Fairness, Efficiency, and Feature-Awareness in the Allocation of Public Goods

Zach Siegel

advised by Auyon Siddiq

UCLA Anderson School of Management

Abstract

We modify classical resource allocation problems by considering heterogeneity among individual beneficiaries via feature-aware *utility functions, with the goal of reducing unequal social welfare between subpopulations. We discuss pitfalls of feature-aware optimization, including an exacerbation of between-group inequality, and we provide examples motivating our idea of* efficient *and* fair *solutions. Finally, we propose an exact method for obtaining solutions with a degree of fairness controlled by a single parameter, and we discuss an approximate approach whose degree of fairness is attenuated in a similar way.*

1 Introduction

Many optimization models assume the perspective of a social planner, allocating resources so as to increase the utility of individuals within a population. Our running example will be facility location, which could refer to establishment of transit stops or polling places during an election. The social planner's goal is to maximize social welfare across a population of *individuals*.

Framing the social planner's task as an optimization problem, a non-subjective proxy for utility might suffice, such as *proximity to nearest facility*. Often, however, individuals have intrinsic *features* that influence the utility they derive from a particular resource allocation. For example, a polling place one mile away may be inaccessible to someone whose income is low and is unlikely to own a car; on the other hand, the utility an individual derives from a transit stop might depend on their income, in that wealthier individuals may be unlikely to use the stop.

Some objective terms may not reflect individuals' utilities, but those that do can be considered together via a *social welfare function*. A sum of individual utilities is a common social welfare function, but this choice relies on subtle philosophical assumptions, and there are other natural but less-common strategies. When individuals' utilities are a function of both resources allocated and *individuals' features*, principled consideration of the social welfare function may be warranted.

It is natural to consider *fairness* in a social planning setting. A fairness-seeking goal might be *equality* across the entire population, meaning low variation in utility. In our setting, individuals have features that are *protected* in a legal or ethical sense, and it is important to ensure that individuals in groups defined by these features have comparable outcomes to others.

Several problems are used as examples throughout, but we start by defining the facility-location problem, in which to explore "fair" allocation of resources within a heterogeneous population. Section 2 reviews notions of fair algorithmic decision-making in the economics, machine-learning, and operations literature. Section 3 provides a taxonomy of the adverse "fair" decision-making and illustrates them through numerical examples, motivating the fairness-inducing social welfare strategies proposed in Section 4.

1.1 Utility Model

We consider a heterogeneous population of N individuals, whose individuals experience utility that is a function of both their *features* and a *resource allocation.* Suppose individual i has a feature vector θ_i, is allocated resources $r_i \in \mathbb{R}$, and experiences utility $u_i = f(\theta_i, r_i)$.

We are not overly concerned with computational methods, and don't impose restrictions on f until Section 4. In this section we establish a running example used throughout, though our treatment of *feature-aware* individual utility and social welfare is widely applicable.

In some cases an individual might experience stochastic binary utility: a successful or unsuccessful outcome, affected but not fully determined by a social planner's decisions. In this case, a utility function could represent the *probability of a successful outcome*:

$$\begin{aligned} u_i &= f(\theta_i, r_i) \\ &= P(\text{individual } i \text{ has a successful outcome} | \theta_i, r_i). \end{aligned}$$

Here, an objective of the form $\sum_{i=1}^{N} u_i$ represents *the expected number of successful outcomes.* A natural for f is a logistic function.

Suppose individual i's feature vector θ_i consists of $\theta_i = (\theta_i^U, \theta_i^P)$, their *unprotected* and *protected* features. For our running example, let $\theta_i^P = g_i \in \{0, 1\}$ denote the group affiliation of individual i, i.e. the protected features consist of a single binary categorical identifier denoting group affiliation. Suppose there are n_U unprotected features and $\theta_i^U = (\theta_i^{U_1}, \ldots, \theta_i^{U_{n_u}})$.

