Combining Leximax Fairness and Efficiency in a Mathematical Programming Model

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

[a] *Carnegie Mellon University*

Abstract

A trade-off between fairness and efficiency is an important element of many practical decisions. We propose a principled and practical method for balancing these two criteria in an optimization model. Following an assessment of existing schemes, we define a set of social welfare functions (SWFs) that combine Rawlsian leximax fairness and utilitarianism and overcome some of the weaknesses of previous approaches. In particular, we regulate the equity/efficiency trade-off with a single parameter that has a meaningful interpretation in practical contexts. We formulate the SWFs using mixed integer constraints and sequentially maximize them subject to constraints that define the problem at hand. We demonstrate the method on problems of realistic size involving healthcare resource allocation and disaster preparation, with solution times of several seconds at most.

Keywords: Ethics in OR, efficiency vs equity, leximax fairness

1. Introduction

Fairness is an important consideration across a wide range of optimization models. It can be a central issue in health care provision, disaster planning, workload allocation, public facility location, telecommunication network management, traffic signal timing, and many other contexts. While it is normally straightforward to formulate an objective function that reflects efficiency or cost, it is not obvious how to express fairness in mathematical form. When both fairness and efficiency are desired, as is typical in practice, there is the additional challenge of mathematically integrating them in a tractable model.

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

For example, when a natural disaster brings down the electric power grid, crisis managers may dispatch crews to urban areas first in order to restore power to more households quickly, thus maximizing efficiency. Yet this may cause rural areas to experience very long blackouts, which could be seen as unfair. A more satisfactory solution might give some amount of priority to rural customers, but without imposing too much harm on the population as a whole. Similarly, traffic signal timing that minimizes total delay may result in impracticably long wait times for traffic on minor streets that cross a main thoroughfare. Again a balance between equity and efficiency may be desirable. The issue can be especially acute in health care. Expensive treatments or research programs that prolong the life of a relatively few gravely ill patients may divert funds from preventive health measures that would spare thousands the suffering brought by less serious diseases.

We undertake in this paper to develop a practical and yet principled approach to balancing efficiency and fairness that can be implemented in a mixed integer/linear programming (MILP) model. While there are many possible measures of fairness, we choose a criterion based ultimately on John Rawls' concept of justice-as-fairness (Rawls 1999). One consequence of the Rawlsian analysis is his famous difference principle, which states roughly that a fair distribution of resources is one that maximizes the welfare of the worst-off. Rawls defends the principle with a social contract argument that can be plausibly extended to lexicographic maximization. That is, the welfare of the worst-off is first maximized subject to resource constraints, whereupon the welfare of the second worst-off is maximized while holding that of the worst-off fixed, and so forth. The Rawlsian perspective has been defended by closely reasoned philosophical arguments in a vast literature (Richardson and Weithman 1999, Freeman 2003).

The Rawlsian argument goes roughly as follows. Let's suppose that all concerned parties adopt an agreed-upon social policy in an original position behind a "veil of ignorance" as to their identity. It must be a policy that all parties can rationally accept upon learning who they are. Rawls argues that no rational decision maker will accept a policy in which she is the least advantaged, unless she would have been even worse off under any other policy. A fair outcome should therefore maximize the welfare of the worst-off. The argument can be employed recursively to defend a leximax criterion. Rawls intended his principle to apply only to the design of social institutions, and to pertain only to the distribution of "primary goods," which are goods that any rational person would want. Yet it can be plausibly extended to distributive justice in general, particularly if it is appropriately combined with an efficiency criterion.

A fundamental question that arises in the integration of equity and efficiency is how to regulate the trade-off between the two. We find in a survey of existing models that it is rarely clear how trade-off parameters can be selected and interpreted in a practical context. However, the modeling scheme of Hooker and Williams (2012) offers a potentially appealing approach to this problem. It governs the trade-off between a Rawlsian maximin and a utilitarian criterion with a single parameter Δ that has the same units as utility and can be related naturally to the problem at hand. The value of Δ is chosen so that parties whose utility is within Δ of the lowest are seen as sufficiently disadvantaged to deserve priority. Larger values of Δ result in greater equity. The model also has a practical mixed integer/linear programming (MILP) formulation.

The Hooker–Williams (H–W) scheme has a serious limitation, however. Because its fairness component is the maximin criterion, the actual utility levels of disadvantaged parties other than the very worst-off have no bearing on social welfare. As a result, the solution can be insensitive to most equity considerations. This outcome is particularly unsatisfactory when resource limitations tightly constrain the benefits available to a few parties, a situation we have found to be common in practice. The H–W model awards what utility it can to the most highly-constrained party, whereupon the welfare of other disadvantaged parties becomes irrelevant, and all solutions become virtually indistinguishable with respect to equity. The fairness criterion plays essentially no role in the determination of an optimum among a potentially large number of outcomes considered equally desirable by the H–W scheme, even for arbitrarily large values of Δ.

A natural way to address this problem is to combine efficiency with a lexicographic criterion rather than a maximin criterion. This allows the utility levels of all disadvantaged parties to factor into social welfare. However, it poses a difficult modeling challenge at both a theoretical and a computational level. We meet the challenge by maximizing a sequence of social welfare functions that, except for the first, are quite different from the single function used in the H–W model. Nonetheless the parameter Δ has a similar interpretation, with $\Delta = 0$ corresponding to a purely utilitarian solution and $\Delta = \infty$ to a purely leximax solution. We also show how to formulate these optimization problems as MILP problems that have substantially different constraint sets and polyhedral properties than the H–W formulation. We solve the sequence of MILP models in a matter of seconds for example problems of realistic size.

The paper is organized as follows. We begin in Section 2 with an assessment of the primary existing schemes for combining equity and efficiency in an optimization model. These include convex combinations of utility with an equity criterion, alpha fairness (and the special case of proportional fairness), the Kalai-Smorodinsky bargaining solution, and threshold models (of which the H–W scheme is an example). We then define a sequence of SWFs that can be maximized, subject to the constraints of the given problem, to obtain a socially optimal solution for a specified tradeoff parameter Δ. A key element of our proposal is a set of practical MILP models for these optimization problems. We also describe a family of valid inequalities that can be added to tighten the models. We extend these results to the common situation in which utility is distributed to groups rather than individuals, such as organizations, regions, or demographic groups. We conclude by demonstrating the practical applicability of our approach on a healthcare resource allocation problem and an emergency preparedness problems. The former allows us to compare results with those reported by Hooker and Williams (2012) on the same problem. The latter is a shelter location and assignment problem of realistic size. We find that our approach yields reasonable and nuanced socially optimal solutions for both problems, with computation times ranging from a fraction of a second to 18 seconds for a given Δ. Two Appendices, provided in an Online Supplement, contain proofs of theorems that were not proved in the body of the paper.

2. Related Work

An optimization model for integrating fairness and efficiency can be viewed as maximizing social welfare function (SWF) $F(\boldsymbol{u})$. The value of the function is interpreted as measuring the desirability of a given distribution $\boldsymbol{u} = (u_1, \ldots, u_n)$ of utilities, where u_i is the amount of utility allocated to party i. The function is maximized subject to resource limitations and other constraints imposed by the application.

2.1. Convex Combinations

The most obvious scheme for combining fairness and efficiency is a convex combination of the two. This corresponds to a SWF of the form

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

where $\Phi(\boldsymbol{u})$ is a fairness measure. A number of functions $\Phi(\boldsymbol{u})$ have been proposed, such as inequality metrics, the Rawlsian maximin principle, and leximax fairness (Cowell 2000, Jenkins and Van Kerm 2011, Karsu and Morton 2015).

A perennial problem with convex combinations is that it is difficult to interpret λ, particularly since $\Phi(\boldsymbol{u})$ is typically measured in units other than utility. For example, if we select the widely-used Gini coefficient $G(\boldsymbol{u})$ as a measure of equity, then we must combine utility with a dimensionless quantity $\Phi(\boldsymbol{u}) = 1 - G(\boldsymbol{u})$, where

$$G(\boldsymbol{u}) = \frac{\displaystyle\sum_{i<j} |u_i - u_j|}{n \displaystyle\sum_i u_i} \tag{2}$$

Another difficulty is that fairness measures are almost always nonlinear, which can pose tractability problems.

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

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

Yet we now have a convex combination of total utility and another equality metric (negative mean absolute difference) in which $\lambda = 1/2$. One may ask why this particular value of λ is suitable.

Since equality is often unsuitable as a fairness measure (Frankfurt 2015, Scanlon 2003), one may wish to use the Rawlsian criterion $\Phi(\boldsymbol{u}) = \min_i\{u_i\}$. It results in a convex combination of quantities that are measured in the same units, but it is again unclear how to select a suitable value of λ. Note that if we index utilities so that $u_1 \leq \cdots \leq u_n$, the convex combination becomes simply a weighted sum $u_1 + (1 - \lambda) \sum_{i>1} u_i$ that gives somewhat more weight to the lowest utility. It is unclear how much more weight is appropriate.

One might also attempt to formulate a convex combination of efficiency with a leximax rather than a maximin criterion. Yet it is unclear how to capture leximax in a function $\Phi(\boldsymbol{u})$ when the utilities cannot be ordered by size in advance. Ogryczak and Sliwinski (2006) show how to formulate leximax in an optimization model without pre-ordering, but this requires coefficients that vary enormously in size and can introduce numerical instability. There is also no evident means for incorporating an efficiency criterion into the model.

2.2. Alpha Fairness and Kalai-Smorodinsky Fairness

Alpha fairness is a parameterized combination of equity and efficiency that does not rely on a convex combination. It is based on an SWF of the form

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

Larger values of α imply a greater emphasis on equity, with $\alpha = 0$ corresponding to a pure utilitarian criterion and $\alpha = \infty$ to a maximin criterion. Lan et al. (2010) provide an axiomatic treatment, and Bertsimas et al. (2012) study worst-case equity/efficiency trade-offs. An interpretation of α is that 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. Yet it is again unclear what kind of trade-off, and therefore what value of α, is appropriate for a given application. There is also the computational issue that $F_\alpha(\boldsymbol{u})$ is nonlinear.

A well-known special case of α-fairness arises when $\alpha = 1$. This results in proportional fairness, which is equivalent to the Nash bargaining solution (Nash 1950). Nash (1950) showed that his bargaining solution for two persons is implied by a set of axioms for utility theory, including a strong and perhaps questionable axiom of cardinal noncomparability across parties (Hooker 2013). Harsanyi (1977), Rubinstein (1982), and Binmore et al. (1986) showed that the Nash solution is the asymptotic outcome of certain rational bargaining procedures, again based on strong assumptions.

Kalai and Smorodinsky (1975) proposed an alternative to the Nash bargaining solution that minimizes each player's relative concession. The approach is defended by Thompson (1994) and is consistent with the contractarian ethical philosophy of Gautier (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. Unfortunately, the K-S criterion can lead to anomalous situations that force overall utility gain to be arbitrarily small when it can be much greater with minimal sacrifice of fairness (Hooker 2013).

