefficient (Section 4.2)	yes	yes	LP	Perhaps the best known measure of inequality. Proportional to the area between the Lorenz curve and a diagonal line representing perfect equality. Lies in the interval [0,1], with 0 indicating perfect equality.
Hoover index (Section 4.2)	yes	yes	LP	The fraction of total utility that must be redistributed to achieve perfect quality. Also related to the Lorenz curve, and proportional to the relative mean deviation.
Fairness for the disadvantaged				
Maximin (Section 5.1)	yes	yes	LP	Maximizes the minimum utility. Based on the Rawlsian principle that inequality is justified only to the extent that it improves the welfare of the worst off. Once maximin is obtained, does not consider the welfare of other disadvantaged individuals.
Leximax (Section 5.1)	yes	yes	LP	Maximizes the welfare of the worst off, then the 2nd worst off, and so forth. Considers the welfare of all disadvantaged individuals but requires solving a sequence of optimization problems.
McLoone index (Section 5.2)	no	yes	MILP	Compares total utility utility of those below the median to what they would enjoy if brought up to the median. Concerned only with the welfare of the lower half.

[1] A convex quadratic programming problem with linear constraints.

Table 2: Summary of fairness criteria, part 2.

Criterion	P–G?	C–M?	Model	Comments
Combining efficiency and fairness – Classical methods				
Alpha fairness (Section 7.1)	yes	yes	NLP[2]	Parameter α regulates equity vs efficiency, with $\alpha = 0$ corresponding to a pure utilitarian and $\alpha = \infty$ to a pure maximin criterion. Unclear how to interpret α in practice.
Proportional fairness (Section 7.1)	yes	yes	NLP[2]	Special case of alpha fairness with $\alpha = 1$, also known as the Nash bargaining solution, and used in engineering applications. Has been justified with axiomatic and bargaining arguments, albeit with a strong interpersonal noncomparability assumption.
Kalai-Smorodinsky bargaining (Section 7.2)	no	no	LP	Maximizes minimum relative concession by maximizing equal fraction of each player's potential gain. Can be defended as outcome of a bargaining procedure and tends to favor those with greater opportunity. Failure of P–G and C–M conditions may be a concern.
Combining efficiency and maximin fairness – Threshold methods				
Utility threshold (Section 8.1)	no	yes	MILP[3]	Uses a maximin criterion until utility cost of fairness becomes too great, and then switches some players to a utilitarian criterion. The break point is controlled by parameter Δ, selected so that players within Δ of the lowest utility are seen as sufficiently disadvantaged to receive greater priority. Equity component is sensitive to the utility level only of the worst-off player.
Equity threshold (Section 8.2)	yes	yes	LP	Uses a utilitarian criterion until inequity becomes too great, and then switches some players to a maximin criterion. The parameter Δ is selected so that players already more than Δ above the lowest utility are not seen as deserving greater utility if the other utilities remain unchanged. The equity component is again sensitive to the utility level only of the worst-off player.

[2] A concave nonlinear maximization problem with linear constraints.
[3] The MILP model of the threshold function is sharp (defines convex hull of feasible set).

Table 3: Summary of fairness criteria, part 3.

Criterion	P–G?	C–M?	Model	Comments
Combining efficiency and leximax fairness – Threshold methods				
Utility threshold, predefined priorities (Section 9.1)	no	no	MILP	Maximizes a utility threshold function that combines utilitarian and maximin criteria, then applies a leximax criterion to optimal solutions if one or more have a utility spread of Δ or less. Makes the strong assumption that priorities of the players can be fixed in advance. SWF is discontinuous for $n \geq 3$, a potential drawback.
Utility threshold, no predefined priorities (Section 9.2)	no	yes	MILP[4]	Solves a sequence of optimization problems in which the kth problem determines the kth smallest utility in the socially optimal solution. Each problem assumes the smallest $k-1$ utilities have been fixed and maximizes a SWF that combines utilitarian and maximin criteria while giving the kth worst-off player priority that is regulated by Δ. Combines leximax and utilitarian criteria and so considers utilities of all disadvantaged players, not just the very worst-off.
Statistical fairness metrics				
Demographic parity (Section 10.1)			LP	The fraction of minority individuals selected for a benefit should be the same as the fraction of majority individuals selected. A very strict criterion that can deny individuals benefits for which they are known to be qualified.
Equalized odds (Section 10.2)			LP	The fraction of (un)qualified individuals who are selected should be the same in minority and majority groups.
Accuracy parity (Section 10.3)			LP	The accuracy rate (fraction of individuals correctly selected or rejected) should be the same in minority and majority groups.
Predictive rate parity (Section 10.4)			MINLP[5]	The fraction of individuals correctly selected should be the same for minority and majority groups. Unclear that the advantages of this criterion (if any) justify solving the difficult optimization problem that results.

[4] A sequence of tractable MILP models is solved. Valid inequalities are identified.

[5] A difficult mixed integer/nonlinear optimization problem.

The vector-valued function $U(\boldsymbol{x})$ defines how the original problem variables $\boldsymbol{x}$ determine the distribution of utilities. In many applications, some of the original variables x_j already represent the utilities we wish to distribute, and there is no need to introduce additional variables u_i. Nonetheless, we will consistently refer to the utilities to be distributed as $u_1, \ldots, u_n$. To simplify notation, we will suppose that the constraint $\boldsymbol{u} = U(\boldsymbol{x})$ is encoded in constraints represented by $(\boldsymbol{u}, \boldsymbol{x}) \in S$, so that the problem (1) becomes simply

$$\max_{\boldsymbol{u},\boldsymbol{x}} \big\{ W(\boldsymbol{u}) \;\big|\; (\boldsymbol{u}, \boldsymbol{x}) \in S \big\} \tag{2}$$

Thus $(\boldsymbol{u}, \boldsymbol{x}) \in S$ if and only if $\boldsymbol{u} = U(\boldsymbol{x})$ and $\boldsymbol{x} \in S_{\boldsymbol{x}}$.

Fairness can also be represented as a constraint by bounding the social welfare function $W(\boldsymbol{u})$. This results in an optimization problem of the form

$$\max_{\boldsymbol{u},\boldsymbol{x}} \big\{ f(\boldsymbol{x}) \;\big|\; W(\boldsymbol{u}) \geq \text{LB},\ (\boldsymbol{u}, \boldsymbol{x}) \in S \big\} \tag{3}$$

To simplify exposition, we will discuss only models of the form (2), but they can be converted to fairness-constrained problems when desired.

A simple medical triage problem provides an illustration of the generic model. There are n patients who require treatment, but subject to a limited budget of B. The cost of treating patient i is c_i. The utility experienced by patient i, measured in quality-adjusted life years, is a_i without treatment and $a_i + b_i$ with treatment. The objective is to allocate treatments in an equitable and effective fashion. If binary variable $x_i = 1$ when patient i is treated, the utility function $\boldsymbol{U}(\boldsymbol{x})$ is given by $U_i(\boldsymbol{x}) = a_i x_i + b_i$ for $i = 1, \ldots, n$. The resulting optimization problem (2) is

$$\max_{\boldsymbol{u},\boldsymbol{x}} \left\{ W(\boldsymbol{u}) \;\middle|\; \begin{array}{l} \displaystyle\sum_i c_i x_i \leq B \\ u_i = a_i + b_i x_i,\ x_i \in \{0,1\},\ \text{all } i \end{array} \right\}$$

The choice of social welfare function $W(\boldsymbol{u})$ should reflect how equity and effectiveness are to be understood and balanced in this context.

A major element of this paper is showing how to write the optimization problem (2) in a form suitable for one of the highly advanced mathematical programming solvers now available. Naturally, the difficulty of (2) depends to a great degree on the nature of the constraints that describe the feasible set $S_{\boldsymbol{x}}$. However, if we suppose that these constraints are (or can be approximated by) linear equations and inequalities over continuous variables, the resulting models have the form indicated in Tables 1–3. Eleven of the 19 are linear programming (LP) models, a problem class that is extremely well solved. Four are mixed integer/linear programming (MILP) problems, which are combinatorial in nature, but for which highly developed solvers are available. Three are nonlinear programming (NLP) problems that are of only moderate difficulty because they minimize convex functions (or maximize concave functions) subject to linear constraints. Indeed, one of these is a highly tractable convex quadratic programming problem with linear constraints. Only one of the models, a nonconvex mixed integer/nonlinear programming (MINLP) problem, poses a substantial and possibly insuperable challenge.

The linearity assumption for constraints is actually quite reasonable in many applications, because it is consistent with a great deal of flexibility to define the problem. Suppose, for example, that the set $S_{\boldsymbol{x}}$ of feasible values of $\boldsymbol{x}$ is convex, and the utility function $U(\boldsymbol{x})$ is linear or concave, as it commonly is when there are nonincreasing returns to scale. In such a case, the feasible set S can be approximated to any desired degree with a linear system $A\boldsymbol{u} + B\boldsymbol{x} \leq \boldsymbol{b}$.

If some of the original problem variables x_i are discrete, however, an otherwise LP problem becomes an MILP problem, as in the medical example just stated. An NLP problem becomes a mixed integer/nonlinear programming (MINLP) problem, which can be quite hard to solve. An MILP problem of course remains an MILP problem.

3. Pigou–Dalton and Chateauneuf–Moyes Conditions

The Pigou–Dalton condition is frequently used to assess social welfare functions, particularly those that measure equality (Dalton 1920, Moulin 2004). It is satisfied when any utility transfer from a better-off party to a worse-off party increases (or does not reduce) social welfare. However, the suitability of this condition for assessing an equity metric has been questioned, for example by Chateauneuf and Moyes (2005). They propose a slightly weaker condition that considers transfers of utility from a better-off class to a worse-off class rather than from one individual to another. The choice of condition is relevant here, because we find that some interesting SWFs satisfy the weaker condition but not the stronger.

Formally, a social welfare function $W(\boldsymbol{u})$ satisfies the Pigou–Dalton condition if $W(\boldsymbol{u} + \epsilon\mathbf{e}_i - \epsilon\mathbf{e}_j) \geq W(\boldsymbol{u})$ for any i, j and any $\epsilon > 0$ for which $u_i + \epsilon \leq u_j - \epsilon$, where $\mathbf{e}_i$, $\mathbf{e}_j$ are the ith and jth unit vectors, respectively. A stricter form of the condition requires $W(\boldsymbol{u} + \epsilon\mathbf{e}_i - \epsilon\mathbf{e}_j) > W(\boldsymbol{u})$, but we use the weaker form.

The Chateauneuf–Moyes (C–M) condition examines the consequences of transferring a given amount of utility from individuals whose utility lies at or above that of a given individual (taking an equal share from each) to those whose utility lies at or below below that of a less fortunate individual (giving an equal share to each). Chateauneuf and Moyes provide arguments for why this type of condition is preferable to Pigou–Dalton. One is the simple observation that while a pairwise Pigou–Dalton transfer reduces inequality between two individuals, it may increase inequality between those individuals and others. A C–M transfer does not incur this problem, because the donor and recipient classes respectively lie completely above and below the rest of the population.

To define the C–M condition formally, let us say that a *C–M transfer* is a transfer of utility from $\boldsymbol{u}$ to $\boldsymbol{u}'$ such that $u_1 \leq \cdots \leq u_n$ as well as $u'_1 \leq \cdots \leq u'_n$, and for some pair of integers ℓ, h with $1 \leq \ell < h \leq n$, we have $u_\ell < u_h$ and

$$\boldsymbol{u}' = \boldsymbol{u} + \frac{\epsilon}{\ell}\sum_{i=1}^{\ell}\boldsymbol{e}_i - \frac{\epsilon}{n-h+1}\sum_{i=h}^{n}\boldsymbol{e}_i$$

for some $\epsilon > 0$. A SWF $W(\boldsymbol{u})$ satisfies the C–M condition if C–M transfers never decrease social welfare. That is, $W(\boldsymbol{u}') \geq W(\boldsymbol{u})$ for any C–M transfer from $\boldsymbol{u}$ to $\boldsymbol{u}'$.

It is difficult to say in general whether a SWF should satisfy either of these conditions, but an indication of whether they do so may be useful in determining whether the SWF is suitable for a particular application.

4. Inequality Measures

The first type of fairness measure we study 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). There is a wide variety of philosophical opinion on the ethical significance of equality, ranging from the view that we have an irreducible obligation to strive for equality, to the view that inequality is unfair only when it reduces total utility (Frankfurt 2015, Parfit 1997, Scanlon 2003). In any event, it is generally acknowledged that equality is not the same concept (or cluster of concepts) as fairness, even when the two are closely related. An equality metric can be appropriate in a context where a specifically egalitarian distribution is the primary goal, without regard for efficiency or other forms of equity.

Inequality measures have been used for inequity averse optimization in a broad range of applications. Examples of these papers are summarized in Karsu and Morton (2015). More recently, inequality measures are also considered in the growing area of algorithmic fairness. For instance, Leonhardt et al. (2018) study Gini coefficient type measures for estimating the disparity in user satisfaction and recommendation quality of recommender systems. Speicher et al. (2018) and Sühr et al. (2019) respectively adopt a generalized entropy index to evaluate the degree of unfairness in predictors trained by machine learning and in two-sided matching platforms.