Define the following coefficient vectors:

$$\begin{aligned} \theta_i &= (\theta_i^U, \theta_i^P) \\ &= (\theta_i^{U_1}, \ldots, \theta_i^{U_{n_U}}, g_i), \\ \beta^U &= (\beta_1^U, \ldots, \beta_{n_U}^U), \\ \beta^P &= (\beta^P), \\ \beta^r &= (\beta^r), \end{aligned}$$

and then utility can be written

$$\begin{aligned} u_i &= f(\theta_i, r_i) \\ &= \frac{1}{1 + e^{-\beta^0 - {\beta^U}^T \theta_i^U - {\beta^P}^T \theta_i^P - \beta^r r_i}} \\ &= P(\text{individual } i \text{ has a successful outcome} | \theta_i, r_i). \end{aligned} \tag{1}$$

1.2 Feature-Agnostic Facility Location

Our primary running example of social planning is that of facility location, though our treatment of feature-aware public resource allocation is applicable to any setting considerate of individual utilities determined by centralized decision-making.

Suppose there are N individuals and M potential facilities of which m can be selected. To minimize the sum of distances to facilities, the following optimization problem suffices:

$$\begin{array}{lll} \min & \displaystyle\sum_{i=1}^{N} r_i & \\ \text{s.t.} & r_i & \geq \displaystyle\sum_{j=1}^{M} d_{ij} x_{ij}, i = 1, \ldots, N \\ & \displaystyle\sum_{j=1}^{M} x_{ij} & = 1, i = 1, \ldots, N \\ & x_{ij} & \leq y_j \ \forall \ i = 1, \ldots, N, j = 1, \ldots, M \\ & \displaystyle\sum_{j=1}^{M} y_j & \leq m \\ & x_{ij} & \geq 0, i = 1, \ldots, N, j = 1, \ldots, M \\ & y_j & \in \{0, 1\}. \end{array} \tag{2}$$

Here, y_j represents whether facility j is selected; x_{ij} represents whether individual i is assigned to facility j; r_i represents individual i's *distance to the closest facility*. In a general sense, r_i represents the *resource allocation* to individual i. Note that the x_{ij} variables will naturally be binary.

1.3 Feature-Aware Facility Location

Now suppose individual i's utility is $u_i = f(\theta_i, r_i)$ where r_i is the distance of individual i from their nearest facility as in (2). Setting aside the form we specified for f in Section 1.1, a more general problem setting is the following:

$$\begin{array}{lll} \max & \sum_{i=1}^{N} u_i & \\ \text{s.t.} & u_i & = f(\theta_i, r_i) \\ & r_i & \geq d_{ij} x_{ij} \\ & \sum_{j=1}^{M} x_{ij} & = 1 \\ & x_{ij} & \leq y_j \ \forall\ i = 1, \ldots, N, j = 1, \ldots, M \\ & \sum_{j=1}^{M} y_j & \leq m \\ & x_{ij} & \geq 0 \\ & y_j & \in \{0, 1\}. \end{array} \tag{3}$$

In order for the r_i in this problem to continue to correspond to the distance from the nearest facility, it is natural to assume $\frac{\partial}{\partial r} f(\theta, r) \leq 0$ (a higher distance yields less utility); in a setting in which r_i represented allocation of a utility-increasing resource, the opposite would be true.

The feature-ignorant facility location formulation (2) is equivalent to $f(\theta_i, r_i) = -r_i$.

1.4 Social Welfare Function

Incorporating individual utilities into a social planner's objective function is the central concern of this work. In Sections 1.1, 1.2, and 1.3, individual utilities are either deterministic or stochastic, and they either ignore or incorporate exogenous individual features, but in all cases individual utilities are *added* to create a social welfare function. This natural formulation is far from the only possibility; Section 2 reviews social welfare criteria in the literature, and Section 4 proposes a social welfare functional form promoting *α-fairness.*

Social welfare functions can be characterized in terms of their functional forms or implicitly via optimality conditions. The running example of a facility-location problem includes discrete decisions and the notion of differential optimality conditions is not well-defined. So, we introduce a second resource-allocation setting with continuous decision variables. Let $\boldsymbol{u} = (u_1, \ldots, u_N)$, and consider a "social welfare function" that is weakly increasing in each u_i:

$$\begin{array}{lll} \max & SW(\boldsymbol{u}) & \\ \text{s.t.} & u_i & = f(r_i) \\ & \sum_{i=1}^{N} r_i & = 1 \\ & r_i & \geq 0. \end{array} \tag{4}$$

where r_i represents the resources allocated to individual i.