2.3. Threshold Models

Williams and Cookson (2000) proposed a pair of 2-person SWFs based on a utility or equity threshold. A utility-threshold model uses the maximin criterion unless the sacrifice in total utility exceeds a threshold, in which case it switches to a utilitarian criterion. An equity-threshold model uses a utilitarian criterion unless inequality becomes excessive, when it switches to maximin. Hooker and Williams (2012) extended the utility-threshold model to the n-person criterion described earlier, and McElfresh and Dickerson (2018) proposed a similar scheme based on a leximax rather than maximin criterion. Our aim in the present paper is likewise to combine leximax and efficiency in a threshold model, but we will argue that it offers two major advantages relative to the McElfresh and Dickerson approach.

The 2-person SWF implied by Williams and Cookson's utility-threshold model can be formulated

$$F(u_1, u_2) = \begin{cases} u_1 + u_2, & \text{if } |u_1 - u_2| \geq \Delta \\ 2\min\{u_1, u_2\} + \Delta, & \text{otherwise} \end{cases} \tag{3}$$

The function is utilitarian when $|u_1 - u_2| \geq \Delta$ and represents a maximin criterion otherwise. Indifference curves (contours) of the SWF are illustrated in Fig. 1. The maximin criterion $\min\{u_1, u_2\}$ is modified in (3) to obtain continuous contours. We will see that maintaining continuity is a major factor in the design of threshold-based SWFs.

The feasible set in Fig. 1 is the portion of the nonnegative quadrant under the curve. It represents all feasible utility outcomes that are permitted by the resource budget and other constraints. The shape of the curve indicates that when party 1's utility reaches a certain point, further improvement requires extraordinary sacrifice by party 2 due to the transfer of resources. The utilitarian solution (black dot in the figure) might therefore be viewed as preferable to the maximin solution (small open circle) and in fact yields slightly more social welfare as indicated by the contours.

Hooker and Williams (2012) extend this social welfare function to n persons as follows:

$$F_1(\boldsymbol{u}) = (n-1)\Delta + nu_{\langle 1\rangle} + \sum_{i=1}^{n} (u_i - u_{\langle 1\rangle} - \Delta)^+ \tag{4}$$

where $(\alpha)^+ = \max\{0, \alpha\}$. Here we adopt the convention that $(u_{\langle 1\rangle}, \ldots, u_{\langle n\rangle})$ is the tuple $(u_1, \ldots, u_n)$ arranged in non-decreasing order. We refer to the function as F_1 because it will be the first in a series of functions $F_1, \ldots, F_n$ we define later. It may be more intuitive to rewrite (4) as

$$F_1(\boldsymbol{u}) = \big(t(\boldsymbol{u}) - 1\big)\Delta + \sum_{i=1}^{t(\mathbf{u})} u_{\langle 1\rangle} + \sum_{i=t(\mathbf{u})+1}^{n} u_{\langle i\rangle}$$

where $t(\boldsymbol{u})$ is defined so that $u_{\langle 1\rangle}, \ldots, u_{\langle t(\boldsymbol{u})\rangle}$ are within Δ of $u_{\langle 1\rangle}$; that is, $u_{\langle i\rangle} - u_{\langle 1\rangle} \leq \Delta$ if and only if $i \leq t(\boldsymbol{u})$. We will refer to utilities $u_{\langle 1\rangle}, \ldots, u_{\langle t(\boldsymbol{u})\rangle}$

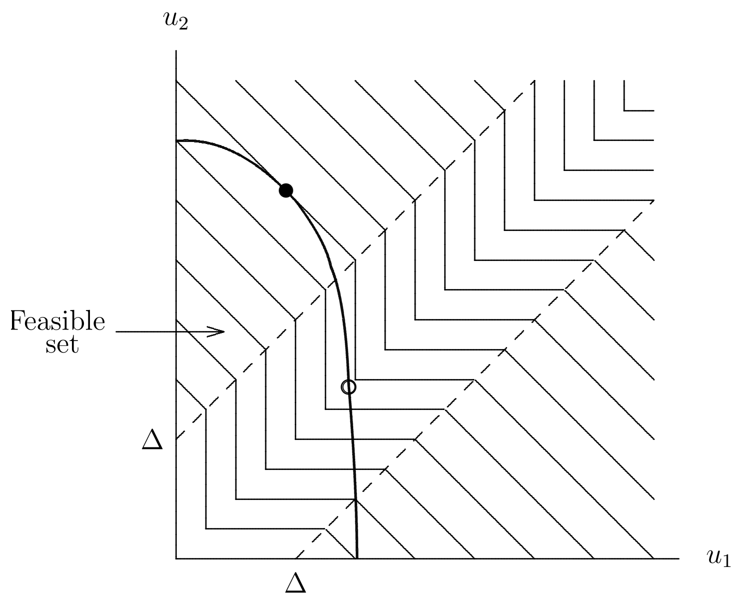


Figure 1: Piecewise linear social welfare contours for 2 persons.

as being *in the fair region* and utilities $u_{\langle t(\boldsymbol{u})+1\rangle}, \ldots, u_{\langle n\rangle}$ as being *in the utilitarian region*. The function $F_1(\boldsymbol{u})$ therefore has the effect of summing all the utilities, but with the proviso that utilities in the fair region are counted as equal to $u_{\langle 1\rangle}$. The term $(t(\boldsymbol{u})-1)\Delta$ is added to ensure continuity of the function.

The parameter Δ therefore has an interpretation that can be described independently of its role in the SWF. Namely, any party with utility within Δ of the lowest is viewed as disadvantaged and deserving of special consideration. The SWF then defines the special consideration to be an identification of the disadvantaged party with the worst-off party, which is given disproportionate weight in the summation of utilities—namely, weight equal to the number of utilities within Δ of the lowest.

A problem with (4), however, is that the actual utility levels of the disadvantaged parties, other than that of the very worst-off, have no effect on the value of the SWF. This is illustrated in the 3-person example of Fig. 2, which shows the contours of $F(u_1, u_2, u_3)$ 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 $F_1(\boldsymbol{u})$, so long as they remain in the fair region. As a result, there are infinitely many utility vectors that maximize social welfare, some of which

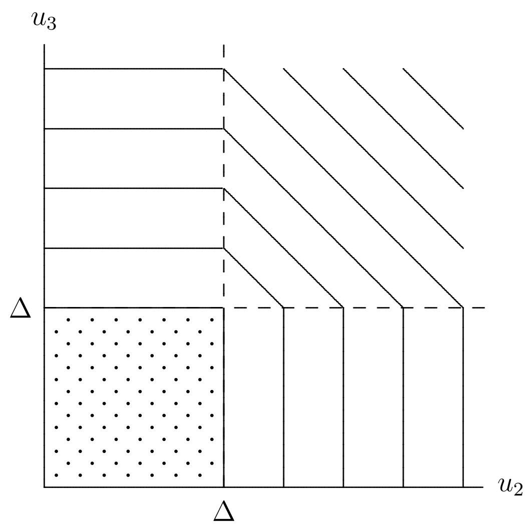


Figure 2: Contours of $F_1(0, u_2, u_3)$. The function is constant in the shaded region.

differ greatly with respect to utilities in the fair region. One can add a tie-breaking term $\epsilon(u_2 + u_3)$ to the social welfare function, where $\epsilon > 0$ is small, so as to maximize utility as a secondary objective. Yet this still does not account for equity considerations within the fair region.

To obtain a threshold model that is sensitive to the actual utility levels of all the disadvantaged parties, one might combine utility with a leximax criterion rather than a maximin criterion. McElfresh and Dickerson (2018) propose one method of doing so in the context of kidney exchange. Their method is related to the H–W approach, but 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

$$F(\boldsymbol{u}) = \begin{cases} nu_1, & \text{if } |u_i - u_j| \leq \Delta \text{ for all } i, j \\ \displaystyle\sum_i u_i + (N^+ - N^-)\Delta, & \text{otherwise} \end{cases} \tag{5}$$

where $N^+ = |\{i \mid u_1 > u_i\}|$ and $N^- = |\{i \mid u_1 < u_i\}|$ and the term $(N^+ - N^-)\Delta$ achieves continuity.

This approach can be seen as having two limitations. One, already noted, is that it is necessary to pre-specify a preference ranking of parties, as in the kidney exchange problem. This is not possible in many applications. Another is that the leximax criterion is not used until optimal solutions of the SWF are already obtained, and then applied only to the optimal solutions. We wish to allow the leximax criterion to play a role in evaluating all the possible solutions. These limitations are overcome by our proposal, described in the next section. An earlier version of our scheme appears in a brief conference paper (Chen and Hooker 2020), which uses somewhat different SWFs.

3. Defining the Social Welfare Functions

To combine leximax and utilitarian criteria in a threshold model, we propose to maximize a sequence of social welfare functions $F_1(\boldsymbol{u}), \ldots, F_n(\boldsymbol{u})$, each of which combines maximin and utilitarian measures. The first function $F_1(\boldsymbol{u})$ is the H–W function (4) defined earlier and is maximized over $\boldsymbol{u} = (u_1, \ldots, u_n)$ to obtain a value for $u_{\langle 1 \rangle}$. Each subsequent function $F_k(\boldsymbol{u})$ is maximized over $u_{\langle k \rangle}, \ldots, u_{\langle n \rangle}$, while fixing utilities $u_{\langle 1 \rangle}, \ldots, u_{\langle k-1 \rangle}$ to the values already obtained, and while giving $u_{\langle k \rangle}$ a certain amount of priority. The solution of this maximization problem determines the value of $u_{\langle k \rangle}$.

The process terminates when maximizing $F_k(\boldsymbol{u})$ yields a value of $u_{\langle k \rangle}$ that lies outside the fair region. At this point, $F_k(\boldsymbol{u})$ is utilitarian, and utilities $u_{\langle k \rangle}, \ldots, u_{\langle n \rangle}$ are determined simultaneously by maximizing $F_k(\boldsymbol{u})$ while fixing $u_{\langle 1 \rangle}, \ldots, u_{\langle k-1 \rangle}$ to the values already obtained. We refer to a utility vector $(u_{\langle 1 \rangle}, \ldots, u_{\langle n \rangle})$ that results from this process as *socially optimal.*

We describe this sequential optimization procedure more precisely in Section 4, but we must first define and explain the functions $F_k(\boldsymbol{u})$ for $k \geq 2$. Three main considerations govern the design of these functions and give them a significantly different character than $F_1(\boldsymbol{u})$.

- The fair region must be viewed as already defined, because $u_{\langle 1 \rangle}$ was fixed by maximizing $F_1(\boldsymbol{u})$.
- The utility $u_{\langle k \rangle}$ should receive less priority as k increases, since the it becomes less disadvantaged relative to the fixed lowest utility $u_{\langle 1 \rangle}$.

- It turns out that the priority given $u_{\langle k\rangle}$ cannot depend on the number of utilities in the fair region, as it does for $k = 1$, because this results in an irreducibly discontinuous SWF. We therefore design $F_k(\boldsymbol{u})$ so that the priority depends only on k.