We present optimization models for relative range, relative mean deviation, coefficient of variation, the Gini coefficient, and the Hoover index. All of them are easily shown to satisfy the Pigou–Dalton condition. The McLoone index can also be regarded as a measure of inequality, but we consider it in the next section as measuring fairness for the disadvantaged.

4.1. Measures of relative dispersion

All of the dispersion measures we consider are normalized by the mean utility so as to be invariant under rescaling of utilities. The *relative range* of utilities is an inequality metric that, when negated, yields the SWF

$$W(\boldsymbol{u}) = -(1/\bar{u})\big(u_{\max} - u_{\min}\big)$$

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 the constraint set implies $\bar{u} > 0$. Then since the SWF is a ratio of affine functions, the formulation of $W(\boldsymbol{u})$ in an optimization model can be linearized using the same change of variable as in linear-fractional programming (Charnes and Cooper 1962). Thus we introduce a scalar variable t and write $\boldsymbol{u} = \boldsymbol{u}'/t$ and $\boldsymbol{x} = \boldsymbol{x}'/t$, which yields the optimization model

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

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. The tractability of this model depends on whether the constraints defining S become harder after the change of variable. The easiest case arises when the constraints are linear, as in linear-fractional programming. If the original constraints are $A\boldsymbol{u} + B\boldsymbol{x} \leq \boldsymbol{b}$, they become another linear system $A\boldsymbol{u}' + B\boldsymbol{x}' \leq t\boldsymbol{b}$ after the variable change.

More generally, if the original constraints have the form $\boldsymbol{g}(\boldsymbol{u}, \boldsymbol{x}) \leq \boldsymbol{b}$ for homogeneous $\boldsymbol{g}$, they retain essentially the same form $\boldsymbol{g}(\boldsymbol{u}', \boldsymbol{x}') \leq t\boldsymbol{b}$ after the variable change.

Another dispersion metric is the *relative mean deviation*, which measures inequality more comprehensively by considering all utilities rather than only the minimum and maximum. The SWF is

$$W(\boldsymbol{u}) = -(1/\bar{u}) \sum_i |u_i - \bar{u}|$$

This can be linearized by using the same change of variables as before:

$$\min_{\boldsymbol{x}', \boldsymbol{u}', \boldsymbol{v}, t} \left\{ \sum_i v_i \;\middle|\; \begin{matrix} -v_i \leq u_i' - \bar{u}' \leq v_i, \text{ all } i \\ \bar{u}' = 1, \; t \geq 0, \; (\boldsymbol{u}', \boldsymbol{x}') \in S' \end{matrix} \right\} \tag{4}$$

where $v_1, \ldots, v_n$ are new variables. This is again an LP problem if S is defined by a linear constraint set.

The *coefficient of variation* is the normalized standard deviation. It may be appropriate when large deviations from the mean are disproportionately significant, but it has the possible drawback of introducing a nonlinear objective function. The SWF is

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

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

$$\min_{\boldsymbol{x}', \boldsymbol{u}', \boldsymbol{v}, t} \left\{ \Big[\frac{1}{n} \sum_i (u_i' - \bar{u}')^2 \Big]^{\frac{1}{2}} \;\middle|\; \begin{matrix} \bar{u}' = 1, \; t \geq 0 \\ (\boldsymbol{u}', \boldsymbol{x}') \in S' \end{matrix} \right\}$$

This is not an LP problem, but we can obtain the same optimal solution by solving it without the exponent $\frac{1}{2}$. If the feasible set S' is convex, this yields a convex nonlinear programming problem in which all local optima are global optima. If S is defined by linear constraints, it can be solved by particularly efficient quadratic programming algorithms that are available in many state-of-the-art optimization packages.

4.2. Gini coefficient and Hoover index

The *Gini coefficient* is by far the best known measure of inequality, as it is routinely used to measure income and wealth inequality (Gini 1912).

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 $W(\boldsymbol{u}) = -G(\boldsymbol{u})$, where

$$G(\boldsymbol{u}) = \frac{1}{2\bar{u}n^2} \sum_{i,j} |u_i - u_j|$$

Again applying the change of variable from linear-fractional programming, the Gini criterion can be linearized:

$$\min_{\boldsymbol{x}',\boldsymbol{u}',V,t} \left\{ \frac{1}{2n^2} \sum_{i,j} v_{ij} \;\middle|\; \begin{array}{c} -v_{ij} \leq u'_i - u'_j \leq v_{ij}, \text{ all } i,j \\ \bar{u}' = 1,\; t \geq 0,\; (\boldsymbol{u}', \boldsymbol{x}') \in S' \end{array} \right\}$$

where v_{ij} is a new variable for all i,j. This is an LP problem if S is defined by linear constraints.

The *Hoover index* is also related to the Lorenz curve, as it is proportional to the maximum vertical distance between the Lorenz curve and a diagonal line representing perfect equality (Hoover 1936). It is also proportional to the relative mean deviation and therefore satisfies the Pigou–Dalton condition. It can be interpreted as the fraction of total utility that would have to be redistributed to achieve perfect equality. The SWF is

$$W(\boldsymbol{u}) = -\frac{1}{2n\bar{u}} \sum_{i} |u_i - \bar{u}|$$

The Hoover index can be minimized by solving the same model (4) as for the relative mean deviation.

5. Fairness for the Disadvantaged

Rather than focus solely on inequality, fairness measures can prioritize the disadvantaged. 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 (surveyed in Freeman 2003, Richardson and Weithman 1999). The difference principle can be plausibly extended to a lexicographic maximum principle. There is also the McLoone index, which is a statistical measure that emphasizes the lot of the less advantaged.

The Rawlsian maximin criterion has been a popular fairness measure for decades. Early works on fair resource allocation, such as bandwidth allocation, often choose the maximin criterion to seek the best possible performance for the worst-off service among services competing for bandwidth (Luss 1999, Ogryczak and Śliwiński 2002, Ogryczak et al. 2008). Recent research has applied the criterion to more diverse problem contexts. For example, Stelmakh et al. (2018) design an algorithm for making paper-reviewer assignment that maximizes the review quality of the most disadvantaged paper, and Nanda et al. (2020) formalize a maximin fairness measure for ridesharing. In addition, the Rawlsian view of fairness is gaining recognition in machine learning as an alternative to the dominant statistical fairness metrics (Hashimoto et al. 2018, Heidari et al. 2019, Shah et al. 2021).

5.1. 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. Rawls argues that 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 objective is readily linearized, as in the optimization model

$$\max_{\boldsymbol{x},\boldsymbol{u},w} \left\{ w \;\middle|\; w \leq u_i, \text{ all } i;\ (\boldsymbol{u},\boldsymbol{x}) \in S \right\}$$

The maximin criterion obviously satisfies the Pigou–Dalton condition, although almost vacuously, because it considers only the smallest utility level.

The maximin criterion can force equality even when doing so is very costly in terms of total utility. Suppose, for example that S is defined only by a budget constraint $\sum_i x_i \leq B$ (with $\boldsymbol{x} \geq \boldsymbol{0}$) and utility functions $u_i = a_i x_i$. Then the maximin solution equalizes the utilities, with each individual experiencing utility $u_0 = B/\sum_i(1/a_i)$. If individual k's welfare is very expensive to provide, perhaps due to an incurable disease, then a_k is very small, and individual k consumes almost all the resources, u_0/a_k. The utility of everyone else is reduced to the same low level u_0 that can be achieved for individual k. One might impose an upper bound d_k on individual k's resource consumption, but then the maximin criterion is satisfied by reducing everyone's utility even more, namely to individual k's utility $a_k d_k$. This leaves unused resources $B - d_k a_k \sum_i (1/a_i)$, but the maximin criterion provides no incentive to distribute them.

The maximin criterion can be plausibly extended to *lexicographic maximization* (leximax), which can remove the problem of leftover resources in the previous example. Leximax is achieved by first maximizing the smallest utility subject to resrouce constraints, then the second smallest, and so forth. While this can avoid leftover resources, it does not avoid the possibly high cost of equality in the absence of constraints that prevent it.

A leximax solution can computed by solving a sequence of optimization problems

$$\max_{\boldsymbol{x},\boldsymbol{u},w} \left\{ w \;\middle|\; \begin{matrix} w \leq u_i, \; u_i \geq \hat{u}_{i_{k-1}}, \; i \in I_k \\ (\boldsymbol{u}, \boldsymbol{x}) \in S \end{matrix} \right\} \tag{5}$$

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 i_k is defined so that

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

If there are two or more utilities $\hat{u}_i$ that achieve the minimum $\min_{i \in I_k}\{\hat{u}_i\}$, it is necessary to enumerate all solutions that result from breaking the tie to be assured of finding a leximax soution. Ogryczak and Sliwinski (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.

5.2. McLoone index

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 the utility distribution has a very long lower tail (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

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

where $\tilde{u}$ 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 $I(\boldsymbol{u}) = \{i \mid u_i \leq \tilde{u}\}$.

The McLoone index violates the Pigou–Dalton condition but satisfies the Chateauneuf–Moyes condition. A violation of Pigou–Dalton can be seen in a 3-person example with $\boldsymbol{u} = (a, 1, b)$, where $0 < a < 1 < b$. The median is 1 and the McLoone index is $\frac{1}{2}(1 + a)$. A utility transfer from u_3 to u_2 yields the utility vector $(a, 1 + \epsilon, b - \epsilon)$, where we suppose $0 < \epsilon \leq \frac{1}{2}(b - 1)$. The median is now $1 + \epsilon$ and the McLoone index is $\frac{1}{2}(1 + a + \epsilon)/(1 + \epsilon)$. The McLoone index has become smaller despite less inequality, a violation of the P–G condition. However, we have the following.

Theorem 1. The McLoone index satisfies the Chateauneuf–Moyes condition.

We can formulate the McLoone index optimization problem as a mixed integer programming (MIP) problem 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

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

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 objective function can be linearized by using the same change of variable as in linear-fractional programming:

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

The model is an MILP problem when the constraints defining S are linear.

6. Convex Combinations

We now move to schemes that combine efficiency and fairness. The most obvious approach is to maximize a convex combination of the two:

$$F(\boldsymbol{u}) = (1-\lambda)\sum_i u_i + \lambda\Phi(\boldsymbol{u})$$

where $\Phi(\boldsymbol{u})$ is an equity measure. A perennial problem with convex combinations is that it is difficult to interpret λ, particularly when $\Phi(\boldsymbol{u})$ is measured in units other than utility. For example, if we use the Gini coefficient $G(\boldsymbol{u})$ as a measure of inequity, then we must combine utility with a dimensionless quantity $\Phi(\boldsymbol{u}) = 1 - G(\boldsymbol{u})$. Larger values of λ give greater weight to equality, but in a practical situation it is unclear how to attribute any meaning to a chosen value of λ.

Eisenhandler and Tzur (2019) use a product rather than a convex combination of utility and $1 - G(\boldsymbol{u})$, which nicely reduces to an SWF that is easily linearized:

$$W(\boldsymbol{u}) = \sum_i u_i - \frac{1}{n}\sum_{i<j}|u_j - u_i|$$

Yet we now have a convex combination of total utility and another equality metric (one that is proportional to the negative mean absolute difference); in particular, it is a convex combination in which $\lambda = 1/2$. This may be reasonable for the intended application, but one may ask why this particular value of λ is suitable, and whether other values should be used in other contexts. Aside from this are the general issues raised by using equality as a surrogate for fairness.

Mostajabdaveh et al. (2019) use a linear combination that is equivalent to $\sum_i u_i + \mu(1 - G(\boldsymbol{u}))\sum_i u_i$, where $\mu \in [0,1]$. This at least combines quantities measured in the same units. Yet we again have the problem of justifying a weight μ. In fact, this combination is equivalent to the convex combination implied by the Eisenhandler and Tzur criterion, except that λ is $\mu/(1+2\mu)$ rather than $\frac{1}{2}$.

One can combine utility with the Rawlsian maximin criterion by using the convex combination

$$W(\boldsymbol{u}) = (1-\lambda)\sum_i u_i + \lambda\min_i\{u_i\} \tag{6}$$

This, like the proposal of Mostajabdaveh et al., combines quantities that are measured in the same units. Yet it is again unclear how to select a suitable value of λ. Note that if we index utilities so that $u_1 \leq \cdots \leq u_n$, (6) is simply a weighted sum $u_1 + (1-\lambda)\sum_{i>1} u_i$ that gives somewhat more weight to the lowest utility. Yet how much more is appropriate?

One can refine criterion (6) by giving gradually decreasing weights $w_1 > w_2 > \cdots > w_n$ to the utilities in an SWF of the form

$$W(\boldsymbol{u}) = \sum_i w_i u_i \tag{7}$$

Yet this only complicates the task of assigning weights. In addition, since we do not know how to index the utilities by size in advance, we have the difficult modeling challenge of ensuring that weight w_i is assigned to the ith smallest utility. There is a long line of work studying this formulation as the objective function for multi-criteria decision making (e.g., Yager 1997, Ogryczak and Śliwiński 2003). Hu and Chen (2020) provide a novel perspective on this SWF in machine learning: they view (7) as the objective function in a classifier training model and establish its correspondence with the commonly studied fairness constrained loss-minimization training models.