We list three social welfare objectives from the literature (theoretically justified in [8] and [2]), ordered from least-to-most "fairness-seeking":

1. Fairness-neutral utilitarian objective, attributed to the social philosopher Jeremy Bentham ("Social Welfare Bentham"):

$$SWB(\boldsymbol{u}) = \sum_{i=1}^{N} u_i, \tag{5}$$

   which in (4) has the interior optimality condition $f'(r_i) = f'(r_j)\ \forall\ i, j$. That is, individuals' marginal utilities should be equal.

2. Fairness-favoring multiplicative objective, referred to as the Nash standard of comparison

or Nash bargaining solution ("Social Welfare Nash"):

$$SWN(\boldsymbol{u}) = \prod_{i=1}^{N} u_i, \tag{6}$$

which clearly leads to the same optima as its logarithm,

$$SWNlog(\boldsymbol{u}) = \sum_{i=1}^{N} \log(u_i). \tag{7}$$

Letting $g(r_i) = \log(f(r_i)) = \log(u_i)$, by the same reasoning as in the utilitarian objective, an interior optimality condition is that

$$\begin{aligned} g'(r_i) &= g'(r_j) \\ &\Downarrow \\ \frac{f'(r_i)}{f(r_i)} &= \frac{f'(r_j)}{f(r_j)}, \end{aligned} \tag{8}$$

which effectively prioritizes lower-utility individuals by associating with the marginal utility a weight inversely proportional to utility.

This solution maximizes the volume of the N-dimensional rectangle (orthotope) with opposite vertices at the origin and the end of the vector $\boldsymbol{u}$, and faces parallel to the axes.

3. Strongly fairness-inducing max-min objective, attributed to the American philosopher John Rawls ("Social Welfare Rawls"):

$$SWR(\boldsymbol{u}) = \min_{i=1,\ldots,N} u_i. \tag{9}$$

Optimizing SWR (maximin) will result in the lowest worst-case inequality of the three objectives listed, but it is indiscriminate between solutions that provide additional benefit to higher-utility individuals, and it may not always be appropriate.

4. Weighting for fairness:

$$SWWeight(\boldsymbol{u}) = \sum_{i=1}^{N} w(\theta_i) u_i. \tag{10}$$

where $w(\theta_i)$ may simply associate a higher weight priority to individuals whose type indicates membership in a marginalized group (affirmative action), or may be a complicated function of a multi-dimensional type that prioritizes individuals who are likely to have a low probability of success.

Optimizing an appropriate social welfare objective can induce a range of notions of fairness, some of which have compelling stochastic interpretations. Recall the stochastic facility location problem (3), where $u_i = f(\theta_i, r_i) = P(\text{individual } i \text{ has a successful outcome} | \theta_i, r_i)$. In this case, the objective $SWB(\boldsymbol{u}) = \sum_{i=1}^{N}$ in (5) is the *expected number of successful outcomes*. Alternatively the objective $SWN(\boldsymbol{u}) = \prod_{i=1}^{N} u_i$ in (6) is *the probability that every individual has a successful outcome*.

Note that the logarithm in the Nash bargaining solution (7) can be replaced with any strictly concave increasing function to achieve a similar *low-utility-aversion*, an analog to risk-aversion, effectively prioritizing low-utility individuals. In fact, [11, 2], and [1] note a family of low-utility-averse functions that includes (5), (7), and (9):

$$SW_{\xi}(\boldsymbol{u}) = \begin{cases} \frac{1}{1-\xi} \sum_i u_i^{1-\xi} : \xi \geq 0, \xi \neq 1 \\ \sum_i \log(u_i) : \xi = 1, \end{cases} \tag{11}$$

which equals the utilitarian (5) when $\xi = 0$, equals (7) when $\xi = 1$, and is equal to the max-min (9) when $\xi \to \infty$. A higher ξ is associated with a stronger sense of "fairness".