To develop SWFs that are somewhat analogous to the H–W function $F_1(\boldsymbol{u})$ while reflecting these considerations, it is helpful to write $F_1(\boldsymbol{u})$ as

$$F_1(\boldsymbol{u}) = t(\boldsymbol{u})u_{\langle 1\rangle} + \big(t(\boldsymbol{u}) - 1\big)\Delta + \sum_{i=t(\boldsymbol{u})+1}^{n} u_{\langle i\rangle}$$

The function assigns weight $t(\boldsymbol{u})$ to utility $u_{\langle 1\rangle}$ and weight 1 to utilities in the utilitarian region. We modify this pattern follows:

$$F_k(\boldsymbol{u}) = \begin{cases} \displaystyle\sum_{i=1}^{k}(n-i+1)u_{\langle i\rangle} + \sum_{i=t(\boldsymbol{u})+1}^{n}(u_{\langle i\rangle} - u_{\langle 1\rangle} - \Delta), & \text{if } t(\boldsymbol{u}) \geq k \\ \displaystyle\sum_{i=1}^{n} u_{\langle i\rangle}, & \text{if } t(\boldsymbol{u}) < k \end{cases} \tag{6}$$

This SWF assigns weight $n - k + 1$ to utility $u_{\langle k\rangle}$ and weight 1 to utilities in the utilitarian region. As desired, the priority given to $u_{\langle k\rangle}$ depends only on k and decreases as k increases.

As the functions $F_k(\boldsymbol{u})$ are sequentially maximized for increasing values of k, each utility $u_{\langle k\rangle}$ in the fair region receives priority at some point in the process. This scheme incorporates lexicographic optimization in the sense that the smaller utilities are determined earlier in the sequence, although rather than maximizing $u_{\langle k\rangle}$ in step k, we maximize a SWF that gives priority to $u_{\langle k\rangle}$. Utilitarianism in incorporated because each maximization problem considers total utility as well as fairness.

For extreme values of Δ, this process yields purely utilitarian or purely leximax solutions. When $\Delta = 0$, we have $t(\boldsymbol{u}) = 1$ for all $\boldsymbol{u}$, and $F_1(\boldsymbol{u})$ reduces to a utilitarian criterion. The fair region is the single point $u_{\langle 1\rangle}$, and we solve the social welfare problem simply by maximizing $F_1(\boldsymbol{u})$, which yields a utilitarian solution. For sufficiently large Δ, $t(\boldsymbol{u}) = n$ for all feasible $\boldsymbol{u}$, and $F_k(\boldsymbol{u})$ is $(n - k + 1)u_{\langle k\rangle}$ plus a constant for $k = 1, \ldots, n$. Since all u_is lie in the fair region, we sequentially maximize $F_k(\boldsymbol{u})$ for $k = 1, \ldots n$ and therefore obtain a pure leximax solution. Intermediate values of Δ combine utilitarian and leximax criteria.

Figure 3 illustrates how maximizing $F_1(\boldsymbol{u}), \ldots, F_n(\boldsymbol{u})$ sequentially is more sensitive to equity than maximizing $F_1(\boldsymbol{u})$, which has the flat region

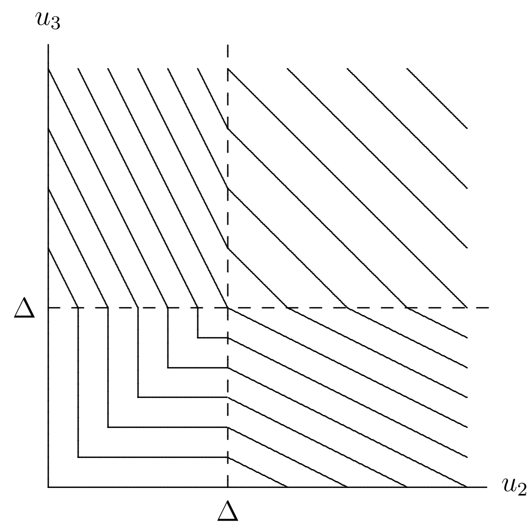


Figure 3: Contours of $F_2(0, u_2, u_3)$ with $\Delta = 3$ and contour interval 1.

shown in Fig. 2, as noted earlier. Suppose we determine a value for u_1 by maximizing $F_1(\boldsymbol{u})$, say $u_1 = 0$. Then the function $F_2(\boldsymbol{u})$ has no flat regions, as is evident in Fig. 3, and the solutions in the flat region of Fig. 2 are now distinguished. Note that the contours are continuous, which can be shown in general.

Theorem 1. The functions $F_k(\boldsymbol{u})$ are continuous for $k = 1, \dots, n$.

Proof. To prove continuity of $F_1(\boldsymbol{u})$, it suffices to show that each term of (4) is continuous, because a sum of continuous functions is continuous. The first term of (4) is a constant, and the second term is continuous because order statistics are continuous functions. Each term of of the summation is continuous because it is the maximum of two continuous functions. To show that $F_k(\boldsymbol{u})$ is continuous for $k \geq 2$, it is convenient to write (6) as

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

which simplifies to

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

Because order statistics are continuous, $u_{\langle k\rangle}$ and $u_{\langle 1\rangle}$ are continuous functions of $\boldsymbol{u}$. Also $\min\{u_{\langle 1\rangle}+\Delta, u_{\langle k\rangle}\}$ and $(u_{\langle i\rangle}-u_{\langle 1\rangle}-\Delta)^+$ are continuous because they are the minimum or maximum of continuous functions. $\square$

4. The Sequential Optimization Procedure

We now describe in detail how one can obtain a socially optimal utility distribution. We first simplify notation by removing the initial constants from $F_k(\boldsymbol{u})$ for $k \geq 2$, resulting in the SWF

$$\bar{F}_k(\boldsymbol{u}) = (n-k+1)u_{\langle k\rangle} + \sum_{i=k}^{n}(u_{\langle i\rangle}-u_{\langle 1\rangle}-\Delta)^+ \tag{7}$$

This obviously has no effect on the optimal solution that results from maximizing the SWF. For convenience, we define $\bar{F}_1(\boldsymbol{u}) = F_1(\boldsymbol{u})$.

We next maximize the social welfare functions $\bar{F}_1(\boldsymbol{u}), \ldots, \bar{F}_n(\boldsymbol{u})$ sequentially, subject to resource constraints, in such a way that maximizing $\bar{F}_k(\boldsymbol{u})$ determines the value of the kth smallest u_i in the socially optimal solution. We therefore maximize $\bar{F}_k(\boldsymbol{u})$ subject to the condition that the unfixed utilities are no smaller than the largest utility already fixed. Then the unfixed u_i with the smallest value in the solution becomes the utility determined by maximizing $\bar{F}_k(\boldsymbol{u})$.

We indicate resource limits by writing $\boldsymbol{u} \in \mathcal{U}$. In practice, they would be formulated in a MILP model by introducing variables and constraints that specify resource limitations and how resource allocations to individual parties translate to utilities. This will be illustrated in our experiments in Section 8.

To state the optimization procedure more precisely, we recursively define a sequence of maximization problems $P_1, \ldots, P_n$, where P_1 maximizes $F_1(\boldsymbol{u})$ subject to $\boldsymbol{u} \in \mathcal{U}$, and P_k for $k = 2, \ldots, n$ is

$$\begin{aligned} &\max\ \bar{F}_k(\boldsymbol{u}) \\ &u_i \geq \bar{u}_{i_{k-1}},\ \ i \in I_k \\ &\boldsymbol{u} \in \mathcal{U} \end{aligned} \tag{8}$$

The indices i_j are defined so that u_{i_j} is the utility determined by solving P_j. Thus

$$i_j = \underset{i \in I_j}{\operatorname{argmin}}\{u_i^{[j]}\}$$

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 denote by $\bar{u}_{i_j} = u_{i_j}^{[j]}$ the solution value obtained for u_{i_j} in P_j. 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

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

5. Mixed Integer Programming Model

For practical solution of the optimization problems P_k, we wish to formulate them as MILP models. We drop the resource constraints $u \in \mathcal{U}$ from problems $P_1, \ldots, P_n$ to obtain $P'_1, \ldots, P'_n$, because we wish to analyze the MILP formulations of the SWFs without the complicating factor of resource constraints. These constraints can later be added to the optimization models before they are solved. In addition, problems $P'_1, \ldots, P'_n$ contain innocuous auxiliary constraints that make the problems MILP representable.

The MILP model for P'_1 follows a different pattern than the models for $P'_2, \ldots, P'_n$, and we therefore treat the two cases separately. Problem P'_1 can be written

$$\begin{aligned} &\max\ z_1 \\ &z_1 \leq n u_{\langle 1 \rangle} + (n-1)\Delta + \sum_{i=2}^{n} (u_{\langle i \rangle} - u_{\langle 1 \rangle} - \Delta)^+ && (a) \\ &u_i \geq 0,\ \text{all } i && (b) \\ &u_i - u_j \leq M,\ \text{all } i, j && (c) \end{aligned} \tag{9}$$

Constraints (c) ensure MILP representability, as explained in Hooker and Williams (2012), because they imply that the hypograph of (9) is a finite union of polyhedra having the same recession cone (Jeroslow 1987). The constraints have no practical import for sufficiently large M, although for theoretical purposes we assume only $M > \Delta$.

The MILP model for P'_1 can be written as follows:

$$\begin{aligned} &\max\ z_1 \\ &z_1 \leq (n-1)\Delta + \sum_{i=1}^{n} v_i && (a) \\ &u_i - \Delta \leq v_i \leq u_i - \Delta\delta_i,\ \ i = 1, \ldots, n && (b) \\ &w \leq v_i \leq w + (M - \Delta)\delta_i,\ \ i = 1, \ldots, n && (c) \\ &u_i \geq 0,\ \ \delta_i \in \{0, 1\},\ i = 1, \ldots, n \end{aligned} \tag{10}$$

The following is proved in Hooker and Williams (2012).

Theorem 2. Model (10) is a correct formulation of P'_1.

When $k \geq 2$, the expression (7) for $\bar{F}_k(\boldsymbol{u})$ implies that problem P'_k can be written

$$\begin{array}{ll} \max\ z_k & \\ z_k \leq (n-k+1)\min\{\bar{u}_{i_1}+\Delta, u_{\langle k\rangle}\} + \sum_{i\in I_k}(u_i - \bar{u}_{i_1} - \Delta)^+ & (a) \\ u_i \geq \bar{u}_{i_{k-1}},\ \ i \in I_k & (b) \\ u_i - \bar{u}_{i_1} \leq M,\ \ i\in I_k & (c) \end{array} \tag{11}$$

The constraints (11c) are included to ensure that the problem is MILP representable. Since $\bar{u}_{i_1}$ is a constant, the hypograph is a union of bounded polyhedra whose recession cones consist of the origin only and are therefore identical.

The MILP model for P'_k when $k = 2, \ldots, n$ is

$$\begin{array}{ll} \max\ z_k & \\ z_k \leq (n-k+1)\sigma + \sum_{i\in I_k} v_i & (a) \\ 0 \leq v_i \leq M\delta_i,\ \ i \in I_k & (b) \\ v_i \leq u_i - \bar{u}_{i_1} - \Delta + M(1-\delta_i),\ \ i\in I_k & (c) \\ \sigma \leq \bar{u}_{i_1} + \Delta & (d) \\ \sigma \leq w & (e) \\ w \leq u_i,\ i \in I_k & (f) \\ u_i \leq w + M(1-\epsilon_i),\ i \in I_k & (g) \\ \sum_{i\in I_k}\epsilon_i = 1 & (h) \\ w \geq \bar{u}_{i_{k-1}} & (i) \\ u_i - \bar{u}_{i_1} \leq M,\ i \in I_k & (j) \\ \delta_i, \epsilon_i \in \{0,1\},\ i\in I_k & \end{array} \tag{12}$$

Theorem 3. Model (12) is a correct formulation of P'_k for $k = 2, \ldots, n$.