7. Alpha Fairness and Kalai-Smorodinksy Bargaining

Alpha fairness and Kalai-Smorodinksy bargaining provide alternative and perhaps more satisfactory means of combing equity and efficiency than convex combinations. Alpha fairness regulates the combination with a continuous parameter α, where larger values of α signify a greater emphasis on equity. A famous special case is the Nash bargaining solution, which corresponds to $\alpha = 1$. Kalai-Smordinsky bargaining, proposed as an alternative to Nash bargaining, allots the parties the largest possible fraction of their potential utility while observing fairness by equalizing that fraction.

7.1. Alpha fairness and Nash bargaining

Alpha fairness (Mo and Walrand 2000, Verloop et al. 2010) is represented by a family of SWFs having the form

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

These SWFs form a continuum that stretches from a utilitarian criterion ($\alpha = 0$) to a maximin criterion as $\alpha \to \infty$. Lan et al. (2010) provide an axiomatic treatment of α-fairness in the context of network resource allocation, and Bertsimas et al. (2012) study worst-case equity/efficiency trade-offs implied by this criterion.

The parameter α can be interpreted as quantifying the equity/efficiency trade-off, because utility u_j must be reduced by $(u_j/u_i)^\alpha$ units to compensate for a unit increase in u_i ($< u_j$) while maintaining constant social welfare. This gives priority to less-advantaged parties, as we desire, with α indicating how much priority. In particular, the fact that $(u_j/u_i)^\alpha > 1$ when $u_i < u_j$ implies that $W_\alpha(\boldsymbol{u})$ satisfies the Pigou–Dalton condition for all α. Yet it is not obvious what kind of trade-off, and therefore what value of α, is appropriate for a given application. There is no apparent interpretation of α independent of its role in the SWF.

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

$$\max_{\boldsymbol{x},\boldsymbol{u}} \left\{ W_\alpha(\boldsymbol{u}) \;\middle|\; (\boldsymbol{u},\boldsymbol{x}) \in S \right\}$$

without reformulation. The objective function is irreducibly nonlinear, but it is concave for all $\alpha \geq 0$. Thus any local optimum is a global optimum if the feasible set is convex. The problem can be solved to optimality by such efficient algorithms as the reduced gradient method, which is a generalization of the simplex method for LP. The fact that $W_\alpha(\boldsymbol{u})$ has a simple closed-form gradient simplifies solution. Maximizing alpha fairness may therefore be tractable for reasonably large instances, particularly if the constraints defining S are linear.

Proportional fairness results from setting $\alpha = 1$ and is often measured by the product $\Pi_i u_i$ rather than its logarithm. Maximizing proportional fairness yields the Nash bargaining solution (Nash 1950), which should not be confused with the Nash equilibrium of game theory. It corresponds to selecting a point $\boldsymbol{u}$ in the feasible set that maximizes the volume of the hyperrectangle with opposite corners at $\boldsymbol{u}$ and the origin. This is illustrated in Fig. 1(a), where each point on the plot represents the utility outcomes for two parties that result from some distribution of resources. The set of feasible utility vectors is the area under the curve. The Nash bargaining solution is the black dot, which is the feasible point that maximizes the area of the shaded rectangle. Proportional fairness is frequently used in engineering, such as for bandwidth allocation in telecommunication networks and traffic signal timing (Mazumdar et al. 1991, Kelly et al. 1998).

Proportional fairness has axiomatic and bargaining-based derivations that might be seen as justifying the parameter setting $\alpha = 1$. For example, Nash (1950) showed that his bargaining solution for two persons is implied

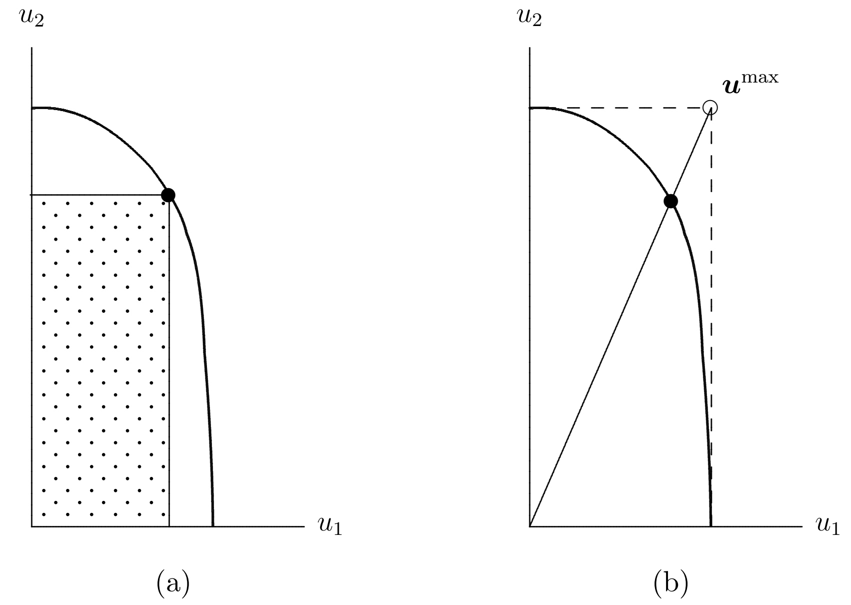


Figure 1: (a) Nash bargaining solution for two players. (b) Kalai-Smorodinsky bargaining solution for two players. In both cases, the default position is the origin.

by a set of axioms for utility theory. Harsanyi (1977), Rubinstein (1982), and Binmore et al. (1986) showed that the Nash solution is the (asymptotic) outcome of certain rational bargaining procedures. Yet the axiomatic derivation relies on a strong axiom of cardinal noncomparability across parties that is central to the proof. The axiom assumes that the ranking of utility vectors is invariant under affine transformations of the form $\phi_i(u_i) = \beta_i u_i + \gamma_i$, which arguably rules out the kind of utility comparisons we need in order to assess fairness (Hooker 2013). Furthermore, the bargaining theories assume that the parties begin with a default utility allocation $\boldsymbol{d} = (d_1, \ldots, d_n)$ on which they fall back if bargaining fails. The proportional fairness SWF then becomes $W(\boldsymbol{u}) = \Pi_i(u_i - d_i)$. An unfair starting point $\boldsymbol{d}$ could lead to an unfair outcome even under a rational bargaining procedure, and even if we grant that rational bargaining from a fair starting point necessarily yields a fair outcome. This weakens the bargaining argument for the fairness of the Nash solution in general.

Another issue with proportional fairness, and alpha fairness in general, is that they can assign equality the same social welfare as arbitrarily extreme inequality. In a 2-player situation, for example, the distribution $\boldsymbol{u} = (s, s)$

has the same social welfare value as (t, T), where

$$t = \begin{cases} s^2/T & \text{if } \alpha = 1 \\ \left(2s^{1-\alpha} - T^{1-\alpha}\right)^{1/(1-\alpha)} & \text{if } \alpha > 1 \text{ and } 2s^{1-\alpha} > T^{1-\alpha} \end{cases}$$

Thus for $\alpha = 1$, we have $t \to 0$ has $T \to \infty$, and for $\alpha > 1$, $t \to 2^{1/(1-\alpha)}s$ as $T \to \infty$, even when social welfare is held fixed. Alpha fairness judges an egalitarian solution to be no better than a solution in which one party has arbitrarily more wealth than the other. This anomaly does not arise when $0 \leq \alpha < 1$.

7.2. Kalai-Smorodinsky bargaining

The Kalai-Smorodinsky (K–S) bargaining solution provides parties the largest possible fraction of their "ideal" utility, subject to the condition that the fraction is the same for all parties (Kalai and Smorodinsky 1975). A party's ideal utility is the maximum feasible utility that party could receive if the utilities of the other parties were ignored. Increases in utility are measured with respect to the default utility allocation.

One motivation for the K–S criterion is that it maximizes total utility while maintaining fairness for all players, where fairness takes into account the fact that allocating utility to some players is more costly than to others. This perspective can be suitable in bargaining contexts, as when labor and management negotiate wages (Alexander 1992). They may see a solution as fair when the two parties make the same relative concession. A technical motivation for the criterion is that it has a monotonicity property that the Nash solution lacks: when the feasible set is enlarged, the negotiated utilities of the players never decrease. This property is not necessarily desirable, as when enlargement allows one player to enjoy much greater utility at a small cost to other players. In any event, the K–S bargaining solution is defended by Thompson (1994) and is arguably consistent with the contractarian ethical philosophy developed by Gauthier (1983).

Mathematically, the objective is to find the largest scalar β such that $\boldsymbol{u} = (1-\beta)\boldsymbol{d} + \beta\boldsymbol{u}^{\max}$ is a feasible utility vector, where each $u_i^{\max}$ is the maximum of u_i over all feasible utility vectors $\boldsymbol{u}$. The bargaining solution is the vector $\boldsymbol{u}$ that maximizes β. Figure 1(b) illustrates the idea for two players when the default position $\boldsymbol{d}$ is the origin. The K–S solution (black dot) is the highest point at which the diagonal line intersects the feasible set.

Formally, the SWF for K–S bargaining might be defined

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

where $u_i^{\max} = \max_{\boldsymbol{x},\boldsymbol{u}}\{u_i \bigm| (\boldsymbol{u},\boldsymbol{x}) \in S\}$ for each i. The SWF clearly violates the Pigou–Dalton condition, because (supposing $\boldsymbol{d} = \boldsymbol{0}$) it regards any utility distribution $u_1, \ldots, u_n$ with ratios different from $u_1^{\max}, \ldots, u_m^{\max}$ as less socially desirable. For example, if we have a 2-person utility distribution $(u_1, u_2) = (\beta u_1^{\max}, \beta u_2^{\max})$ with $u_1^{\max} \neq u_2^{\max}$ for some β with $0 < \beta \le 1$, then a utility transfer that tends to equalize the distribution reduces social welfare. The Chaueauneuf–Moyes condition is violated for the same reason. These facts should be carefully considered before the K–S solution is used in applications. On the other hand, the optimization problem for the K–S criterion is straightforward:

$$\max_{\beta,\boldsymbol{x},\boldsymbol{u}} \left\{\beta \bigm| \boldsymbol{u} = (1-\beta)\boldsymbol{d} + \beta \boldsymbol{u}^{\max},\ (\boldsymbol{u},\boldsymbol{x}) \in S,\ \beta \le 1\right\}$$

Axiomatic justifications are given for the K–S solution by Kalai and Smorodinsky as well as by Thompson, but they again rely on a strong axiom of cardinal noncomparability. A bargaining justification might be given by arguing that it is rational for each player to minimize relative concession, and repeated rounds of bargaining will lead under suitable conditions to an equilibrium in which their relative concessions are equal and minimized.

On the other hand, the K–S scheme may allocate far more utility to an individual whose welfare is easily improved than to one who is less fortunate. For example, it may allocate treatment resources to persons suffering from the common cold to provide them the same fraction of their maximum health potential as patients with chronic kidney failure. The K–S model offers no means to prevent this kind of outcome by adjusting the trade-off between equity and efficiency, as is possible with alpha fairness.

More generally, one can ask why the potential utility that fortune or fate has granted to some individuals should necessarily be relevant to a fair allocation. Perhaps fairness sometimes demands a contrasting approach: rather than rewarding fortunate individuals strictly in proportion to their potential, we should give greater emphasis to improving the lot of those in less fortunate circumstances (Dworkin 1981a, 1981b, 2000; Barry 1988).

8. Threshold Criteria with Maximin Fairness

Williams and Cookson (2000) suggest two ways to combine utilitarian and maximin objectives using threshold criteria. One, based on a utility threshold, begins with a maximin criterion but switches to a utilitarian criterion when the overall utility cost of fairness becomes too great. The other, based on an equity threshold, begins with utilitarianism and switches to a maximin criterion when inequity becomes too great.

These proposals were originally defined only for two persons, and it is not obvious how to extend them to multiple parties. Hooker and Williams (2013) provide an n-person extension for the utility-threshold criterion, formulate it as a mixed integer programming problem, study its polyhedral properties, and apply it to a healthcare provision problem. After summarizing this work, we suggest an n-person extension of the equity-threshold criterion. It is more straightforward to formulate and can, in fact, yield a linear programming model.

An advantage of the threshold criteria is that they regulate the equity-efficiency trade-off with a parameter Δ that has a practical meaning in the n-person models. When a utility threshold is used, parties with utility within Δ of the worst-off are regarded as disadvantaged and deserving of special priority. When an equity threshold is used, parties whose utility is already more than Δ above the lowest are not regarded as deserving greater utility if the other utilities remain unchanged.