We propose the weighted social welfare function

(10) for several reasons. For one, adding weights is unlikely to require different computational methods, and so this may provide a flexible fairness-inducing strategy (as opposed to e.g. composing a logarithm with utility as in the Nash bargaining solution (7)). Adding weights eludes a compelling stochastic interpretation like (6) in the preceding paragraphs, but the optimality conditions in (8) can be reproduced using an appropriate weighting function, and so presumably the same "fairness-seeking" results will follow. We argue that this affirmative action-minded weighting scheme can in fact be more flexible and powerful in achieving a notion of fairness. This form is the basis for our proposed fairness-seeking method in Section 4.

2 Review of Fairness Criteria

In addition to the different notions of fairness promoted by a choice of social welfare function, described in Section 1.4, many indices have been proposed to empirically measure fairness, along with strategies to promote it in optimization. Of central concern is always a balance of some notion of "accuracy" (or "utilitarianism" or "efficiency") with the notion "fairness". We will advocate for and against some strategies that have been proposed for making this tradeoff, we will explore what we consider adverse outcomes in Section 3, and we will propose fairness-inducing strategies in Section 4.

Most of the indices that follow can be utilized in optimization in the following ways:

- as an objective function, with accuracy constrained to exceed some threshold,
- as a constraint, with accuracy as the stated objective,
- as a term in a constraint that also includes accuracy,
- as one objective among several, at least one of which measures accuracy, in a multi-objective optimization problem.

The most important distinction between the following metrics is that some measure *between-group unfairness* and others measure *within-group unfairness* (or *individual unfairness*), as delineated in [10].

2.1 Between-Group Unfairness

In the following examples, we consider the case of a population partitioned into two groups, where group membership of individual i is denoted by indicator variable g_i, constituting the individual's protected features: $\theta_i^P = g_i$.

A review of fair machine learning is provided in [3], in which three notions of between-group fairness are delineated: *anti-classification* where "protected attributes—like race, gender, and their proxies—are not explicitly used to make decisions", *classification parity* where "common measures of predictive performance (e.g., false positive and false negative rates) are equal across groups defined by the protected attributes", and *calibration* where "conditional on risk estimates, outcomes are independent of protected attributes".

Some of the same authors as [3], in their preview work [4], discuss the cost of fairness in a decision process that utilizes exogenous risk scores. They also note that in a decision problem, utility may be fruitfully decomposed into an "immediate" and and long-term utilities, reflecting that any "accuracy" or "utilitarian" objective need not be the only concern in resource allocation. They prove that the utility-maximizing "fair" decision process is to use different risk score thresholds for different subpopulations.

[12] treats *classification parity* as *disparate mistreatment* and proposes the covariance between individual *type* and *misclassification rate* (e.g. false positive rate) as a reasonable proxy to minimize. This

work is similar to ours in that it deals with decision-making informed by an exogenous predictive model. Analogous to their strategy addressing misclassification parity, instead of enforcing ϵ-parity in resource allocation, meaning

$$|E(f(\theta, r)|g = 0) - E(f(\theta, r|g = 1))| < \epsilon,$$

one can instead limit covariance between group membership and outcome:

$$|Cov(f(\theta, r), g)| < \epsilon,$$

and this covariance can be written simply:

$$\begin{aligned} Cov(f(\theta, r), g)) &= E((g - \overline{g})(f(\theta, r) - \overline{f})) \\ &= E((g - \overline{g})f(\theta, r)) - \underbrace{E(g - \overline{g})}_{=0}\overline{f} \\ &= \frac{1}{N}\sum_{i=1}^{N}(g_i - \overline{g})f(\theta_i, r_i). \end{aligned}$$

This fairness-inducing objective is simple and intuitive, though we demonstrate drawbacks of objectives like this in Section 3. An advantage of this covariance is that it measures the correlation between outcome and type even when g is not binary-categorical, but is an arbitrary feature vector.

In [14], a learning algorithm is proposed that simultaneously aims to minimize between-group unfairness and within-group unfairness by ensuring that individuals in a protected group achieve similar outcomes to the general population, and that individuals with similar unprotected features receive similar outcomes, articulated through what they term a *Lipschitz condition*, proposed in their previous work [6].