Proof. We first show that given any $(\boldsymbol{u}, z_k)$ that is feasible for (11), where $u_{i_j} = \bar{u}_{i_j}$ for $j = 1, \ldots, k-1$, there exist $\boldsymbol{v}, \boldsymbol{\delta}, \boldsymbol{\epsilon}, w, \sigma$ for which $(\boldsymbol{u}, z_k, \boldsymbol{v}, \boldsymbol{\delta}, \boldsymbol{\epsilon}, w, \sigma)$ is feasible for (12). Constraint (12j) follows directly from (11c). To satisfy the remaining constraints in (12), we set

$$\begin{aligned} &(\delta_i, \epsilon_i, v_i) = \left\{ \begin{array}{ll} (0,0,0), & \text{if } u_i - \bar{u}_{i_1} \le \Delta \text{ and } i \ne \kappa \\ (0,1,0), & \text{if } u_i - \bar{u}_{i_1} \le \Delta \text{ and } i = \kappa \\ (1,0,u_i - \bar{u}_{i_1} - \Delta), & \text{if } u_i - \bar{u}_{i_1} > \Delta \text{ and } i \ne \kappa \\ (1,1,u_i - \bar{u}_{i_1} - \Delta), & \text{if } u_i - \bar{u}_{i_1} > \Delta \text{ and } i = \kappa \end{array} \right\},\ i \in I_k \\ &w = u_\kappa \\ &\sigma = \min\{\bar{u}_{i_1} + \Delta, u_\kappa\} \end{aligned} \tag{13}$$

where κ is an arbitrarily chosen index in I_k such that $u_\kappa = u_{\langle k \rangle}$. It is easily checked that these assignments satisfy constraints (b)–(h). They satisfy (i) because (11b) implies that $u_\kappa \ge \bar{u}_{i_{k-1}}$. To show they satisfy (12a), we note that (12a) is implied by (11a) because $\min\{\bar{u}_{i_1} + \Delta, u_\kappa\} \le \sigma$ and $(u_i - \bar{u}_{i_1} - \Delta)^+ \le v_i$ for $i \in I_k$. Since (11a) is satisfied by $(\boldsymbol{u}, z)$, it follows that (12a) is satisfied by (13).

For the converse, we show that for any $(\boldsymbol{u}, z_k, \boldsymbol{v}, \boldsymbol{\delta}, \boldsymbol{\epsilon}, w, \sigma)$ that satisfies (12), $(\boldsymbol{u}, z_k)$ satisfies (11). Constraint (11b) follows from (12f) and (12i), and (11c) is identical to (12j). To verify that (11a) is satisfied, we let κ be the index for which $\epsilon_\kappa = 1$, which is unique due to (12g). It suffices to show that (12a) implies (11a) when the remaining constraints of (12) are satisfied. For this it suffices to show that

$$\sigma \le \min\{\bar{u}_{i_1} + \Delta, u_\kappa\} \tag{14}$$

$$v_i \le (u_i - \bar{u}_{i_1} - \Delta)^+,\ i \in I_k \tag{15}$$

(14) follows from (d), (e), and (f) of (12). (15) follows from (b) and (c) of (12). This proves the theorem. □

6. Valid Inequalities

In this section, we identify some valid inequalities that can strengthen the MILP model of P'_k for $k \ge 2$. The MILP model (10) for P'_1 is already sharp, meaning that the inequality constraints of the model describe the convex hull of the feasible set, and there is therefore no benefit in adding valid inequalities. The sharpness property may be lost when budget constraints are added, but the resulting model may remain a relatively tight formulation. When $n \le 3$, the models P'_k for $k \ge 2$ become sharp when the valid inequalities described below are added. This is not true when $n \ge 4$, but the valid inequalities nonetheless tighten the formulation.

The sharpness of the MILP model (10) for P_1' is proved in Hooker and Williams (2012). We present a simpler proof in Appendix 2.[1]

Theorem 4. The MILP model (10) is a sharp representation of P_1' (9).

We now describe a class of valid inequalities that can be added to the MILP model (12) of P_k' for $k \geq 2$ to tighten the formulation.

Theorem 5. The following inequalities are valid for P_k' for $k \geq 2$:

$$z_k \leq \sum_{i \in I_k} u_i \tag{16}$$

$$z_k \leq (n-k+1)u_i + \beta \sum_{j \in I_k \setminus \{i\}} (u_j - \bar{u}_{i_{k-1}}), \;\; i \in I_k \tag{17}$$

where

$$\beta = \frac{M - \Delta}{M - (\bar{u}_{i_{k-1}} - \bar{u}_{i_1})} = \Big(1 - \frac{\Delta}{M}\Big)\Big(1 - \frac{\bar{u}_{i_{k-1}} - \bar{u}_{i_1}}{M}\Big)^{-1} \tag{18}$$

Proof. It suffices to show that for any $(\boldsymbol{u}, z_k, \boldsymbol{v}, \boldsymbol{\delta}, \boldsymbol{\epsilon}, w)$ that satisfies (12), where $u_{i_j} = \bar{u}_{i_j}$ for $j = 1, \ldots, k-1$, the vector $\boldsymbol{u}$ satisfies (16) and (17). Since we know from Theorem 3 that $\boldsymbol{u}$ is feasible in (11), it suffices to show that (11) implies (16) and (17). To derive (16), we write (11a) as

$$z_k \leq \sum_{i \in I_k} \Big(\min\{\bar{u}_{i_1} + \Delta, u_{\langle k \rangle}\} + (u_i - \bar{u}_{i_1} - \Delta)^+ \Big) \tag{19}$$

For any term i in the summation, we consider two cases. If $u_i \leq \bar{u}_{i_1} + \Delta$, then $u_{\langle k \rangle} \leq \bar{u}_{i_1} + \Delta$ (because $u_{\langle k \rangle} \leq u_i$), and the term reduces to $u_{\langle k \rangle}$. If $u_i > \bar{u}_{i_1} + \Delta$, term i becomes

$$\min\{\bar{u}_{i_1} + \Delta, u_{\langle k \rangle}\} + (u_i - \bar{u}_{i_1} - \Delta) = \min\{0, u_{\langle k \rangle} - \bar{u}_{i_1} - \Delta\} + u_i \leq u_i$$

In either case, term i is less than or equal to u_i, and (16) follows.

To establish (17), it is enough to show that (17) is implied by (11) for each $i \in I_k$. We consider the same two cases as before.

Case 1: $u_i - \bar{u}_{i_1} \leq \Delta$, which implies $u_{\langle k \rangle} - \bar{u}_{i_1} \leq \Delta$. Since $\boldsymbol{u}$ satisfies (11a), we have

$$z_k \leq (n-k+1)u_{\langle k \rangle} + \sum_{\substack{j \in I_k \setminus \{i\} \\ u_j - \bar{u}_{i_1} > \Delta}} (u_j - \bar{u}_{i_1} - \Delta) \tag{20}$$

[1] The proof in Hooker and Williams (2012) can be simplified by using only the multipliers $\alpha_i = \frac{M}{n\Delta}\left(a_i - 1 + \frac{\Delta}{M}\right)$ for $i = 1, \ldots, n$, because each $a_i \geq 1 - \Delta/M$. The multipliers β_{ij} in their proof are unnecessary.

It suffices to show that this implies

$$z_k \leq (n-k+1)u_i + \beta\Big(\sum_{\substack{j\in I_k\setminus\{i\} \\ u_j-\bar{u}_{i_1}\leq\Delta}} (u_j - \bar{u}_{i_{k-1}}) + \sum_{\substack{j\in I_k\setminus\{i\} \\ u_j-\bar{u}_{i_1}>\Delta}} (u_j - \bar{u}_{i_{k-1}})\Big), \tag{21}$$

because (21) is equivalent to the desired inequality (17). But (20) implies (21) because $u_{\langle k\rangle} \leq u_i$ by definition of $u_{\langle k\rangle}$, $u_j - \bar{u}_{i_{k-1}} \geq 0$ for all $j \in I_k$ due to (11b), and it can be shown that

$$\beta(u_j - \bar{u}_{i_{k-1}}) \geq u_j - \bar{u}_{i_1} - \Delta \tag{22}$$

for any $j \in I_k$. To show (22), we note that the definition of β implies the following identity:

$$\bar{u}_{i_1} - \bar{u}_{i_{k-1}} + \Delta = (1-\beta)(M + \bar{u}_{i_1} - \bar{u}_{i_{k-1}}).$$

Adding $(1-\beta)\bar{u}_{i_{k-1}}$ to both sides, we obtain

$$\bar{u}_{i_{k-1}} - \beta\bar{u}_{i_{k-1}} + \Delta = (1-\beta)(M + \bar{u}_{i_1}) \geq (1-\beta)u_j, \tag{23}$$

where the inequality holds because $M + \bar{u}_{i_1} \geq u_j$ due to (11c). We obtain (22) by rearranging (23).

Case 2: $u_i - \bar{u}_{i_1} > \Delta$. It again suffices to show that (11) implies (21). Due to the case hypothesis, we have from (11a) that

$$z_k \leq (n-k+1)\min\{\bar{u}_1 + \Delta, u_{\langle k\rangle}\} + (u_i - \bar{u}_{i_1} - \Delta) + \sum_{\substack{j\in I_k\setminus\{i\} \\ u_j-\bar{u}_{i_1}>\Delta}} (u_j - \bar{u}_{i_1} - \Delta)^+$$

This can be written

$$\begin{aligned} z_k \leq (n-k+1)u_i - (n-k+1)\Big(u_i - \min\{\bar{u}_1 + \Delta, u_{\langle k\rangle}\}\Big) \\ +(u_i - \bar{u}_{i_1} - \Delta) + \sum_{\substack{j\in I_k\setminus\{i\} \\ u_j-\bar{u}_{i_1}>\Delta}} (u_j - \bar{u}_{i_1} - \Delta)^+ \end{aligned}$$

which can be written

$$\begin{aligned} z_k \leq (n-k+1)u_i - (n-k)\Big(u_i - \min\{\bar{u}_1 + \Delta, u_{\langle k\rangle}\}\Big) \\ -\Big(\bar{u}_1 + \Delta - \min\{\bar{u}_1 + \Delta, u_{\langle k\rangle}\}\Big) + \sum_{\substack{j\in I_k\setminus\{i\} \\ u_j-\bar{u}_{i_1}>\Delta}} (u_j - \bar{u}_{i_1} - \Delta)^+ \end{aligned} \tag{24}$$

The second term is nonpositive because $u_i > \bar{u}_1 + \Delta$ by the case hypothesis, and $u_i \geq u_{\langle k\rangle}$. The third term is clearly nonpositive. Thus (24) implies (21) because $u_j - \bar{u}_{i_{k-1}} \geq 0$ and (22) holds for $j \in I_k$ as before. □

7. Modeling Groups of Individuals

In many applications, utility is naturally allocated to groups rather than individuals, where individuals within each group receive an equal allocation. This occurs in the examples of Section 8, in which groups correspond to classes of patients with the same disease/prognosis or to neighborhood populations. In other applications, the number of individuals may be too large for practical solution, since problem P_i must be solved for each individual i. In such cases, individuals can typically be grouped into a few classes within which the individual differences are small or irrelevant, thus making the problem tractable and the results easier to digest. We therefore modify the above SWFs to accommodate groups rather than individuals. The theorems are proved in Appendix 2.