8.1. Utility-threshold criterion

The 2-person utility-threshold model of Williams and Cookson uses a maximin criterion when the two utilities are sufficiently close to each other, specifically $|u_1 - u_2| \leq \Delta$, and otherwise it uses a utilitarian criterion. This is illustrated in Fig. 2, 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

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

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 $W(\boldsymbol{u})$ to n parties as follows. The utility u_i of party 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 party whose utility is in the fair region is considered sufficiently disadvantaged to deserve priority. The generalized SWF $W(\boldsymbol{u})$ counts all utilities in the fair region

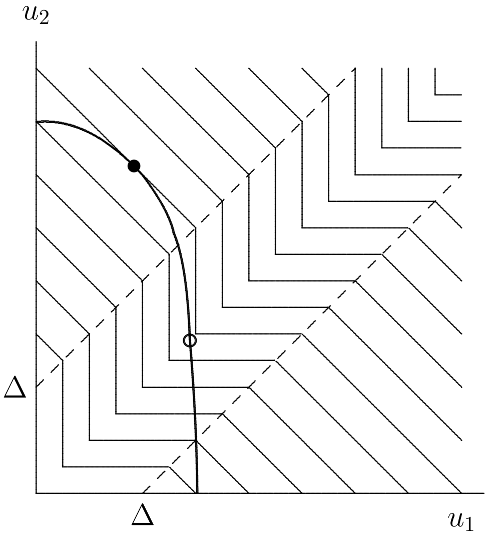


Figure 2: Contours for the utility-threshold SWF.

as equal to $u_{\min}$, so that they are treated in solidarity with the worst-off, and all other utilities as themselves. Copies of Δ are added to the SWF to ensure continuity of $W(\boldsymbol{u})$.

$$W(\boldsymbol{u}) = (n-1)\Delta + \sum_{i=1}^{n} \max\left\{u_i - \Delta, u_{min}\right\} \tag{8}$$

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.

Hooker and Williams extend $W(\boldsymbol{u})$ to problems in which utility is distributed to groups of different sizes, where each member of the group receives the same utility. This is useful when allocating resources to geographic regions, demographic groups, organizations, and so forth. 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 $W^g(\boldsymbol{u})$ considers a group i to be in the fair region when its per capita u_i is within Δ of u_{min}.

$$W^g(\boldsymbol{u}) = \Big(\sum_i s_i - 1\Big)\Delta + \sum_i s_i \max\left\{u_i - \Delta, u_{min}\right\} \tag{9}$$

Tractable MIP models are formulated for maximizing $W(\boldsymbol{u})$ and $W^g(\boldsymbol{u})$ subject to auxiliary constraints $u_i - u_j \leq M$ required for MIP representability.

The model for maximizing $W^g(\boldsymbol{u})$ is

$$\max_{\boldsymbol{x},\boldsymbol{u},\boldsymbol{\delta},\boldsymbol{v},w,z} \left\{ \Big(\sum_i s_i\Big)\Delta + \sum_i s_i v_i \;\middle|\; \begin{array}{c} u_i - \Delta \le v_i \le u_i - \Delta\delta_i, \text{ all } i \\ w \le v_i \le w + (M-\Delta)\delta_i, \text{ all } i \\ u_i - u_i \le M, \text{ all } i,j \\ u_i \ge 0, \; \delta_i \in \{0,1\}, \text{all } i \\ (\boldsymbol{u},\boldsymbol{x}) \in S \end{array} \right\} \tag{10}$$

The model for individuals is obtained by setting $s_i = 1$ for all i. This is an MILP model when the constraints $(\boldsymbol{u},\boldsymbol{x}) \in S$ are linear. Hooker and Williams prove that this representation of $W^g(\boldsymbol{u})$ is sharp (i.e., its continuous relation describes the convex hull of the feasible set) and is therefore the tightest possible linear model. Sharpness may, of course, be lost when the constraints $(\boldsymbol{u},\boldsymbol{x}) \in S$ are added. The practicality of the model was verified with experiments on a healthcare resource allocation instance of realistic size.

Gerdessen et al. (2018) make several observations regarding properties of the SWF (8). In particular, the solutions obtained by varying Δ need not all lie on the Pareto frontier defined by the convex combination (6) of utilitarian and maximin objectives. This is in fact to be expected, because the convex combination balances total utility with only the welfare of the worst-off party, while (8) takes into account how many parties are disadvantaged (i.e, in the fair region).

A weakness of the utility-threshold criteria (8) and (9) is that the actual utility levels of the disadvantaged parties, other than the very worst-off, have no effect on the value of the SWF. This is illustrated in the 3-person example of Fig. 3, which shows the contours of $W(u_1, u_2, u_3)$ with $\Delta = 3$ and u_1 fixed to zero. The SWF is constant in the shaded region, meaning that the utilities allocated to persons 2 and 3 have no effect on social welfare as measured by $W(\boldsymbol{u})$, so long as they remain in the fair region. As a result, many solutions that deliver the same social welfare differ greatly with respect to equity. This problem is addressed in Section 9 by combining a utilitarian with a leximax criterion.

It is stated in Karsu and Morton (2015) that $W(\boldsymbol{u})$ satisfies the Pigou–Dalton condition, but this is true only for $n = 2$. Figure 3 provides a counterexample for $n = 3$. The move from point A to point B represents a utility transfer from a better-off party to a worse-off party but strictly reduces social welfare. Yet $W(\boldsymbol{u})$ satisfies the slightly weaker C–M condition.

Theorem 2. The utility-threshold social welfare function $W(\boldsymbol{u})$ satisfies the Chateauneuf–Moyes condition.

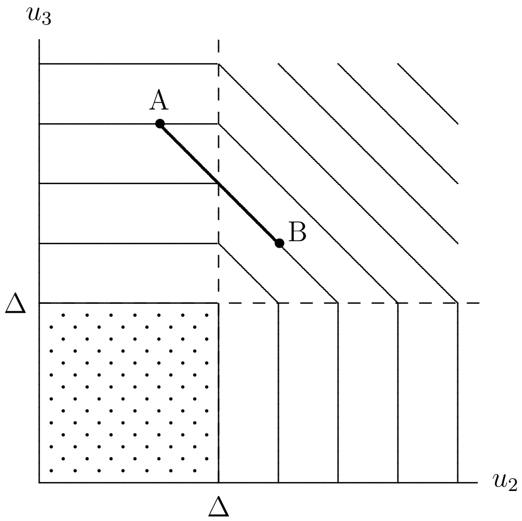


Figure 3: Contours of the utility-threshold SWF $W(0, u_2, u_3)$. The function is constant in the shaded region.

The utility-threshold criterion also escapes an anomaly that, as noted earlier, characterizes alpha fairness. It cannot assign equality the same social value as arbitrarily extreme inequality. In a 2-person context, for example, an egalitarian distribution $\boldsymbol{u} = (s, s)$ can have the same social value as a distribution in which one party has no utility and the other $\Delta + 2s$, but the gap can be no greater than this.

8.2. Equity-threshold criterion

Williams and Cookson define the 2-person equity-threshold SWF to be utilitarian when $|u_1 - u_2| \leq \Delta$ and otherwise maximin. In Fig. 4, the utilitarian solution (open dot) is unfair to person 1, and the welfare-maximizing solution is more egalitarian (black dot).

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

We generalize this SWF to n parties in a manner similar to the Hooker–Williams approach. The main difference is that we now say a utility u_i belongs to the fair region if $u_i - u_{min} \geq \Delta$, otherwise it is in the utilitarian region. Yet we continue to count utilities in the fair region as equal to u_{min} and those in the utilitarian region utilities as themselves. This yields the

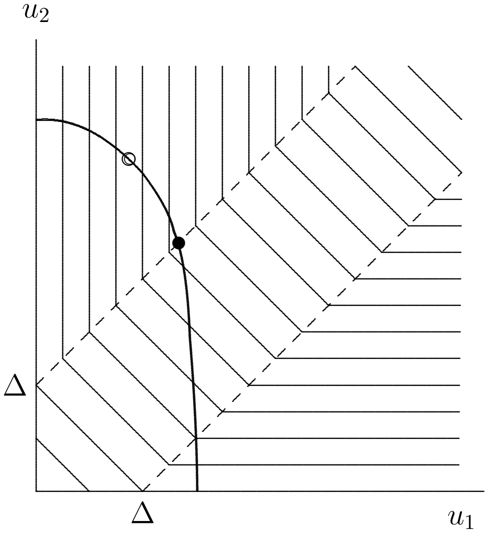


Figure 4: Contours for the equity-threshold SWF.

SWFs

$$W(\boldsymbol{u}) = n\Delta + \sum_{i=1}^{n} \min\{u_i - \Delta, u_{min}\} \tag{11}$$

$$W^g(\boldsymbol{u}) = \Big(\sum_{i=1}^{n} s_i\Big)\Delta + \sum_{i=1}^{n} s_i \min\{u_i - \Delta, u_{min}\} \tag{12}$$

As before, $W^g(\boldsymbol{u})$ is designed for distribution over groups.

These SWFs have two main effects. One is that a utilitarian criterion is applied to everyone whose utility is within Δ of the lowest. The other is that increasing a utility that is already more than Δ greater than the lowest adds nothing to social welfare if the other utilities remain unchanged. Like the utility-threshold criterion, the equity-threshold criterion can equate solutions that have very different equity characteristics. This is illustrated in Fig. 5, where all the solutions in the shaded region have the same social welfare.

While the utility-threshold SWF satisfies only the Chateauneuf–Moyes condition, we have the following:

Theorem 3. The equity-threshold SWF $W(\boldsymbol{u})$ satisfies the Pigou–Dalton condition and therefore the Chateauneuf–Moyes condition.

Unlike the utility-threshold SWF, the equity-threshold SWF has a simple

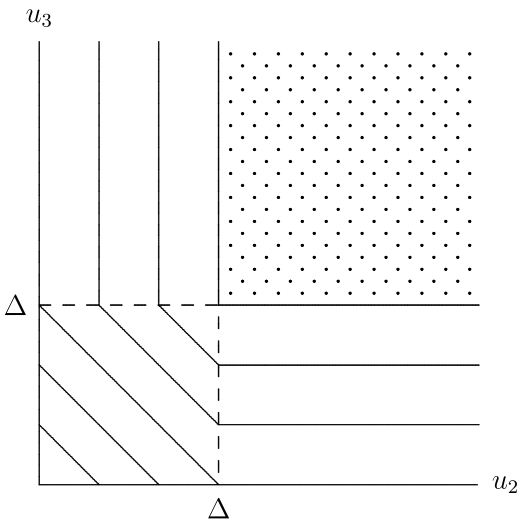


Figure 5: Contours of the equity-threshold SWF $W(0, u_2, u_3)$. The function is constant in the shaded region.

linear model.

$$\max_{\boldsymbol{x},\boldsymbol{u},\boldsymbol{v},w,z} \left\{ n\Delta + \sum_i v_i \;\middle|\; \begin{array}{c} v_i \leq w \leq u_i, \text{ all } i \\ v_i \leq u_i - \Delta, \text{ all } i \\ w \geq 0, v_i \geq 0, \text{ all } i \\ (\boldsymbol{u}, \boldsymbol{x}) \in S \end{array} \right\}$$

The formulation for the group SWF $W^g(\boldsymbol{u})$ is the same, except that the objective function is

$$\Big(\sum_i s_i\Big)\Delta + \sum_i s_i v_i$$

These formulations are LP problems when linear constraints define S.

9. Threshold Criteria with Leximax Fairness

As pointed out in the previous section, threshold-based combinations that rely on maximin fairness are sensitive to the utility level of only the very worst-off party. The resulting SWFs equate distributions that can differ substantially in their equity characteristics. This tends to become a problem in practice when the constraint set severely restricts the maximum utility of some individual. The solution will almost certainly assign this person the maximum utility, regardless of what the rest of the problem is like. The equity situation of other disadvantaged parties become irrelevant, so long as their utilities are within Δ of the lowest. As a result, equity plays almost no role in the solution. This situation can be addressed to a great degree by replacing maximin fairness with leximax fairness. We consider two proposals for doing so, both extensions of the Hooker–Williams approach. One assumes that utility recipients can be ranked by priority in advance. The other makes no such assumption and obtains a socially optimal distribution by maximizing a sequence of SWFs, each of which combines utility and a maximin criterion.

9.1. Predetermined preference order

McElfresh and Dickerson (2018) propose a method for combining utilitarian and leximax criteria in the context of kidney exchange. It relies on the assumption that the parties can be given a preference ordering in advance. It first maximizes a SWF that combines utilitarian and maximin criteria in a way that treats the most-preferred party as the worst-off. If all optimal solutions of this problem lie in the utilitarian region, a utilitarian criterion is used to select one of the optimal solutions. (Here, a utility vector $\boldsymbol{u}$ is said to be in the fair region if $\max_i\{u_i\} - \min_i\{u_i\} \leq \Delta$, and otherwise in the utilitarian region.) Otherwise a leximax criterion is used for all of the optimal solutions, subject to the preference ordering (i.e., maximize u_1 first, then u_2 etc.). If we index the parties in order of decreasing preference, the SWF is

$$W(\boldsymbol{u}) = \left\{ \begin{array}{ll} nu_1, & \text{if } |u_i - u_j| \leq \Delta \text{ for all } i, j \\ \displaystyle\sum_i u_i + \text{sgn}(u_1 - u_i)\Delta, & \text{otherwise} \end{array} \right. \tag{13}$$

McElfresh and Dickerson state that $W(\boldsymbol{u})$ has continuous contours, but this is true only for $n = 2$. For a counterexample with $n = 3$, we note that $W(0, 0, \Delta + \epsilon) = \epsilon$ and $W(0, \epsilon, \Delta + \epsilon) = 2\epsilon - \Delta$ for arbitrarily small $\epsilon > 0$. The discontinuity of the SWF raises questions regarding its suitability for application, since a slight change in the utility distribution could bring about a large and unexpected change in the measurement of social welfare.

While $W(\boldsymbol{u})$ satisfies the Pigou–Dalton condition for $n = 2$ (if one considers only utility transfers from u_2 to u_1), it violates both the P–G and Chateauneuf–Moyes conditions when $n = 3$. For example, a C–M transfer that converts (u_1, u_2, u_3) from $(\epsilon, 0, \Delta + \epsilon)$ to $(\epsilon, \epsilon, \Delta)$ reduces social welfare from $2\epsilon + \Delta$ to ϵ.

McElfresh and Dickerson maximize $W(\boldsymbol{u})$ using an algorithm that is specialized to the kidney exchange problem, but we can state a general mixed integer model.

$$\max_{\substack{\boldsymbol{u},\boldsymbol{x}\\ w_1,w_2\\ \boldsymbol{y},\phi,\boldsymbol{\delta}}} \left\{ w_1 + w_2 \;\middle|\; \begin{array}{l} w_1 \le nu_1,\ w_1 \le M\phi \\ w_2 \le \sum_i (u_i + y_i),\ w_2 \le M(1-\phi) \\ u_i - u_j - \Delta \le M(1-\phi),\ \text{all } i,j \\ y_i \le \Delta,\ y_i \le -\Delta + M\delta_i,\ u_i - u_1 \le M(1-\delta_i),\ \text{all } i \\ (\boldsymbol{u},\boldsymbol{x}) \in S;\ \phi,\delta_i \in \{0,1\},\ \text{all } i \end{array} \right\}$$

Two additional issues should be considered. One is the need for preassigned priorities. While it is possible to specify in advance a preference ranking of parties in some applications, such as the kidney exchange problem, this is not possible in many applications. Also the leximax criterion is not used until optimal solutions of the SWF are already obtained, and then applied only to the optimal solutions. It may be preferable to use a leximax criterion when considering all feasible distributions, rather than those that are already optimal in some sense.

9.2. A sequence of social welfare functions

Chen and Hooker (2020a, 2020b) avoid assuming a pre-determined preference ordering of recipients by maximizing a sequence of social welfare functions $W_1(\boldsymbol{u}), \ldots, W_n(\boldsymbol{u})$. The SWFs successively give priority to the worst-off recipient, the second worst-off, and so forth, while in each case considering the impact on total utility by means of a threshold criterion. The first function $W_1(\boldsymbol{u})$ is identical to the Hooker–Williams function in (11), and the remainder are defined as follows:

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

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 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. Very similar SWFs are given for groups of individuals. It is shown that all of these SWFs are continuous.

A socially optimal distribution is found by first solving a problem P_1 given by

$$\max_{\boldsymbol{u},\boldsymbol{x}} \left\{ W_1(\boldsymbol{u},\boldsymbol{x}) \;\middle|\; |u_i - u_j| \leq M, \text{ all } i,j;\ (\boldsymbol{u},\boldsymbol{x}) \in S \right\} \tag{14}$$

and then solving problems P_k given by

$$\max_{\boldsymbol{u},\boldsymbol{x}} \left\{ W_k(\boldsymbol{u},\boldsymbol{x}) \;\middle|\; \begin{array}{c} u_{i_j} = \bar{u}_{i_j},\ j = 1,\ldots,k-1 \\ u_i \geq \bar{u}_{i_{k-1}},\ u_i - \bar{u}_{i_1} \leq M,\ i \in I_k \\ (\boldsymbol{u},\boldsymbol{x}) \in S \end{array} \right\} \tag{15}$$

The indices i_j are defined so that u_{i_j} is the utility determined by solving P_j. In particular, u_{i_j} is the utility with the smallest value among the unfixed utilities in an optimal solution obtained by solving P_j. Thus

$$i_j = \operatorname*{arg\,min}_{i \in I_j} \left\{ u_i^{[j]} \right\}$$

where $\boldsymbol{u}^{[j]}$ is an optimal solution of P_j and $I_j = \{1,\ldots,n\} \setminus \{i_1,\ldots,i_{j-1}\}$. We need only solve P_k for $k = 1,\ldots,K+1$, where K is the largest k for which $\bar{u}_{i_k} \leq \bar{u}_{i_1} + \Delta$. The solution of the social welfare problem is then

$$u_i = \begin{cases} \bar{u}_i & \text{for } i = i_1,\ldots,i_{K-1} \\ u_i^{[K]} & \text{for } i \in I_K \end{cases}$$

The MILP model for solving P_1 with groups is (10). Using notation similar to that for the goal programming model (5), the MILP formulation for solving P_k with groups, $k \geq 2$, is

$$\max_{\substack{\boldsymbol{x},\boldsymbol{u},\boldsymbol{\delta},\boldsymbol{\epsilon} \\ \boldsymbol{v},w,\tau,z}} \left\{ z \;\middle|\; \begin{array}{c} z \leq \big(\sum_{i \in I_k} s_i - 1\big)\tau + \sum_{i \in I_k} s_i v_i \\ 0 \leq v_i \leq M\delta_i,\ i \in I_k \\ v_i \leq u_i - \hat{u}_{i_1} - \Delta + M(1-\delta_i),\ i \in I_k \\ \tau \leq \hat{u}_{i_1} + \Delta,\ \tau \leq w,\ w \geq \hat{u}_{i_1} \\ w \leq u_i \leq w + M(1-\epsilon_i),\ i \in I_k \\ u_i - \hat{u}_{i_1} \leq M,\ i \in I_k \\ \sum_{i \in I_k} \epsilon_i = 1;\ \delta_i, \epsilon_i \in \{0,1\}, i \in I_k \\ (\boldsymbol{u},\boldsymbol{x}) \in S \end{array} \right\} \tag{16}$$

While this is not a sharp model in general for $k \geq 2$, Chen and Hooker identify valid inequalities that can strengthen the linear relaxation of P_k:

$$z_k \le \sum_{i \in I_k} s_i u_i \tag{17}$$

$$z_k \le \Big(\sum_{j \in I_k} s_i \Big) u_j + \beta \sum_{j \in I_k \setminus \{i\}} s_j (u_j - \bar{u}_{i_{k-1}}), \;\; i \in I_k \tag{18}$$

where

$$\beta = \frac{M - \Delta}{M - (\bar{u}_{i_{k-1}} - \bar{u}_{i_1})} = \Big(1 - \frac{\Delta}{M}\Big)\Big(1 - \frac{\bar{u}_{i_{k-1}} - \bar{u}_{i_1}}{M}\Big)^{-1}$$

Formulations (10) and (16) are used to solve healthcare resource and earthquake shelter location problems of realistic size in a matter of seconds for a given value Δ.

We have already seen that $W_1(\boldsymbol{u})$ can violate the Pigou-Dalton condition but satisfies the Chateauneuf–Moyes condition. The same is true for $W_k(\boldsymbol{u})$ for $k \ge 2$. These functions fail Pigou-Dalton in counterexamples similar to Fig. 7. Regarding the C–M condition, we have the following:

Theorem 4. The social welfare functions $W_k(\boldsymbol{u})$ satisfy the Chateauneuf–Moyes condition for $k = 2, \ldots, n$.

Although each $W_k(\boldsymbol{u})$ satisfies the C–M condition, we show that feasible sets can be contrived in which a C–M transfer may not transform a socially optimal solution to another socially optimal solution. For example, suppose $n = 4$, $\Delta = 5$, and the feasible set consists only of the three vectors on the left:

$$\begin{array}{ll} \boldsymbol{u}^1 = (1,2,8,9) & (24,15,27,35) \\ \boldsymbol{u}^2 = (2,3,7,8) & (24,18,32,39) \\ \boldsymbol{u}^3 = (1,2,3,12) & (25,16,22,28) \end{array}$$

The corresponding values $(W_1(\boldsymbol{u}), \ldots, W_4(\boldsymbol{u}))$ are shown on the right. Distribution $\boldsymbol{u}^2$ results from applying a C–M transfer to the socially optimal distribution $\boldsymbol{u}^1$, but $\boldsymbol{u}^2$ is not socially optimal because $u_1 = 2$ in no optimal solution of P_1. Rather, the unique optimal solution of P_1 is $\boldsymbol{u}^3$, in which $u_1 = 1$. This situation can occur when a socially optimal distribution $\boldsymbol{u}$ is not an optimal solution of P_1. In the example, $\boldsymbol{u}^1$ is not an optimal solution of P_1. It is unclear whether this should be seen as a weakness of the Chen–Hooker approach, or as evidence of the inherent complexity of balancing equity and efficiency.

10. Statistical Fairness Metrics

The mathematical formulation of equity has become a major issue in the field of machine learning, because machine learning algorithms are employed to make high-stake decisions and require precisely coded criteria for assessing whether those decisions are fair. As summarized in Mehrabi et al. (2019), fairness in machine learning seeks to eliminate "any prejudice or favoritism toward an individual or a group based on their inherent or acquired characteristics." One well-known example that motivates extensive interest in machine learning fairness is the series of research efforts on whether the COMPAS software, supported by a recidivism risk prediction algorithm, is biased against African-Americans (Angwin et al. 2016, Dieterich et al. 2016, Chouldechova 2017). The focus of fair machine learning has been almost entirely on mitigating this kind of bias and ensuring that certain minority groups, often defined by law, receive fair treatment. The AI community has seized upon traditional statistical measures of classification error to detect bias, so that it can be avoided when possible.

In a typical scenario, the machine makes yes–no decisions as to who receives a certain benefit, such as a mortgage loan, a job interview, parole, and so forth. A large training set is used to train the machine to select appropriate individuals as reliably as possible, based on various features they possess. The aim is to predict who will pay their mortgage, become a valued employee, or avoid future crimes. These tasks conventionally use supervised learning methods to train predictive models from labelled data. In particular, majority of the literature on fair machine learning studies fairness in classification, and we focus on this setup as well.

A fairness test compares decisions for a minority or protected group with those for the remainder of the population. Four outcomes are possible for each individual: a true positive (the machine correctly selects the individual for a benefit), a false positive (it incorrectly selects), a true negative (it correctly rejects), and a false negative (it incorrectly rejects). We will refer to the number of individuals in these four groups, respectively, as TP, FP, TN, and FN. Various metrics involving these four statistics are compared between the minority group and the rest of the population, each yielding a measure of parity between the groups.

We will set $a_i = 1$ when individual i should be selected, and $a_i = 0$ otherwise. We let N be an index set for individuals in the protected group, and N' for those in the remainder of the population. Rather than a vector $\boldsymbol{u}$ of utilities distributed across individuals, we have a vector $\boldsymbol{\delta} = (\delta_1, \ldots, \boldsymbol{\delta}_n)$ of individual 0–1 decisions, where $\delta_i = 1$ indicates that individual i is selected.

We can view social welfare as a function $W(\boldsymbol{\delta})$ of these decisions rather than a function $W(\boldsymbol{u})$ of utilities. Of course, one could view $\boldsymbol{\delta}$ as a simplified representation of utilities in which each individual receives utility 0 or 1. Typically, bounds are put on $W(\boldsymbol{\delta})$ rather than maximizing $W(\boldsymbol{\delta})$, thus leading to an optimization problem (3) that maximizes some other objective subject to these bounds.

Unfortunately, it is far from clear how the fairness of a decision vector $\boldsymbol{\delta}$ should be measured. There are a wide variety of classification error metrics, many of which are pairwise incompatible, with no consensus on which is most suitable for any given application (e.g. Kleinberg et al. 2016, Friedler et al. 2016). In addition, the focus on classification error affords a rather narrow perspective on the fairness problem, because the underlying concern is generally distributive justice in a broader sense. Discrimination against a minority group is normally seen as undesirable because it results in an unjust distribution of utilities. Finally, there is no obvious criterion for which groups should be designated as protected, unless one is content to recognize only those sanctioned by law.

The AI community might well consider the option of training machines to maximize a more comprehensive measure of social welfare, such as one of those discussed in previous sections, to better align fairness concepts with social well-being. We are already beginning to see some movement in this direction (Heidari et al. 2018, Corbett-Davies and Goel 2018, Heidari et al. 2019, Hu and Chen 2020). The classification vector $\boldsymbol{\delta}$ can be viewed as a set of decision variables on which utilities depend, perhaps as given by a utility function $\boldsymbol{u} = \boldsymbol{U}(\boldsymbol{\delta})$, and social welfare assessed by a function $W(\boldsymbol{u})$ as in model (2). In the simplest case, one could set $U_i(\delta_i) = c_i\delta_i + d_i$, where $c_i + d_i$ is the utility experienced by individual i if selected, and d_i if not selected. Of course, legal requirements may dictate that bounds are placed directly on one of the parity measures.