We suppose there are n groups of possibly different sizes. We let u_i denote the utility of each individual in group i and s_i the number of individuals in the group. Following Hooker and Williams, the SWF F_1 becomes

$$G_1(\boldsymbol{u}) = \Big(\sum_{i=1}^{n} s_i - 1\Big)\Delta + \Big(\sum_{i=1}^{n} s_i\Big)u_{\langle 1\rangle} + \sum_{i=1}^{n} s_i(u_i - u_{\langle 1\rangle} - \Delta)^+ \tag{25}$$

Hooker and Williams prove the following.

Theorem 6. The problem P_1', modified for groups, is equivalent to the MILP model

$$\begin{aligned} &\max\ z_1 \\ &z_1 \le \Big(\sum_{i=1}^{n} s_i - 1\Big)\Delta + \sum_{i=1}^{n} s_i v_i \\ &u_i - \Delta \le v_i \le u_i - \Delta\delta_i,\ \ i = 1,\dots,n \\ &w \le v_i \le w + (M-\Delta)\delta_i,\ \ i = 1,\dots,n \\ &u_i \ge 0,\ \ \epsilon_i, \delta_i \in \{0,1\},\ i = 1,\dots,n \end{aligned} \tag{26}$$

It is shown in Appendix 2 that for $k \ge 2$, we can adapt $\bar{F}_k(\boldsymbol{u})$ to groups as follows:

$$\bar{G}_k(\boldsymbol{u}) = \Big(\sum_{i=k}^{n} s_{\langle i\rangle}\Big)\min\{u_{\langle 1\rangle} + \Delta, u_{\langle k\rangle}\} + \sum_{i=k}^{n} s_{\langle i\rangle}(u_{\langle i\rangle} - u_{\langle 1\rangle} - \Delta)^+ \tag{27}$$

Theorem 7. The functions $\bar{G}_k(\boldsymbol{u})$ are continuous in $u_{\langle k\rangle}, \dots, u_{\langle n\rangle}$ for $k = 1, \dots, n$.

The MILP model is very similar to the one we developed for (11):

$$
\begin{aligned}
&\max\ z_k \\
&z_k \leq \Big(\sum_{i\in I_k} s_i\Big)\sigma + \sum_{i\in I_k} s_i v_i \quad (a) \\
&(12b)\text{–}(12j) \qquad\qquad\qquad\quad (b)\text{–}(j) \\
&\delta_i, \epsilon_i \in \{0,1\},\ i \in I_k
\end{aligned}
\tag{28}
$$

Theorem 8. The problem P'_k, reformulated for groups, is equivalent to (28) for $k = 2, \ldots, n$.

Hooker and Williams (2012) prove that (26) is a sharp representation of $P_1^{'}$ reformulated for groups. We present a simpler proof in Appendix 2.

Theorem 9. The MILP model (26) is a sharp representation of P'_1 reformulated for groups.

Finally, we describe a set of valid inequalities for the MILP model (28) for $k \geq 2$.

Theorem 10. The following inequalities are valid for the group problem P'_k for $k \geq 2$:

$$
z_k \leq \sum_{i\in I_k} s_i u_i \tag{29}
$$

$$
z_k \leq \Big(\sum_{j\in I_k} s_i\Big) u_j + \beta \sum_{j\in I_k\setminus\{i\}} s_j(u_j - \bar{u}_{i_{k-1}}),\quad i \in I_k \tag{30}
$$

where β is given by (18).

8. Applications

We now implement our approach on a healthcare resource allocation problem and a disaster management problem. We solve all MILP instances using Gurobi 8.1.1 on a desktop PC running Windows 10.

8.1. Healthcare Resource Allocation

A proper balance between fairness and efficiency is crucial in the allocation of healthcare resources. Hooker and Williams (2012) study a problem in which treatments are made available to patients on the basis of their disease and prognosis. In discussing this case, we caution that the results we report should not be taken as general recommendations for the allocation of medical resources. They are based on cost and clinical data specific to a particular set of circumstances. We use this example because it allows comparison with the published H–W results on the same problem instance.

Patients are divided into groups based on their disease and prognosis. There is one treatment potentially available to each patient group, and for policy consistency, it is provided to either all or none of the group members. Binary variable y_i is 1 if group i receives the recommended treatment and 0 otherwise. The average utility u_i experienced by members of group i is measured in terms of quality adjusted life years (QALYs); q_i is the net gain in QALYs for a member of group i when receiving the recommended treatment, and α_i is the expected QALYs experienced with medical management without the treatment. Thus

$$u_i = \alpha_i + q_i y_i, \;\; i = 1, \dots, n \tag{31}$$

The budget constraint is

$$\sum_i^n s_i c_i y_i \leq B \tag{32}$$

where s_i is the group size, c_i the cost of treating one patient in group i, and B the total available budget. The budget is set so as to force some hard decisions. The constraints (31)–(32), along with $y_i \in \{0, 1\}$ for $i = 1, \dots, n$, are added to the MILP models (26) and (28).

The H-W results are reproduced here in Table 1, in which the columns indicate solutions values of y_i for the 33 patient groups and various ranges of Δ. The treatments are pacemaker implant, hip replacement, aortic valve replacement, coronary artery bypass grafting (CABG), heart and kidney transplant, and kidney dialysis. Three types pf CABG surgery are distinguished (left main, double, and triple bypass), and kidney dialysis patients are distinguished by years of life expectancy with dialysis. Most of these categories are further divided into one, two, or three patient groups representing the degree of severity of the disease. The last column indicates the average number of QALYs per patient for each Δ range.

The results contain several interesting features, but most obvious is the transfer of resources from heart bypass surgery to dialysis as Δ increases. Kidney dialysis is quite costly, because the treatment is ongoing rather than a one-time event such as surgery. The payoff in QALYs per unit cost is therefore relatively low, and bypass surgery is selected when the utilitarian objective dominates (smaller values of Δ). As Δ increases, resources are transferred to dialysis patients, who are the worst off without treatment; heart bypass patients tend to have a fairly long life expectancy without the surgery. However, the less serious dialysis patients are treated first as Δ

increases, the opposite of what one should expect. This will be corrected in our leximax model.

Another problematic aspect of these results is that the average QALYs per patient decrease relatively little as Δ increases, as can be seen in the last column of Table 1. This is due to the fact that the H-W method does not take into account the utility levels of patients in the fair region (i.e., within Δ of the lowest), except for the very lowest. This results in a large space of alternate optimal solutions, many of them quite different from each other. To deal with this indeterminacy, the H-W experiments break ties by adding $\epsilon \sum_i s_i u_i$ to the objective function. This means that utilities in the fair region (except the lowest) are treated in a utilitarian fashion. Thus as Δ increases, the solution becomes basically utilitarian again, except that the welfare of the very worst-off patient is maximized.

The results of our model appear in Table 2. The computation time for a given Δ is negligible, almost always less than 0.5 second, even though there are 33 groups. The solution is significantly different from that of the H-W model. We note first that the average utility per patient drops considerably as Δ increases, indicating that equity plays a larger role for $\Delta > 0$ than in the H-W solution. Kidney dialysis enters the solution for much smaller values of Δ, and the more seriously ill kidney patients enter first, the reverse of what occurs in the H-W solution. This reflects the fact that our solution is sensitive to the utility levels of all disadvantaged patient groups rather than only the very worst-off.

There are other differences with the H-W solution. Heart bypass surgery remains in the solution for the most seriously ill patients, with some exceptions, through the entire range of Δ. This is again because the solution is sensitive to their disadvantaged position even though they are not the worst-off. Pacemakers now drop out of the solution for large Δ, even though pacemaker implantation is relatively inexpensive. This is because the pacemaker patients are better off without treatment than any of the other patients and therefore cease to receive priority as equity becomes more important. In general, these results indicate that incorporating leximax rather than maximin fairness in a social welfare function yields solutions that more adequately reflect equity considerations.

8.2. Shelter Location and Assignment

Disaster preparation and post-disaster response are important elements of humanitarian operations, in which equity is an essential consideration. We apply our approach to the shelter location and assignment problem investigated in Mostajabdaveh et al. (2019). There are two sets of decisions:

where to construct shelters in preparation for natural disasters, and how to assign one shelter to each potentially affected residential area. Mostajabdaveh et al. solve a model with multiple scenarios representing possible demands and street disruptions, but we simplify the problem by removing the stochastic element so as to clarify the equity/efficiency trade-off.

Utility is measured as negative cost, where cost is taken to be the travel distance between a residential area and its assigned shelter. A conventional efficiency objective is to minimize the average travel distance among all individuals. Optimizing this objective alone forces people in some areas to travel a long distance to their shelter, whereas equalizing convenience of access results in much greater total distance traveled.

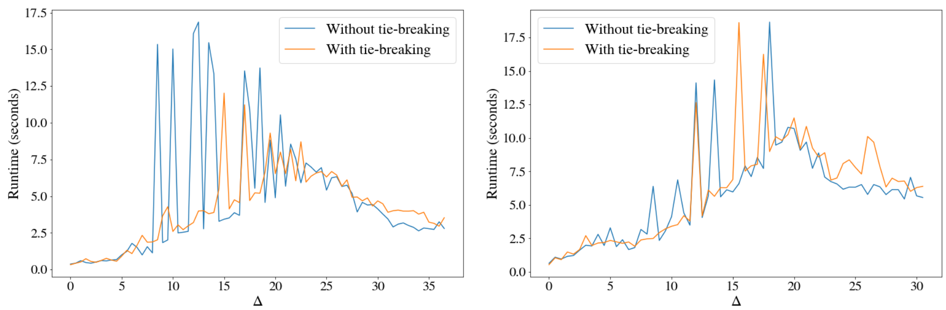
To formulate a MILP model of the problem, suppose m is the number of candidate locations for shelter, and n is the number of population areas. For $j \in \{1, \ldots, m\}$, c_j is the shelter capacity, and e_j is the cost of opening a shelter at location j. For $i \in \{1, \ldots, n\}$, s_i is the population of area i. We suppose that each person living in area i must travel a distance of D_{ij} to reach location j. Binary decision variable $y_j = 1$ when a shelter is open at location j, and binary variable $X_{ij} = 1$ when all persons living in area i are assigned to shelter j. The total cost of opening shelters must not exceed B. The model of Mostajabdaveh et al. assumes that each shelter is large enough for the entire population of any individual area, so that it is unnecessary to split areas between shelters. The resulting constraints are

$$\begin{aligned}
&\sum_{j=1}^{m} X_{ij} = 1, \;\; i = 1, \ldots, n \\
&\sum_{i=1}^{n} s_i X_{ij} \leq c_j y_j, \;\; j = 1, \ldots, m \\
&\sum_{j=1}^{m} e_j y_j \leq B \\
&X_{ij},\; y_j \in \{0, 1\}, \;\; i = 1, \ldots, n,\; j = 1, \ldots, m
\end{aligned} \tag{33}$$

The utility u_i of each person in area i is defined by the following constraints:

$$u_i = -\sum_{j=1}^{m} D_{ij} X_{ij}, \;\; i = 1, \ldots, n \tag{34}$$

Constraints (33)–(34) are added to the MILP models (26) and (28). We solve the problem with and without a tie breaking term in the objective function; the resulting solutions are very similar.



(a) Instance cp92 ($n = 50$, $m = 25$)

(b) Instance cp122 ($n = 50$, $m = 50$)

Figure 4: Runtime in shelter allocation example

We generate problem instances with one of the methods used by Mostajabdaveh et al. Instances of the capacitated warehouse location problem from Beasley (1988) are converted to shelter location instances by identifying shelters with warehouses, residential areas with customers, and population counts s_i with customer demands. The distance D_{ij} is taken to be C_{ij}/s_i, where C_{ij} is the cost of meeting all of customer i's demand from warehouse j. We use instances cp92 (25 locations) and cp122 (50 locations), both having 50 customers. The budget B is set to 150000 for cp92 and 300000 for cp122.

Fig. 4 shows that the run time is greater for intermediate values of Δ, less than 10 seconds in most cases but never more than 18 seconds. The resulting socially optimal solutions appear in Figs. 5 and 6. These plots show the evolution of per capita utility in individual neighborhoods as Δ increases. The shaded region indicates which utilities are in the fair region (within Δ of the worst). We see immediately that the problem is highly constrained, because the lowest utilities quickly reach a plateau and remain at a low level even for large Δ values. These neighborhoods are located at a considerable distance from candidate shelter locations, and so they remain disadvantaged even when given high priority. In this type of situation, it is particularly important to use a leximax rather than a maximin criterion of fairness, so as to take into account the situation of disadvantaged neighborhoods other than the very worst-off. If a maximin criterion were used, only the most distant neighborhood would have a bearing on equity. Other neighborhoods with utilities in the fair region would be treated arbitrarily or (if there is tie breaking) so as to minimize total travel distance, the latter resulting in an essentially utilitarian solution even for large Δ.

By contrast, the leximax component yields the desired result, because

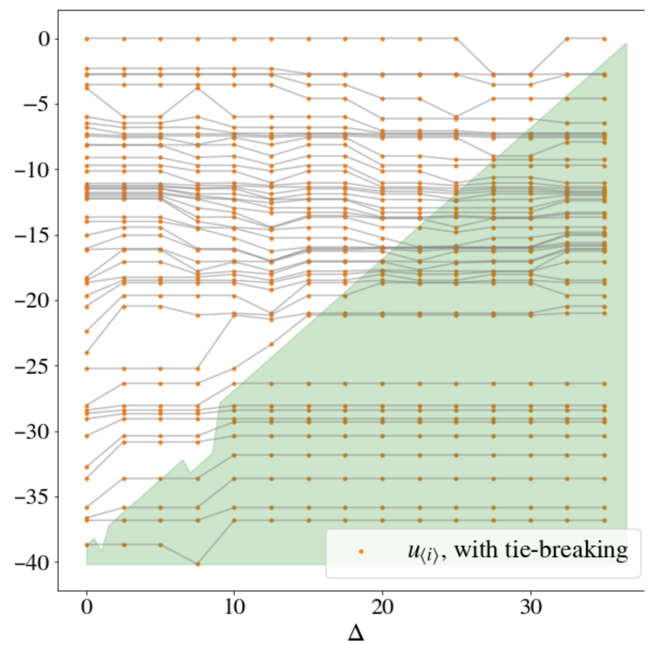


Figure 5: Utility distributions in shelter allocation instance CP92 ($n = 50$, $m = 25$)

several of the less advantaged neighborhoods improve their status as Δ increases. At the same time, some of the more privileged neighborhoods lose utility as they begin to sacrifice somewhat for the sake of the more remote neighborhoods. Interestingly, the less advantaged neighborhoods typically start to benefit from an increasing Δ shortly before they enter the fair region. The reason for this is that when a low utility level enters the fair region, it immediately becomes suboptimal due to the greater weight it receives there. It is therefore pushed up to one of the discrete feasible levels outside the fair region.

9. Conclusion

We propose a new systematic approach to balancing efficiency and equity in an optimization model, in which utility serves as the efficiency measure and Rawlsian leximax fairness as the equity measure. A parameter Δ regulates the equity/efficienty trade-off by dividing the feasible utility range into a *fair* region and a *utilitarian* region, where the fair region consists of utilities within Δ of the utility of the worst-off party. Leximax fairness

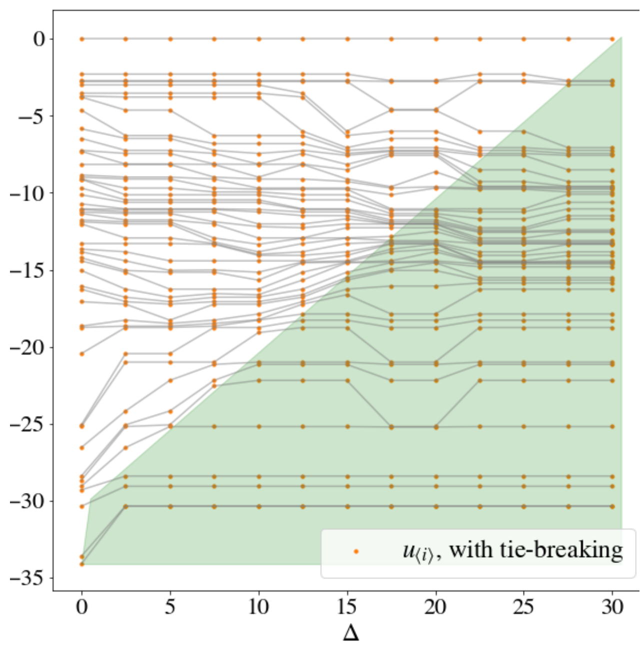


Figure 6: Utility distributions in shelter allocation instance CP122 ($n = 50$, $m = 50$)

is the dominating objective in the fair region, and utility dominates otherwise. Thus a single parameter allows a decision maker to control the balance between equity and efficiency by deciding which parties are sufficiently disadvantaged—that is, sufficiently near the worst-off—to deserve some degree of priority.

For a given optimization model, we combine equity and efficiency by solving a sequence of optimization problems, each of which maximizes a SWF subject to constraints from the original model. The SWFs are formulated using mixed integer constraints. They successively give priority to the worst-off, the second worst-off, and so on, with the degree of priority gradually decreasing relative to utilities in the utilitarian region.

As proof of concept, we apply our method to health resource allocation and disaster preparation problems. The solution time is at most a matter of seconds for a given value of Δ. We find that the solutions are not only sensitive to equity considerations but reveal complex and subtle trade-offs. This suggests that the modeling approach developed here can potentially serve as a useful mathematical tool for balancing fairness and efficiency in real-world situations.

Table 1: Results of healthcare example from Table 2 of Hooker and Williams (2012). A 1 indicates that the treatment is given to all members of a patient group, and a 0 indicates that the treatment is given to none.

Δ	Pace-	Hip	Aortic		CABG		Heart	Kidney			Kidney dialysis			Avg.
range	maker	repl.	valve	L	3	2	trans.	trans.	<1	1-2	2-5	5-10	>10	QALYs.
0-3.3	111	111	111	111	111	111	1	11	0	0	000	0000	000	7.54
3.4-4.0	111	111	111	111	111	111	0	11	1	0	000	0000	000	7.54
4.0-4.4	111	111	111	111	111	111	0	01	1	0	000	0000	001	7.51
4.5-5.01	111	011	111	111	111	111	1	01	1	0	000	0000	011	7.43
5.02-5.55	111	011	011	111	111	111	1	01	1	0	000	0001	011	7.36
5.56-5.58	111	011	011	111	111	011	0	01	1	0	000	0001	111	7.36
5.59	111	011	011	110	111	111	0	01	1	0	000	0001	111	7.20
5.60-13.1	111	111	111	101	000	000	1	11	1	0	111	1111	111	7.06
13.2-14.2	111	011	111	011	000	000	1	11	1	1	111	1111	111	7.03
14.3-15.4	111	111	111	011	000	000	1	11	1	1	101	1111	111	7.13
15.5 up	111	011	111	011	001	000	1	11	1	0	011	1111	111	7.19

Table 2: Results of healthcare example.

Δ	Pace-	Hip	Aortic		CABG		Heart	Kidney			Kidney dialysis			Avg.
range	maker	repl.	valve	L	3	2	trans.	trans.	<1	1-2	2-5	5-10	>10	QALYs.
0-0.2	111	111	111	111	111	111	1	11	0	0	000	0000	000	7.544
0.3	111	111	111	111	111	011	1	11	0	0	100	0000	001	7.542
0.4	111	111	111	111	011	011	1	11	0	1	100	1000	000	7.516
0.5	111	111	111	111	011	011	0	11	0	1	110	1000	000	7.514
0.6	111	111	111	111	011	001	1	01	0	1	111	1100	000	7.453
0.7	111	111	111	111	011	001	0	00	0	1	111	1100	100	7.405
0.8	111	111	111	011	011	001	1	00	0	1	111	1100	110	7.354
0.9	111	111	111	011	001	001	1	00	0	1	111	1101	110	7.283
1-1.9	111	111	011	011	001	001	1	00	0	1	111	1101	111	7.226
2-3.3	111	111	111	011	001	000	1	00	0	1	111	1111	111	7.206
3.4-4.55	011	111	111	011	001	000	1	00	1	1	111	1111	111	7.091
4.56-5.06	001	111	111	011	001	001	1	00	1	1	111	1111	111	6.940
5.07-5.34	011	111	111	011	001	000	1	00	1	1	111	1111	111	7.091
5.35-6.59	111	111	111	001	001	000	1	01	1	1	111	1111	111	7.123
6.60-8.43	011	111	111	011	001	000	1	00	1	1	111	1111	111	7.091
8.44-11.5	001	111	111	011	001	001	1	00	1	1	111	1111	111	6.940
11.6-13	010	111	111	011	001	001	1	00	1	1	111	1111	111	6.802
13 up	000	110	110	011	001	001	1	00	1	1	111	1111	111	5.942

References

[1] Beasley, J. E. (1988). An algorithm for solving large capacitated warehouse location problems. *European Journal of Operational Research*, 33(3):314–325.

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

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

[4] 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).

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

[6] 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.

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

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

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

[10] Gautier, D. (1983). *Morals by Agreement*. Oxford University Press.

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

[12] 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.

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

[14] 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.

[15] Jeroslow, R. (1987). Representability in mixed integer programming, I: Characterization results. *Discrete Applied Mathematics, 17:223-243.*

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

[17] Karsu, O. and Morton, A. (2015). Inequity averse optimization in operational research. *European Journal of Operational Research*, pages 343–359.

[18] 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.

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

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

[21] Nash, J. (1950). The bargaining problem. *Econometrica*, 18:155—-162.

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

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

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

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

[26] 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.

[27] 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.