In any event, the discussion below is restricted to fairness metrics $W(\boldsymbol{\delta})$ defined directly in terms of the decision vector $\boldsymbol{\delta}$. We consider four of the best known metrics: demographic parity, equalized odds, accuracy parity, and predictive rate party. For brevity, we refer to individuals in the protected group as minority individuals, and those in the remainder of the population as majority individuals. We do not discuss the Pigou-Dalton and Chateauneuf–Moyes conditions in this context, because they do not appear to be meaningful for 0–1 decision vectors.

We also omit some of the fairness metrics that have been proposed for machine learning. Most of them are surveyed in Verma and Rubin (2018) and are similar to those discussed here. Beyond these, the Matthews correlation coefficient (Matthews 1975, Chicco and Jurman 2020) is often regarded as the most comprehensive measure of classification accuracy, but it corresponds to a complicated, nonconvex SWF that could be quite difficult to optimize. Counterfactual fairness (Kusner et al. 2017, Russell et al. 2017) aims to select minority applicants with the same probability that would apply if they had been majority applicants. For example, financial irresponsibility of a mortgage applicant, which cannot be directly observed, may correlate with residence in a low-income neighborhood. This may lead to bias against minority applicants whose residence in the neighborhood is due to social conditions that have nothing to do with financial irresponsibility. Counterfactual fairness strives to avoid this confounding of factors by constructing a causal network and using Bayesian inference to isolate the effect of financial responsibility (Pearl 2000, Pearl et al. 2016). It is unclear at this point how to incorporate this scheme into an optimization model. Beyond fairness in classification and supervised learning, recent research has also seen progress on fairness in unsupervised learning (e.g., Abraham et al. 2019, Deepak and Abraham 2020) and reinforcement learning (e.g., Weng 2019, Siddique et al. 2020). These machine learning frameworks are generally difficult to interpret as optimization models and tend to require customized fairness definitions.

10.1. Demographic parity

The simplest bias metric is based on *demographic parity*, also known as proportional/statistical parity. It is achieved when the fraction of minority individuals selected is the same as the fraction of majority individuals selected. It is defined by comparing the ratio

$$(\mathrm{TP}+\mathrm{FP})/(\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN})$$

across the two groups. The associated social welfare function is $W(\boldsymbol{\delta}) = 1 - |B(\boldsymbol{\delta})|$, where

$$B(\boldsymbol{\delta}) = \frac{1}{|N|}\sum_{i\in N}\delta_i - \frac{1}{|N'|}\sum_{i\in N'}\delta_i$$

Thus $0 \leq W(\boldsymbol{\delta}) \leq 1$, and complete parity is obtained when $W(\boldsymbol{\delta}) = 1$. This SWF is easily linearized and therefore gives rise to an MILP problem when the problem constraints are linear:

$$\min_{\boldsymbol{\delta},\boldsymbol{x},w}\left\{ w \;\middle|\; -w \leq B(\boldsymbol{\delta}) \leq w,\ (\boldsymbol{\delta},\boldsymbol{x}) \in S,\ \boldsymbol{\delta} \in \{0,1\}^n \right\} \tag{19}$$

Since Dwork et al. (2012) proposed the use of demographic parity for fairness in classification, it has been widely studied and applied. Existing classification algorithms seeking demographic parity guarantees almost never impose the criterion exactly via the formulation in (19) due to the integer variables $\boldsymbol{\delta}$. Instead, one strategy is to use continuous relaxations of the exact definition $B(\boldsymbol{\delta})$. For instance, Zafar et al. (2017) define a convex proxy for demographic parity by replacing the discrete variables $\boldsymbol{\delta}$ with the continuous decision boundaries of the trained model, and Olfat and Aswani (2018) substitute the decision boundaries with covariance matrices to formulate a stronger but non-convex proxy of demographic parity. Another strategy is to treat a given classification algorithm as a black box and design separate pre-processing or post-processing schemes to attain fairness guarantees. As an example, Agarwal et al. (2018) develop a systematic approach that reduces fair classification to a sequence of cost-sensitive classification, and derive theoretical guarantees on the generated classifier for a variety of fairness measures including demographic parity, equalized odds and accuracy parity.

Despite its popularity, critics of demographic parity view the measure as unsuitable for most practical purposes because it requires strict equality of outcomes. For example, it discriminates against a minority group that happens to be to be more qualified for loans than the majority on the average. It requires that a minority individual receive a loan with no greater probability than a majority individual.

10.2. Equalized odds

The *equalized odds* metric is based on two related but distinct criteria. One is that the fraction of *qualified* minority persons selected is the same as the fraction of qualified majority persons selected (Hardt et al. 2016). The other is that the fraction of *unqualified* minority persons selected is the same as the fraction of unqualified majority persons selected (Zafar et al. 2017). The former is also known as *equality of opportunity* and is defined by comparing the ratio TP/(TP + FN). It has the SWF $W(\boldsymbol{\delta}) = 1 - |B(\boldsymbol{\delta})|$ across the two groups, where

$$B(\boldsymbol{\delta}) = \frac{\sum_{i\in N} a_i\delta_i}{\sum_{i\in N} a_i} - \frac{\sum_{i\in N'} a_i\delta_i}{\sum_{i\in N'} a_i} \tag{20}$$

The latter criterion is based on the ratio FP/(FP + TN) and again has the SWF $W(\boldsymbol{\delta}) = 1 - |B(\boldsymbol{\delta})|$, but with

$$B(\boldsymbol{\delta}) = \frac{\sum_{i\in N}(1-a_i)\delta_i}{\sum_{i\in N}(1-a_i)} - \frac{\sum_{i\in N'}(1-a_i)\delta_i}{\sum_{i\in N'}(1-a_i)} \tag{21}$$

Both are easily linearized and give rise to the optimization problem (19). Similar to the case of demographic parity, these exact formulations are rarely used to train classification models. Hardt et al. (2016) design post-processing schemes to adjust the outcomes of unfair classifiers to attain equalized odds guarantees. Zafar et al. (2017) study an in-processing perspective and propose tractable proxies for (20) and (21) by replacing $\boldsymbol{\delta}$ with continuous approximations.

10.3. Accuracy parity

The two-sided evaluation in equalized odds can be obviated simply by measuring the fraction of predictions that are accurate, which is the ratio

$$(\mathrm{TP} + \mathrm{TN})/(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})$$

The SWF is $W(\boldsymbol{u}) = 1 - |B(\boldsymbol{\delta})|$, where

$$B(\boldsymbol{\delta}) = \frac{1}{|N|} \sum_{i \in N} \big(a_i \delta_i + (1 - a_i)(1 - \delta_i)\big) - \frac{1}{|N'|} \sum_{i \in N'} \big(a_i \delta_i + (1 - a_i)(1 - \delta_i)\big)$$

The optimization problem is again (19). Accuracy parity is less studied than the previous two measures, perhaps because it does not distinguish between true positives and true negatives. It is less often used in the design of fair classifiers than as a tool to evaluate existing classifiers. For example, Berk et al. (2018) list accuracy parity as one of the meaningful fairness definitions in criminal justice risk assessment.

10.4. Predictive rate parity

When one wishes to compare what fraction individuals selected from each group should have been selected, the relevant measure is *predictive rate parity*, defined as TP/(TP + FP). The SWF is $W(\boldsymbol{\delta}) = |1 - B(\boldsymbol{\delta})|$, with

$$B(\boldsymbol{\delta}) = \frac{\sum_{i \in N} a_i \delta_i}{\sum_{1 \in N} \delta_i} - \frac{\sum_{i \in N'} a_i \delta_i}{\sum_{1 \in N'} \delta_i}$$

The optimization model is again (19), but it poses a difficult optimization problem because variables occur in the denominator. A change of variables similar to that in linear–fractional programming is unhelpful for two reasons. One is that the two ratios in $B(\boldsymbol{\delta})$ give rise to two scaling factors t, t' that create a nonconvex bilinear term tt' even in a linear constraint set. The other is that rescaling destroys the integrality of the 0–1 variables δ_i. We therefore appear to have an irreducibly difficult problem in nonlinear integer programming.

Predictive parity is primarily considered in risk assessment contexts, such as, recidivism prediction (Dieterich et al. 2016, Chouldechova 2017), child maltreatment screening (Chouldechova et al. 2018). In any case, it is unclear why predictive rate parity would be preferable in a given application than one of the measures discussed above. Accuracy parity, for example, would seem to be at least as suitable, and it creates no computational difficulties.

11. General Guidelines

There is no one best approach to formulating equity and fairness in an optimization model. Fairness is a collection of concepts, many of them rather vague, that can be found in popular culture, academic literature, and legal settings. Nonetheless, the various formulations surveyed here have characteristics that may be more or less suitable for the type of fairness one wishes to achieve in a given context. We conclude with an overview of these characteristics to assist one in exploring the equity landscape. We encourage the reader to consult the more detailed discussion provided earlier, and perhaps cited literature, before settling on a choice of model for a particular application.

Inequality metrics (Section 4) are of limited applicability because they take no account of absolute welfare levels. Even if relative welfare is all that matters, there may be an ethical difference between a distribution with extremes at the bottom end and one with extremes at the top end, and inequality measures do not distinguish these. Nonetheless, inequality measures can be appropriate if they truly represent the only criterion of interest. The *relative range* suits applications in which one simply wants to avoid extreme outliers. The *relative mean deviation* measures dispersion across the entire distribution. It is proportional to the *Hoover index*, which is the fraction of total utility that must be redistributed to achieve perfect equality. The *coefficient of variation* and *Gini coefficient* have the advantage that they are widely used, and there is a general appreciation of what they say about a distribution. All of these measures but the coefficient of variation have simple linear models. The nonlinearity of the latter seems an unwarranted complication, unless something about an application calls for this particular measure.

Fairness criteria can reflect concern for the disadvantaged as well as inequality (Section 5). A famous example is the Rawlsian difference principle, which gives rise to the *maximin* criterion. It is backed by a highly developed social contract argument that can have considerable intuitive appeal. However, the principle is intended only for the design of social institutions and can have surprising implications when applied to welfare distribution in general. For example, if improving the welfare of certain individuals is very expensive, perhaps due to incurable disease, the maximin principle can require a massive resource transfer that reduces everyone else to the same level of suffering. Limiting the transfer does not help, because it reduces utility even further and, worse, can allow some resources to go unused. The latter difficulty, but only it, can be remedied by extending the maximin to a *leximax* principle. A very different option is to use the *McLoone index*, a statistical criterion that measures the extent to which those in the lower half of the utility distribution are deprived. It appears in discussions of educational equality and other public policy matters.

Pure fairness measures can be appropriate when there is no need to balance fairness against the overall welfare of the population. However, practical situations frequently call for both equity and efficiency to be explicitly considered. One way to strive for both is simply to maximize a convex combination of the two (Section 6). Yet it is highly unclear how to adjust their relative weights, particularly when they are measured in different units. Efficiency is measured in units of utility, while most of the equity objectives discussed so far are dimensionless.

Alpha fairness and Kalai-Smorodinsky bargaining offer more principled solutions to the equity-efficiency trade-off (Section 7). The parameter α in *alpha fairness* regulates the trade-off on a scale that ranges from a purely utilitarian to a purely maximin criterion. Axiomatic and bargaining justifications have been offered for this SWF, particularly for $\alpha = 1$ (proportional fairness, or the Nash bargaining solution). However, these justifications are perhaps less relevant to practice than the mere fact that one can continuously adjust the trade-off to suit the occasion. Alpha fairness has, in fact, seen fairly wide employment in engineering, despite the nonlinearity of the SWF. Yet while α can be interpreted in terms of welfare-preserving utility transfers, it is still unobvious how to justify any particular choice for its value. Also, alpha fairness can assign the same social welfare to equality as to extreme inequality (when $\alpha \geq 1$), although this becomes a practical issue only for certain types of problem constraints.

The *Kalai-Smorodinsky* solution avoids this last issue entirely but poses another. It is suitable for bargaining situations when the parties concerned see equal relative concessions to be fair, as when buyer and seller negotiate a price, or labor and management negotiate wages. However, it may be unsuitable when some individuals have less utility potential due to physical impairment or some other factor beyond their control. In such cases, fairness may require special consideration for those who suffer misfortune, as in several other schemes considered here. Also K–S bargaining offers no parameter to adjust the equity-efficiency trade-off, and it violates both the Pigou-Dalton and Chateauneuf-Moyes conditions.

Threshold SWFs (Section 8) combine utilitarian and maximin criteria using a parameter Δ that is easier to interpret in practice than the α of alpha fairness. They also avoid the alpha fairness model's anomaly of sometimes regarding equality as ethically equivalent to extreme inequality. A *utility-threshold* model is suitable when equity is the initial concern, but one does not wish to pay too high a cost for fairness. This may occur, for example, in health-related or politically sensitive contexts. The parameter Δ is chosen so that disadvantaged parties whose utility is with Δ of the lowest are seen as deserving special priority. The SWF satisfies the C–M condition and has a mixed integer model that is readily solved in practice. An *equity-threshold* model is better suited for situations in which efficiency is the initial concern, but one does not want to create excessive inequality. This may be the situation in traffic management, telecommunications, or disaster recovery. In this context, the parameter Δ has a somewhat different meaning: it is chosen in such a way that one wishes to recognize no social benefit in improving the lot of well-off individuals whose utility is already more than Δ greater than the lowest, if the other utilities remain unchanged. The SWF satisfies both the P–G and C–M conditions and has an easily solved linear model.