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

Appendices

for "Combining Leximax Fairness and Efficiency in a Mathematical Programming Model"
by Violet (Xinying) Chen and J. N. Hooker

Appendix 1

In this Appendix, we prove that $F_k(\boldsymbol{u})$ satisfies the Chateauneuf-Moyes condition for all k. While it is widely recognized condition that an equality measure should satisfy the Pigou-Dalton condition, which requires that any utility transfer from a better-off party to a worse-off party increases (or at least does not decrease) social welfare, this is not obviously true of $F_k(\boldsymbol{u})$, since it is not an equality measure or even a fairness measure. In any event, Chateauneuf and Moyes (2005) have defined a weaker form of the Pigou-Dalton condition that all the functions $F_k(\boldsymbol{u})$ satisfy. It is based on transfers of utility from a better-off class to a worse-off class rather than from one individual to another. Specifically, it examines the consequences of transferring a given amount of utility from individuals whose utility lies above any given threshold (taking an equal share from each) to those whose utility lies below any given threshold (giving an equal share to each). Arguably, only such transfers should be considered, because a removal of utility from the upper range should remove at least as much from the best-off individual as from other well-off individuals, and an endowment of utility on the lower range should benefit the worst-off individual at least as much as other badly-off individuals. The Chateauneuf-Moyes (C-M) condition requires that such transfers result in at least as much social welfare.

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

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

A SWF $F(\boldsymbol{u})$ satisfies the C-M condition if C-M transfers never decrease social welfare. That is, for any $\boldsymbol{u}$ and any C-M transfer from $\boldsymbol{u}$ to $\boldsymbol{u}'$,

$$F(\boldsymbol{u}') \geq F(\boldsymbol{u}) \tag{A.1}$$

for sufficiently small $\epsilon > 0$.

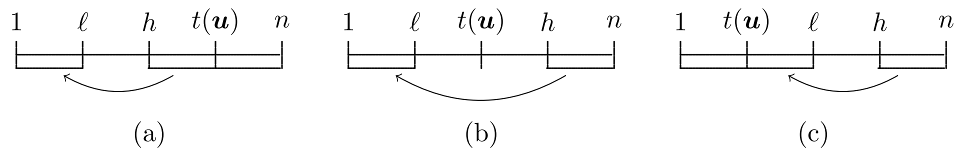


Figure A.1: Illustration of proof of Theorem 11.

Table A.1: Verifying the Chateauneuf-Moyes condition for $F_1(\boldsymbol{u})$

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

Theorem 11. The Hooker-Williams social welfare function $F_1(\boldsymbol{u})$ satisfies the Chateauneuf-Moyes condition.

Proof. It suffices to show that (A.1) holds for any $\boldsymbol{u}$ and sufficiently small $\epsilon > 0$. There are three types of utility transfer, illustrated in Fig. A.1: (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 A.1. It is clear on inspection of Fig. A.1 that the gain is at least ϵ in each case, and the loss never more than ϵ. The C-M condition is therefore satisfied. □

Theorem 12. The social welfare functions $F_k(\boldsymbol{u})$ satisfy the Chateauneuf-Moyes condition for $k = 2, \ldots, n$.

Proof. 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. A.3, in which $k \leq t(\boldsymbol{u})$. It is convenient to write $F_k(\boldsymbol{u})$ in the following form:

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

The resulting gain by individuals $1, \ldots \ell$, and loss by individuals $h, \ldots, n$, are indicated in Table A.3. In cases (b)–(f), it is clear on inspection of Fig. A.3 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

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

This inequality is clearly satisfied when the following is false:

$$\frac{1}{\ell} + \frac{1}{n - \ell} \geq 1 \tag{A.3}$$

We therefore assume (A.3) is true. Since (A.2) is clearly satisfied when $\ell = 1$, we suppose $\ell \geq 2$, in which case (A.3) 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 (A.2) 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. □

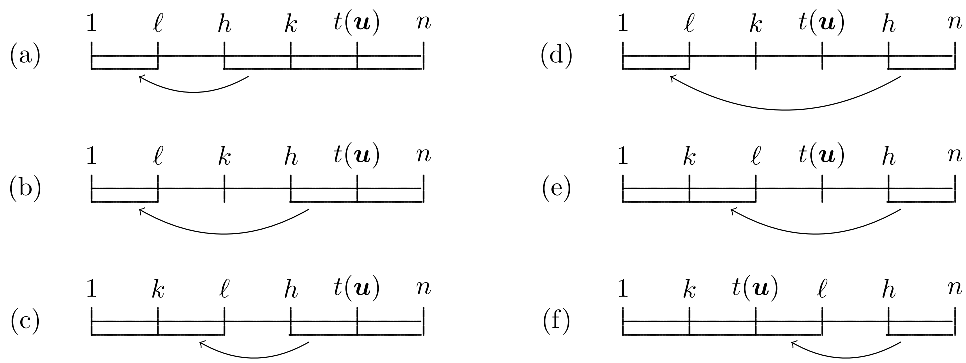


Figure A.3: Illustration of proof of Theorem 12.

Appendix 2

In this Appendix, we derive the group-related SWFs $G_2(\boldsymbol{u}), \ldots, G_m(\boldsymbol{u})$ and prove the relevant theorems. We obtain the SWFs by treating the group members as individuals and applying the the SWFs $F_k(\boldsymbol{u})$ for individuals, with the assumption that all individuals in a group have the same utility.

We begin by deriving $G_1(\boldsymbol{u})$. Let $u'_{i'}$ be the utility of *individual* i', and let u_i be the utility of each individual in *group* i. There are n' individuals and n groups. Let s_i be the size of group i, so that

$$n' = \sum_{i=1}^{n} s_i \tag{A.4}$$

Then

$$F_1(\boldsymbol{u}') = n'u'_{\langle 1\rangle} + (n'-1)\Delta + \sum_{i'=1}^{n'} (u'_{i'} - u'_{\langle 1\rangle} - \Delta)^+$$

Since $u_{\langle 1\rangle} = u'_{\langle 1\rangle}$ and group i has size s_i, we have

$$G_1(\boldsymbol{u}) = \Big(\sum_{i=1}^{n} s_i\Big) u_{\langle 1\rangle} + \Big(\sum_{i=1}^{n} s_i - 1\Big)\Delta + \sum_{i=1}^{n} s_i (u_i - u_{\langle 1\rangle} - \Delta)^+$$

This is the formula used in Hooker and Williams (2012).

Hooker and Williams prove that (10) is a sharp representation of P_1', and (26) a sharp representation of P_1' reformulated for groups. We present a simpler proof of both theorems. It is necessary only to prove the latter, because the former is a special case of it.

Table A.3: 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)	$\frac{1}{\ell}\Big(t(\boldsymbol{u}) + \sum_{i=2}^{k}(n-i+1) + \ell - t(\boldsymbol{u})\Big)\epsilon \geq \epsilon$	$\frac{n-h+1}{n-h+1}\epsilon = \epsilon$

Proof of Theorems 4 and 9. We prove Theorem 9, of which Theorem 4 is a special case in which $s_i = 1$ for each i. It suffices to show that any inequality $z_1 \leq \boldsymbol{a}^T\boldsymbol{u} + b$ that is valid for P_1' is a surrogate (nonnegative linear combination) of inequalities in (26). Let

$$N = \sum_{i=1}^{n} s_i$$

We first show that the following is a surrogate of (26) for any i:

$$z_1 \leq (N-1)\Delta + \Big(s_i + (N - s_i)\frac{\Delta}{M}\Big)u_i + \Big(1 - \frac{\Delta}{M}\Big)\sum_{j \neq i} s_j u_j. \tag{A.5}$$

We then show that $z_1 \geq \boldsymbol{a}^T\boldsymbol{u} + b$ is a surrogate of the inequalities (A.5). The theorem follows.

To show that (A.5) is a surrogate of (26), we first note that the following is a linear combination of the upper bounds on v_i in (26b) and (26c), using multipliers $1/\Delta$ and $1/(M - \Delta)$, respectively:

$$v_j \leq \frac{\Delta}{M}w + \Big(1 - \frac{\Delta}{M}\Big)u_j. \tag{A.6}$$

We also have the following from (26b) and (26c):

$$v_i \leq u_i. \tag{A.7}$$

$$w \le v_i. \tag{A.8}$$

We now obtain the following, for any given i and j, as a linear combination of (A.6) and (A.8), using multipliers 1 and Δ/M, respectively:

$$v_j \le \frac{\Delta}{M} v_i + \Big(1 - \frac{\Delta}{M}\Big) u_j. \tag{A.9}$$

Finally, we obtain (A.5) for any given i by summing (26a) with multiplier 1, (A.7) with multiplier

$$s_i + (N - s_i)\frac{\Delta}{M}$$

and (A.9) over all $j \ne i$ with multiplier s_j.

It remains to show that $z_1 \le \boldsymbol{a}^T\boldsymbol{u} + b$ is a surrogate of (A.5) for $i = 1, \ldots, n$. We first observe that $(\boldsymbol{u}, z) = (\boldsymbol{0}, (N-1)\Delta)$ is feasible in P_1' and must therefore satisfy $z_1 \le \boldsymbol{a}^T\boldsymbol{u} + b$, which implies $b \ge (N-1)\Delta$. We can assume without loss of generality that $b = (N-1)\Delta$, since otherwise we can add an appropriate multiple of the valid inequality $0 \le b$ to obtain the desired inequality $z_1 \le \boldsymbol{a}^T\boldsymbol{u} + b$. We also note that $(\boldsymbol{u}, z) = (M, \ldots, M, NM + (N-1)\Delta)$ is feasible and must satisfy $z_1 \le \boldsymbol{a}^T\boldsymbol{u} + (N-1)\Delta$, which means

$$\sum_{j=1}^{m} a_j \ge N \tag{A.10}$$

Finally, we note that $(\boldsymbol{u}, z) = (M\boldsymbol{e}_i, (N-1)\Delta + s_i(M-\Delta))$ is feasible for P_1', where $\boldsymbol{e}_i$ is the ith unit vector. Substituting this into $z_1 \le \boldsymbol{a}^T\boldsymbol{u} + (N-1)\Delta$, we obtain

$$a_i \ge \Big(1 - \frac{\Delta}{M}\Big) s_i \tag{A.11}$$

Due to (A.10), we can suppose without loss of generality that $\sum_{j=1}^m a_j = N$, since otherwise we can add appropriate multiples of the valid inequalities $0 \le a_j$ to obtain $z_1 \le \boldsymbol{a}^T\boldsymbol{u} + b$.