Threshold models that combine efficiency with the maximin criterion inherit the tendency of the latter to ignore the actual utility levels of the disadvantaged other than the very worst-off. This may result in less sensitivity to equity than desired, particularly when the utility of some individuals is severely limited *a priori* by the constraint set. Two utility-threshold models address this issue by combining efficiency with a leximax rather a maximin criterion (Section 9). One assumes a *predefined preference ordering* for the parties, which may be suitable for some situations, such as organ transplants. However, the SWF is discontinuous, and it satisfies neither the P–G nor the C–M conditions. Another model makes no assumptions regarding preference, but it maximizes a *sequence of SWFs* to balance efficiency and leximax fairness. It uses the same parameter Δ as maximin-based utility-threshold model. The sequential SWFs are continuous, and they again satisfy the C–M condition and have mixed integer models that are readily solved in practice. This approach yielded markedly superior solutions, relative to a maximin-based threshold model, in a healthcare problem where the utility of some patients is severely limited by poor prognosis, and an earthquake shelter location problem in which the utility of some neighborhoods is severely limited by their remoteness from all of the potential shelter locations. At this writing, no *equity*-threshold models have been developed to combine efficiency with leximax fairness, although it appears that this could be done along similar lines.

Statistical bias measures (Section 10) are widely used in machine learning to judge whether a protected subpopulation, such as a minority group, is treated fairly. These measures do not attempt to take account of overall welfare, and they assess distributive justice in a rather restricted sense. Rather than evaluate a distribution of utilities across the population, they examine how yes and no decisions are distributed between the protected and control groups, as for example in the granting of mortgage loans, job interviews, school admissions, or parole. Many of these statistical metrics are pairwise incompatible, and there is no consensus as to which are appropriate for a given application. Indeed, most were originally developed to measure predictive accuracy, not fairness.

To take some examples, *demographic parity* compares the fraction of individuals accepted in the two groups. It is often too strict because it fails to recognize group differences in qualifications. *Equalized odds* compares the fraction of qualified (or unqualified) individuals accepted. *Accuracy parity* compares the fraction of individuals correctly classified (by acceptance or rejection). *Predictive rate parity* compares the fraction of selected individuals who are correctly selected. The computational tractability of minimizing bias varies widely. The first three SWFs mentioned here have easy linear models. The fourth poses an extremely difficult mixed integer/nonlinear programming problem, which hardly seems worth solving, because there is no clear reason for using this metric rather than another. The *Matthews correlation coefficient*, perhaps the most comprehensive bias measure, is even more challenging computationally. *Counterfactual fairness* is a very different concept based on causal networks, and its formulation as a social welfare maximization problem is currently a research issue.

The standard approach to fairness in machine learning is to maximize predictive accuracy subject to a constraint on bias. Yet this not only relies on a narrow conception of utility (by identifying it with predictive accuracy), but it provides no criterion for balancing utility and equity. Equity itself is assessed only with respect to a protected group than across an entire utility distribution. Indeed, it is unclear on what principle groups should be selected for protection, unless one is content to consider only those mandated by law. An alternative approach would be to maximize social welfare more broadly, rather than predictive accuracy, when training a neural network or building a rule base, perhaps using one of the SWFs surveyed here. The SWF could reflect the utilitarian benefits of accuracy as well as other utilitarian and equity-oriented factors.

References

[1] Abraham, S. S., Deepak P, and Sundaram, S. S. (2019). Fairness in clustering with multiple sensitive attributes. *arXiv preprint 1910.05113.*

[2] Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. (2018). A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69. PMLR.

[3] Alexander, C. (1992). The Kalai–Smorodinsky bargaining solution in wage negotiations. *Journal of the Operational Research Society*, 43:779–786.

[4] Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica.* 23 May.

[5] Atkinson, A. B. (1975). *The Economics of Inequality.* Clarendon Press.

[6] Barry, B. (1988). Equal opportunity and moral arbitrariness. In Bowie, N. E., editor, *Equal Opportunity*, pages 23–44. Westview Press.

[7] Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. (2018). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 1:42.

[8] Bertsimas, D., Farias, V., and Trichakis, N. (2012). On the efficiency-fairness trade-off. *Management Science*, 58:2234–2250.

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

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

[11] Chateauneuf, A. and Moyes, P. (2005). Measuring inequality without the Pigou-Dalton condition. WIDER Working Paper Series 2005/02, World Institute for Development Economic Research (UNU-WIDER).

[12] Chen, V. and Hooker, J. N. (2020a). Balancing fairness and efficiency in an optimization model. *ArXiv preprint 2006.05963*.

[13] Chen, V. and Hooker, J. N. (2020b). A just approach balancing Rawlsian leximax fairness and utilitarianism. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 221–227.

[14] Chicco, D. and Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21.

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

[16] Chouldechova, A., Benavides-Prado, D., Fialko, O., and Vaithianathan, R. (2018). A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*, pages 134–148. PMLR.

[17] Corbett-Davies, S. and Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint 1808.00023*.

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

[19] Cowell, F. A. and Kuga, K. (1981). Additivity and the entropy concept: An axiomatic approach to inequality measure. *Journal of Economic Theory*, 25:131–143.

[20] Dalton, H. (1920). The measurement of the inequality of incomes. *Economic Journal*, 30:348–461.

[21] Deepak, P. and Abraham, S. S. (2020). Representativity fairness in clustering. In *WebSci*, pages 202–211.

[22] Dieterich, W., Mendoza, C., and Brennan, T. (2016). COMPAS risk scales: Demonstrating accuracy equity and predictive parity. Technical report, Northpointe Inc. Research Department. 8 July.

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

[24] Dworkin, R. (1981a). What is equality? Part 1: Equality of resources. *Philosophy and Public Affairs*, 10:185–246.

[25] Dworkin, R. (1981b). What is equality? Part 2: Equality of welfare. *Philosophy and Public Affairs*, 10:283–345.

[26] Dworkin, R. (2000). *Sovereign Virtue*. Harvard University Press.

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

[28] Frankfurt, H. G. (2015). *On Inequality*. Princeton University Press.

[29] Freeman, S., editor (2003). *The Cambridge Companion to Rawls*. Cambridge University Press.

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

[31] Gauthier, D. (1983). *Morals by Agreement*. Oxford University Press.

[32] Georgopoulos, P., Elkhatib, Y., Broadbent, M., Mu, M., and Race, N. (2013). Towards network-wide QoE fairness using openflow-assisted adaptive video streaming. In *Proceedings of the 2013 ACM SIGCOMM Workshop on Future Human-centric Multimedia Networking*, pages 15–20.

[33] Gerdessen, J. C., Kanellopoulos, A., and Claassen, G. (2018). "Combining equity and utilitarianism"—Additional insights into a novel approach. *International Transactions in Operational Research*, 25:983–1000.

[34] Gini, C. (1912). *Variabilità e mutabilità*. P. Cuppini, reprinted 1955 in E. Pizetti abd T. Salvemini, eds., *Memorie di metodologica statistica*, Rome: Libreria Eredi Virgilio Veschi.

[35] Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In *Proceedings of 30th International Conference on Neural Information Processing*, pages 3323–3331.

[36] Harsanyi, J. C. (1977). *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*. Cambridge University Press.

[37] Hashimoto, T., Srivastava, M., Namkoong, H., and Liang, P. (2018). Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938. PMLR.

[38] Heidari, H., Ferrari, C., Gummadi, K. P., and Krause, A. (2018). Fairness behind a veil of ignorance: A welfare analysis for automated decision making. In *Advances in Neural Information Processing Systems*, pages 1265–1276.

[39] Heidari, H., Loi, M., Gummadi, K. P., and Krause, A. (2019). 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.

[40] Hooker, J. N. (2013). Moral implications of rational choice theories. In Lütge, C., editor, *Handbook of the Philosophical Foundations of Business Ethics*, pages 1459–1476. Springer.

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

[42] Hoover, E. M. (1936). The measurement of industrial localization. *Review of Economics and Statistics*, 18:162–171.

[43] Hoßfeld, T., Skorin-Kapov, L., Heegaard, P. E., and Varela, M. (2018). A new QoE fairness index for QoE management. *Quality and User Experience*, 3.

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

[45] Jain, R., Chiu, D. M., and Hawe, W. (1984). A quantitative measure of fairness and discrimination for resource allocation in shared computer systems. Technical Report TR–301, Eastern Research Laboratory, DEC, Hudson, MA.

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

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

[48] Karsu, O. and Morton, A. (2015). Inequality averse optimization in operational research. *European Journal of Operational Research*, 245:343–359.

[49] Kelly, F. P., Maulloo, A. K., and Tan, D. K. H. (1998). Rate control for communication networks: Shadow prices, proportional fairness and stability. *Journal of the Operational Research Society*, 49(3):237–252.

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

[51] Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. In *Proceedings of Advances in Neural Information Processing Systems*.

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

[53] Leonhardt, J., Anand, A., and Khosla, M. (2018). User fairness in recommender systems. In *Companion Proceedings of the Web Conference 2018*, pages 101–102.

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

[55] Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) – Protein Structure*, 405:442–451.

[56] Mazumdar, R., Mason, L., and Douligeris, C. (1991). Fairness in network optimal flow control: Optimality of product forms. *IEEE Transactions on Communications*, 39(5):775–782.

[57] McElfresh, C. and Dickerson, J. (2018). Balancing lexicographic fairness and a utilitarian objective with application to kidney exchange. *32nd AAAI Conference on Artificial Intelligence*, pages 1161–1168.

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

[59] Mehta, R. (2020). Recursive quadratic programming for constrained nonlinear optimization of session throughput in multiple-flow network topologies. *Engineering Reports*, 2:1–14.

[60] Mo, J. and Walrand, J. (2000). Fair end-to-end window-based congestion control. *IEEE/ACM Transactions on Networking*, 8:556–567.

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

[62] Moulin, H. (2004). *Fair Division and Collective Welfare*. MIT Press.

[63] Nanda, V., Xu, P., Sankararaman, K. A., Dickerson, J., and Srinivasan, A. (2020). 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.

[64] Nash, J. (1950). The bargaining problem. *Econometrica*, 18:155–162.

[65] Ogryczak, W., Luss, H., Pióro, M., Nace, D., and Tomaszewski, A. (2014). Fair optimization and networks: A survey. *Journal of Applied Mathematics*, 2014:1–25.

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

[67] Ogryczak, W. and Śliwiński, T. (2003). On solving linear programs with the ordered weighted averaging objective. *European Journal of Operational Research*, 148(1):80–91.

[68] Ogryczak, W. and Sliwinski, T. (2006). On direct methods for lexicographic min-max optimization. In Gervasi, O., Kumar, V., Tan, C., Taniar, D., Laganà, A., Mun, Y., and Choo, H., editors, *Computational Science and Its Applications (ICCSA)*, LNCS 3982, pages 802–811. Springer.

[69] Ogryczak, W., Wierzbicki, A., and Milewski, M. (2008). A multi-criteria approach to fair and efficient bandwidth allocation. *Omega*, 36(3):451–463. Special Issue on Multiple Criteria Decision Making for Engineering.

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

[71] Parfit, D. (1997). Equality and priority. *Ratio*, pages 201–221.

[72] Pearl, J. (2000). *Causality: Models, Reasoning and Inference.* Cambridge University Press.

[73] Pearl, J., Glymour, M., and Jewell, N. (2016). *Causal Inference in Statistics: A Primer.* Wiley.

[74] Pokhrel, S. R., Panda, M., Vu, H. L., and Mandjes, M. (2016). TCP performance over Wi–Fi: Joint impact of buffer and channel losses. *IEEE Transactions on Mobile Computing*, 15:1279–1291.

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

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

[77] Rubinstein, A. (1982). Perfect equilibrium in a bargaining model. *Econometrica*, 50:97—109.

[78] Russell, C., Kusner, M. J., Loftus, J. R., and Silva, R. (2017). When worlds collide: Integrating different counterfactual assumptions in fairness. In *Proceedings of 31st International Conference on Neural Information Processing Systems*, pages 6417–6426.

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

[80] Shah, K., Gupta, P., Deshpande, A., and Bhattacharyya, C. (2021). Rawlsian fair adaptation of deep learning classifiers. *arXiv preprint 2105.14890.*

[81] Siddique, U., Weng, P., and Zimmer, M. (2020). Learning fair policies in multi-objective (deep) reinforcement learning with average and discounted rewards. In *International Conference on Machine Learning*, pages 8905–8915. PMLR.

[82] Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K. P., Singla, A., Weller, A., and Zafar, M. B. (2018). A unified approach to quantifying algorithmic unfairness: Measuring individual &group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2239–2248.

[83] Stelmakh, I., Shah, N. B., and Singh, A. (2018). Peerreview4all: Fair and accurate reviewer assignment in peer review. *arXiv preprint 1806.06237*.

[84] Sühr, T., Biega, A. J., Zehlike, M., Gummadi, K. P., and Chakraborty, A. (2019). Two-sided fairness for repeated matchings in two-sided markets: A case study of a ride-hailing platform. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3082–3092.

[85] Theil, H. (1967). *Economics and Information Theory.* North-Holland.

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

[87] Verloop, I. M., Ayesta, U., and Borst, S. (2010). Monotonicity properties for multi-class queueing systems. *Discrete Event Dynamic Systems*, 20:473–509.

[88] Verma, S. and Rubin, J. (2018). Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness (FairWare)*, pages 1–7.

[89] Weng, P. (2019). Fairness in reinforcement learning. *arXiv preprint 1907.10323*.

[90] Williams, A. and Cookson, R. (2000). Equity in Health. *Culyer, A.J. and Newhouse, J.P. editors, Handbook of Health Economics.*

[91] Yager, R. (1997). On the analytic representation of the leximin ordering and its application to flexible constraint propagation. *European Journal of Operational Research*, 102(1):176 – 192.

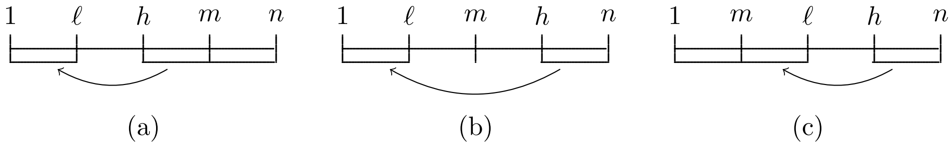
[92] Zafar, M. B., Valera, I., Rodrigues, M. G., and Gummadi, K. P. (2017). Fairness beyond disparate treatment and disparate impact: Learning classification without disparate mistreatment. In *Proceedings of 26th International Conference on World Wide Web*, pages 1171–1180.

Appendix

Proof of Theorem 1. Consider a utility distribution $\boldsymbol{u} = (u_1, \ldots, u_n)$ with $u_1 \leq \cdots \leq u_n$, and let $\tilde{u} = u_m$ be the median. There are three types of C–M utility transfers, illustrated in Fig. 6: (a) $\ell < h \leq m$, (b) $\ell \leq m < h$, and (c) $m < \ell < h$. Since a C–M transfer does not reorder the utilities, the new value of u_m is the median. If we let $U = \sum_{i=1}^{m} u_i$, the McLoone indices before and after the transfer are as in Table 4. It is easily checked that, in each case, the transfer does not reduce the index. This follows directly from algebraic manipulation in cases (a) and (b), and from the fact that $U \leq m u_m$ in case (c).

Proof of Theorem 2. Let $t(\boldsymbol{u})$ denote the number of utilities in the fair region for a given $\boldsymbol{u}$. We can distinguish three types of C–M utility transfer, illustrated in Fig. 7: (a) $\ell < h \leq t(\boldsymbol{u})$, (b) $\ell \leq t(\boldsymbol{u}) < h$, and (c) $t(\boldsymbol{u}) < \ell < h$. The resulting utility gain by individuals $1, \ldots \ell$, and loss by individuals $h, \ldots, n$, are indicated in Table 5. It is clear on inspection of Fig. 7 that the gain is at least ϵ in each case, and the loss never more than ϵ. The C-M condition is therefore satisfied.

Proof of Theorem 3. Given a utility distribution $\boldsymbol{u}$, let $\boldsymbol{u}'$ be the result of a Pigou-Dalton transfer of utility $\epsilon > 0$ from u_h to u_ℓ, where $u_\ell + \epsilon \leq u_h - \epsilon$.



(a)

(b)

(c)

Figure 6: Illustration of proof of Theorem 1.

Table 4: McLoone indices before and after a C–M transfer.

Case	Before transfer	After transfer
(a)	$\dfrac{U}{m u_m}$	$\dfrac{U + (n-m)\epsilon/(n-h+1)}{m\big[u_m - \epsilon/(n-h+1)\big]}$
(b)	$\dfrac{U}{m u_m}$	$\dfrac{U + \epsilon}{m U_m}$
(c)	$\dfrac{U}{m u_m}$	$\dfrac{U + m\epsilon/\ell}{m\big(u_m + \epsilon/\ell\big)}$

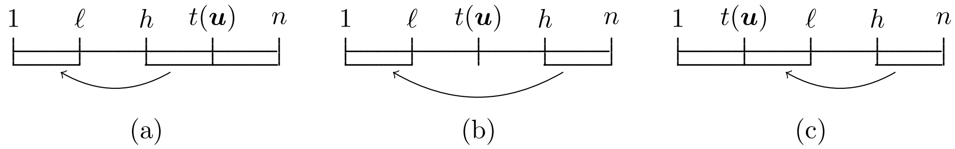


Figure 7: Illustration of proof of Theorem 2.

Table 5: Verifying the Chateauneuf–Moyes condition for a utility-threshold SWF

Case	Gain	Loss
(a)	$\dfrac{t(\boldsymbol{u})}{\ell}\epsilon > \epsilon$	$\dfrac{n-t(\boldsymbol{u})}{n-h+1}\epsilon < \epsilon$
(b)	$\dfrac{t(\boldsymbol{u})}{\ell}\epsilon > \epsilon$	ϵ
(c)	ϵ	ϵ

There are three cases to consider:

(a) $u_\ell, u_h \leq u_{\min} + \Delta$

(b) $u_\ell, u_h > u_{\min} + \Delta$

(c) $u_\ell \leq u_{\min} + \Delta$ and $u_h > u_{\min} + \Delta$

Case (c) breaks down into three subcases, where the following relations hold as well:

(c1) $u_\ell + \epsilon \leq u_{\min} + \Delta$ and $u_h > u_{\min} + \Delta$

(c2) $u_\ell + \epsilon, u_h - \epsilon \leq u_{\min} + \Delta$

(c3) $u_\ell + \epsilon, u_h - \epsilon > u_{\min} + \Delta$

Proof of Theorem 4. It is clear that a sufficiently small utility-invariant transfer satisfies the C-M condition when $k > t(\boldsymbol{u})$, because in this case $F_k(\boldsymbol{u})$ is simply utilitarian. We therefore need only consider the six cases illustrated in Fig. 8, in which $k \leq t(\boldsymbol{u})$. It is convenient to write $F_k(\boldsymbol{u})$ in the following form:

$$F_k(\boldsymbol{u}) = t(\boldsymbol{u})u_{\langle 1\rangle} + \sum_{i=2}^{k}(n-i+1)u_{\langle i\rangle} + \sum_{i=t(\boldsymbol{u})+1}^{n}(u_{\langle i\rangle} - \Delta)$$

Table 6: Verifying the Pigou-Dalton condition for an equity-threshold SWF

Case	$W(\boldsymbol{u})$	$W(\boldsymbol{u}')$
(a)	$(n-2)\Delta + u_\ell + u_h + U$	$(n-2)\Delta + u_\ell + u_h + U$
(b)	$n\Delta + 2u_{\min} + U$	$n\Delta + 2u_{\min} + U$
(c1)	$(n-1)\Delta + u_\ell + u_{\min} + U$	$(n-1)\Delta + u_\ell + u_{\min} + \epsilon + U$
(c2)	$(n-1)\Delta + u_\ell + u_{\min} + U$	$(n-2)\Delta + u_\ell + u_h + 2\epsilon + U$
(c3)	$(n-1)\Delta + u_\ell + u_{\min} + U$	$(n-1)\Delta + 2u_{\min} + U$

The resulting gain by individuals $1,\ldots\ell$, and loss by individuals $h,\ldots,n$, are indicated in Table 7. In cases (b)–(f), it is clear on inspection of Fig. 8 that the gain is more than ϵ in each case, and the loss never more than ϵ. In case (a), we note first that the gain can be written

$$n - \frac{\ell-1}{2} - \frac{n-t(\boldsymbol{u})}{\ell}$$

To show that the loss is no greater than the gain, it suffices to show this when $h = \ell+1$, since $h \geq \ell+1$ and the loss is nonincreasing with respect to h. Thus it suffices to show

$$n - \frac{\ell-1}{2} - \frac{n-t(\boldsymbol{u})}{\ell} \geq \frac{1}{n-\ell}\Big(\sum_{i=\ell+1}^{k}(n-i+1) + n - t(\boldsymbol{u})\Big)$$

Since $k \leq t(\boldsymbol{u})$ and each term of the summation is at most $n-\ell$, it suffices to show

$$n - \frac{\ell-1}{2} - \frac{n-t(\boldsymbol{u})}{\ell} \geq \frac{\big(t(\boldsymbol{u})-\ell\big)(n-\ell) + n - t(\boldsymbol{u})}{n-\ell}$$

Rearranging, we obtain

$$\big(n-t(\boldsymbol{u})\big)\Big(\frac{1}{\ell} + \frac{1}{n-\ell} - 1\Big) \leq \frac{\ell+1}{2} \tag{1}$$

This inequality is clearly satisfied when the following is false:

$$\frac{1}{\ell} + \frac{1}{n-\ell} \geq 1 \tag{2}$$

We therefore assume (2) is true. Since (1) is clearly satisfied when $\ell = 1$, we suppose $\ell \geq 2$, in which case (2) implies $n < \ell^2/(\ell-1)$. Since $\ell < h \leq n$, we can state

$$\ell + 1 \leq n < \frac{\ell^2}{\ell-1}$$

or $\ell^2 - 1 \leq n(\ell - 1) < \ell^2$. Since n and ℓ are positive integers, this implies $n = \ell + 1$, in which case (1) reduces to

$$\frac{\ell + 1 - t(\boldsymbol{u})}{\ell} \leq \frac{\ell + 1}{2}$$

This holds because $t(\boldsymbol{u}) \geq \ell + 1$, and the theorem follows. □

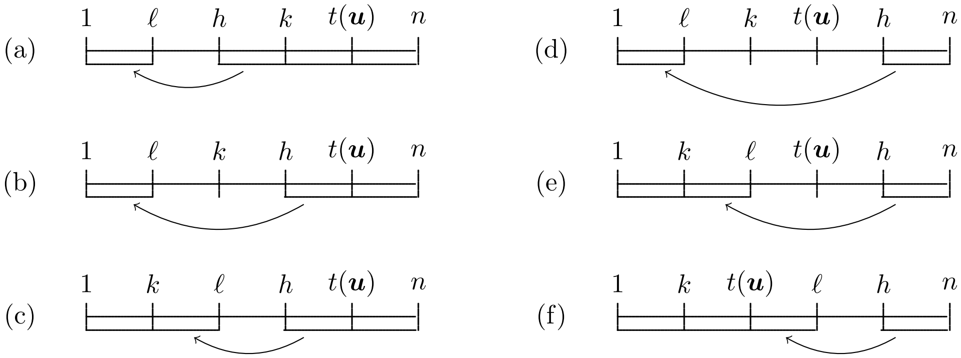


Figure 8: Illustration of proof of Theorem 4.

Table 7: Verifying the Chateauneuf–Moyes condition for $F_k(\boldsymbol{u})$

Case	Gain	Loss
(a)	$\frac{1}{\ell}\Big(t(\boldsymbol{u}) + \sum_{i=2}^{\ell}(n-i+1)\Big)\epsilon$	$\frac{1}{n-h+1}\Big(\sum_{i=h}^{k}(n-i+1) + n - t(\boldsymbol{u})\Big)\epsilon$
(b)	$\frac{1}{\ell}\Big(t(\boldsymbol{u}) + \sum_{i=2}^{\ell}(n-i+1)\Big)\epsilon \geq \frac{t(\boldsymbol{u})}{\ell}\epsilon > \epsilon$	$\frac{n-t(\boldsymbol{u})}{n-h+1}\epsilon < \epsilon$
(c)	$\frac{1}{\ell}\Big(t(\boldsymbol{u}) + \sum_{i=2}^{k}(n-i+1)\Big)\epsilon \geq \frac{t(\boldsymbol{u})}{\ell}\epsilon > \epsilon$	$\frac{n-t(\boldsymbol{u})}{n-h+1}\epsilon < \epsilon$
(d)	$\frac{1}{\ell}\Big(t(\boldsymbol{u}) + \sum_{i=2}^{\ell}(n-i+1)\Big)\epsilon \geq \frac{t(\boldsymbol{u})}{\ell}\epsilon > \epsilon$	$\frac{n-h+1}{n-h+1}\epsilon = \epsilon$
(e)	$\frac{1}{\ell}\Big(t(\boldsymbol{u}) + \sum_{i=2}^{k}(n-i+1)\Big)\epsilon \geq \frac{t(\boldsymbol{u})}{\ell}\epsilon > \epsilon$	$\frac{n-h+1}{n-h+1}\epsilon = \epsilon$
(f)	$\frac{1}{\ell}\Big(t(\boldsymbol{u}) + \sum_{i=2}^{k}(n-i+1) + \ell - t(\boldsymbol{u})\Big)\epsilon \geq \epsilon$	$\frac{n-h+1}{n-h+1}\epsilon = \epsilon$