To obtain $z \le \boldsymbol{a}^T\boldsymbol{u} + b$ as a surrogate of (A.5), we sum (A.5) over all j using the multipliers

$$\alpha_i = \frac{M}{N\Delta}\Big(a_i - \Big(1 - \frac{\Delta}{M}\Big)s_i\Big) \tag{A.12}$$

for each i. It is easily checked that $\sum_{i=1}^m \alpha_i = 1$, so that the linear combination has the form

$$z \le \boldsymbol{d}^T\boldsymbol{u} + (N-1)\Delta \tag{A.13}$$

We wish to show that $\boldsymbol{d} = \boldsymbol{a}$. Note that

$$d_i = \Big(s_i + N\frac{\Delta}{M}\Big)\alpha_i + \Big(1 - \frac{\Delta}{M}\Big)s_i \sum_{j\neq i} \alpha_j$$

Using the fact that $\sum_{j=1}^m \alpha_j = 1$, this becomes

$$d_i = N\frac{\Delta}{M}\alpha_i + \Big(1 - \frac{\Delta}{M}\Big)s_i$$

which immediately reduces to $d_i = a_i$. We conclude that (A.13) is a linear combination of the inequalities (A.5) using multipliers α_i. It remains to show that each α_i is nonnegative, but this follows from (A.11) and (A.12). □

We now derive $\bar{G}_k(\boldsymbol{u})$ for $k \geq 2$. Recall that the SWF for individuals is

$$\bar{F}_{k'}(\boldsymbol{u}') = (n' - k' + 1)\min\{u'_{\langle 1\rangle} + \Delta, u'_{\langle k'\rangle}\} + \sum_{i'=k'}^{n'} (u'_{\langle i'\rangle} - \bar{u}'_{\langle 1\rangle} - \Delta)^+ \quad \text{(A.14)}$$

To obtain $\bar{G}_k(\boldsymbol{u})$, we again assume the individuals in each group i have the same utility u_i. The first individual in (A.14) that belongs to group k is individual

$$k' = 1 + \sum_{j=1}^{k-1} s_{i_j} \quad \text{(A.15)}$$

Due to (A.4) and (A.15), the first term on the RHS of (A.14) is

$$\Big(n' - 1 - \sum_{j=1}^{k-1} s_{i_j} + 1\Big)\min\{u_{\langle k\rangle} + \Delta, u_{\langle k\rangle}\} = \Big(\sum_{i=k}^{n} s_{\langle i\rangle}\Big)\min\{u_{\langle k\rangle} + \Delta, u_{\langle k\rangle}\}$$

since all the utilities in a group are the same. Thus we have

$$\bar{G}_k(\boldsymbol{u}) = \Big(\sum_{i=k}^{n} s_{\langle i\rangle}\Big)\min\{u_{\langle 1\rangle} + \Delta, u_{\langle k\rangle}\} + \sum_{i=k}^{n} s_{\langle i\rangle}(u_{\langle i\rangle} - u_{\langle 1\rangle} - \Delta)^+ \quad \text{(A.16)}$$

We show as follows that $\bar{G}_k(\boldsymbol{u})$ is continuous in $u_{\langle k\rangle}, \ldots, u_{\langle n\rangle}$.

Proof of Theorem 7. It suffices to show each term of (A.16) is a continuous function of $u_{\langle k\rangle}, \ldots, u_{\langle n\rangle}$, with $u_{\langle 1\rangle}, \ldots, u_{\langle k-1\rangle}$ and the corresponding group sizes $s_{\langle 1\rangle}, \ldots, s_{\langle k-1\rangle}$ fixed. The first term is continuous because it is equal to a constant time the maximum of order statistics $u_{\langle 1\rangle}$ and $u_{\langle k\rangle}$, which are continuous functions of $\boldsymbol{u}$. Similarly, each term of the summation is a constant times the maximum of a continuous expression and zero. □

We can now establish that the MILP model (28) is correct.

Proof of Theorem 8. We first show that given any $(\boldsymbol{u}, z_k)$ that is feasible for (11), where $u_{i_j} = \bar{u}_{i_j}$ for $j = 1, \ldots, k-1$, there exist $\boldsymbol{v}, \boldsymbol{\delta}, \boldsymbol{\epsilon}, w, \sigma$ for which $(\boldsymbol{u}, z_k, \boldsymbol{v}, \boldsymbol{\delta}, \boldsymbol{\epsilon}, w, \sigma)$ is feasible for (28). Constraint (28j) follows directly from (11c). To satisfy the remaining constraints in (28), we assign values to $\boldsymbol{v}, \boldsymbol{\delta}, \boldsymbol{\epsilon}, w, \sigma$ as in (13), where κ is an arbitrarily chosen index in I_k such that $u_\kappa = u_{\langle k \rangle}$. It is easily checked that these assignments satisfy constraints (28b)–(28h). They satisfy (28i) because (11b) implies that $u_\kappa \geq \bar{u}_{i_{k-1}}$. To show they satisfy (28a), we note that (28a) is implied by (11a) because $\min\{\bar{u}_{i_1} + \Delta, u_\kappa\} \leq \sigma$ and $(u_i - \bar{u}_{i_1} - \Delta)^+ \leq v_i$ for $i \in I_k$. Since (11a) is satisfied by $(\boldsymbol{u}, z)$, it follows that (28a) is satisfied by (13).

For the converse, we show that for any $(\boldsymbol{u}, z_k, \boldsymbol{v}, \boldsymbol{\delta}, \boldsymbol{\epsilon}, w, \sigma)$ that satisfies (28), $(\boldsymbol{u}, z_k)$ satisfies (11). Constraint (11b) follows from (12f) and (12i), and (11c) is identical to (28j). To verify that (11a) is satisfied, we let κ be the index for which $\epsilon_\kappa = 1$, which is unique due to (28g). It suffices to show that (28a) implies (11a) when the remaining constraints of (28) are satisfied. For this it suffices to show that

$$\sigma \leq \min\{\bar{u}_{i_1} + \Delta, u_\kappa\} \tag{A.17}$$

$$v_i \leq (u_i - \bar{u}_{i_1} - \Delta)^+, \; i \in I_k \tag{A.18}$$

(A.17) follows from (d), (e), and (f) of (28). (A.18) follows from (b) and (c) of (28). This proves the theorem. □

Finally, we show that (29)–(30) are valid inequalities for the group version of P_k' for $k \geq 2$, which is

$$\begin{array}{lr} \max \; z_k & \\ z_k \leq \Big(\sum_{i \in I_k} s_i\Big) \min\{\bar{u}_{i_1} + \Delta, u_{\langle k \rangle}\} + \sum_{i \in I_k} s_i (u_i - \bar{u}_{i_1} - \Delta)^+ & (a) \\ u_i \geq \bar{u}_{i_{k-1}}, \;\; i \in I_k & (b) \\ u_i - \bar{u}_{i_1} \leq M, \;\; i \in I_k & (c) \end{array} \tag{A.19}$$

Proof of Theorem 10. It suffices to show that for any $(\boldsymbol{u}, z_k, \boldsymbol{v}, \boldsymbol{\delta}, \boldsymbol{\epsilon}, w)$ that satisfies (28), where $u_{i_j} = \bar{u}_{i_j}$ for $j = 1, \ldots, k-1$, the vector $\boldsymbol{u}$ satisfies (16) and (17). Since we know from Theorem 8 that $\boldsymbol{u}$ is feasible in (A.19), it suffices to show that (A.19) implies (29) and (30). To derive (29), we write (A.19a) as

$$z_k \le \sum_{i\in I_k} s_i\Big(\min\{\bar{u}_{i_1}+\Delta, u_{\langle k\rangle}\} + (u_i - \bar{u}_{i_1} - \Delta)^+\Big) \tag{A.20}$$

For any term i in the summation, we consider two cases. If $u_i \le \bar{u}_{i_1}+\Delta$, then $u_{\langle k\rangle} \le \bar{u}_{i_1}+\Delta$ (because $u_{\langle k\rangle} \le u_i$), and the term reduces to $s_i u_{\langle k\rangle}$. If $u_i > \bar{u}_{i_1}+\Delta$, term i becomes

$$s_i\Big(\min\{\bar{u}_{i_1}+\Delta, u_{\langle k\rangle}\} + (u_i - \bar{u}_{i_1} - \Delta)\Big) = s_i\Big(\min\{0, u_{\langle k\rangle} - \bar{u}_{i_1} - \Delta\} + u_i\Big) \le s_i u_i$$

In either case, term i is less than or equal to u_i, and (29) follows.

To establish (30), it is enough to show that (30) is implied by (A.19) for each $i \in I_k$. We consider the same two cases as before.

Case 1: $u_i - \bar{u}_{i_1} \le \Delta$, which implies $u_{\langle k\rangle} - \bar{u}_{i_1} \le \Delta$. Since $\boldsymbol{u}$ satisfies (A.19a), we have

$$z_k \le \Big(\sum_{j\in I_k} s_j\Big) u_{\langle k\rangle} + \sum_{\substack{j\in I_k\setminus\{i\}\\ u_j-\bar{u}_{i_1}>\Delta}} s_j(u_j - \bar{u}_{i_1} - \Delta) \tag{A.21}$$

It suffices to show that this implies

$$z_k \le \Big(\sum_{j\in I_k} s_j\Big) u_i + \beta\Big(\sum_{\substack{j\in I_k\setminus\{i\}\\ u_j-\bar{u}_{i_1}\le\Delta}} s_j(u_j - \bar{u}_{i_{k-1}}) + \sum_{\substack{j\in I_k\setminus\{i\}\\ u_j-\bar{u}_{i_1}>\Delta}} s_j(u_j - \bar{u}_{i_{k-1}})\Big), \tag{A.22}$$

because (A.22) is equivalent to the desired inequality (30). But (A.21) implies (A.22) because $u_{\langle k\rangle} \le u_i$ by definition of $u_{\langle k\rangle}$, $u_j - \bar{u}_{i_{k-1}} \ge 0$ for all $j \in I_k$ due to (A.19b), and (A.3) for all $j \in I_k$.

Case 2: $u_i - \bar{u}_{i_1} > \Delta$. It again suffices to show that (A.19) implies (A.22). Due to the case hypothesis, we have from (A.19a) that

$$z_k \le \Big(\sum_{j\in I_k} s_j\Big) \min\{\bar{u}_1+\Delta, u_{\langle k\rangle}\} + s_i(u_i - \bar{u}_{i_1} - \Delta) + \sum_{\substack{j\in I_k\setminus\{i\}\\ u_j-\bar{u}_{i_1}>\Delta}} s_j(u_j - \bar{u}_{i_1} - \Delta)$$

This can be written

$$\begin{aligned} z_k \le \Big(\sum_{j\in I_k} s_j\Big) u_i - \Big(\sum_{j\in I_k} s_j\Big)\Big(u_i - \min\{\bar{u}_1+\Delta, u_{\langle k\rangle}\}\Big) \\ + s_i(u_i - \bar{u}_{i_1} - \Delta) + \sum_{\substack{j\in I_k\setminus\{i\}\\ u_j-\bar{u}_{i_1}>\Delta}} s_j(u_j - \bar{u}_{i_1} - \Delta) \end{aligned}$$

which can be written

$$
\begin{aligned}
z_k \leq \Big(\sum_{j\in I_k} s_j\Big)u_i - \Big(\sum_{j\in I_k\setminus\{i\}} s_j\Big)\Big(u_i - \min\{\bar{u}_1 + \Delta, u_{\langle k\rangle}\}\Big) \\
- s_i\Big(\bar{u}_1 + \Delta - \min\{\bar{u}_1 + \Delta, u_{\langle k\rangle}\}\Big) + \sum_{\substack{j\in I_k\setminus\{i\} \\ u_j - \bar{u}_{i_1} > \Delta}} s_j(u_j - \bar{u}_{i_1} - \Delta)
\end{aligned}
\tag{A.23}
$$

The second term is nonpositive because $u_i > \bar{u}_1 + \Delta$ by the case hypothesis, and $u_i \geq u_{\langle k\rangle}$. The third term is clearly nonpositive. Thus (A.23) implies (A.22) because $u_j - \bar{u}_{i_{k-1}} \geq 0$ and (22) holds for $j \in I_k$ as before. $\square$