In [6], rather than condensing inequality throughout a population to a single index, between-group equality is conceived of as a constraint on the distance between utility distributions conditioned on feature values. That is, statistical parity up to bias ϵ holds for groups S and T with distributions μ_S and μ_T if $D(\mu_S, \mu_T) \leq \epsilon$, where D is a distance metric on the space of probability distributions. They point out that the Lipscitz condition is stronger than statistical parity.

In [6], their goal is to propose a *fair affirmative action* scheme, which is similar in spirit to the goals of this paper. Their intermediate representation of individuals is also similar in spirit to our proposed extension of automating the fairness-inducing weight-assignment, as discussed in section 4.

2.2 Within-Group Unfairness

While between-group unfairness has prominent social and legal motivations, within-group unfairness is another important measure of equitable resource allocation, especially when groups are not well-defined or "protected". This notion is of unfairness is measured by the Gini coefficient and the similar McCloone index, and is the subject of the common idea of "inequality".

In some cases, within-group and between-group fairness are coth considered as objectives. In [10], the trade-off between the two is explored along with an objective called a "generalized entropy index" that captures both; in their context, reducing between-group unfairness is proven to be guaranteed to increase *within-group* unfairness. In [6] and [14], both are minimized through two novel algorithmic approaches, the latter of which, like our approach, minimizes a linear combination of objectives with coefficients governing the trade-off between fairness and overall utility (in their case, accuracy).

In [1] (cited repeatedly by [8]), the authors also propose the Nash standard of comparison (objectives (6) and (7)), which allocates resources to individuals who would experience the highest percent change in utility, and favors the worst-off. The authors describe the sum of logarithms objective as achieving a *propor-*

tional fairness criterion only if the space of possible utility distributions is convex; otherwise, more generally, this fairness criterion is satisfied by utility vector $\boldsymbol{u}$ if and only if, for any other utility vector $\boldsymbol{u'}$, the "aggregate proportional change" is negative:

$$\sum_{i=1}^{N} \frac{u_i' - u_i}{u_i} \leq 0 \; \forall \; \boldsymbol{u'} \in U, \tag{12}$$

where U represents the set of all feasible utility distributions.

In [11], the social welfare functions in the family (11) (which includes the utilitarian objective (5) at $\xi = 0$, the Nash objective (6) at $\xi = 1$, and the max-min objective (9) as $\xi \to \infty$) are applied to a network optimization problem to promote node equality, interrogating whether a higher level of fairness, attenuated by the parameter ξ, necessarily implies a lower global utility in terms of network throughput.

[2] is very close in spirit to this paper, as their central goal is to balance fairness and efficiency through multiple objectives. Their objectives are a utilitarian social welfare function and a max-min utility, while in section 4 we use a utilitarian solution and a subpopulation-restricted utilitarian solution. Ultimately, they are able to formulate as a MILP a "leximax" approach that iteratively yields as much utility as possible for the worst-off individual. We are concerned with the utility of a disadvantaged subpopulation and in fact advocate against always prioritizing the lowest-utility individuals in Section 3.

2.3 Price of Fairness

In [1], the authors propose that the fairness-neutral utilitarian objective (5) (with no fairness-inducing constraints) as a baseline for a "price of fairness". That is, given an optimum with respect to a fairness-neutral utilitarian objective $\boldsymbol{u}^*_{system}$ and another optimum with respect to a fairness-inducing problem formulation $\boldsymbol{u}^*_{fair}$, the price of fairness is defined as the percentage decrease in utilitarian objective (5):

$$POF = \frac{SWB(\boldsymbol{u}^*_{system}) - SWB(\boldsymbol{u}^*_{fair})}{SWB(\boldsymbol{u}^*_{system})}, \tag{13}$$

where $SWB(\boldsymbol{u}) = \sum_{i=1}^{N} u_i$. They axiomatically define a "fair" classifier as one that is *Pareto efficient*, in that solutions from a fair classifier cannot be dominated by another, and given this limitation, they can bound the price of fairness in some settings. In contrast, [12] reviews observations that several notions of fairness not be simultaneously satisfiable.

It is taken as a given in [6] and [14] that to achieve between-group fairness, classifiers must, to a point, systematically misclassify individuals from at least one group. A "cost of fairness" is defined in [4]; a "price of fairness" is defined in [1]; a "price of equity" is defined in [8]; "Balancing Fairness and Efficiency" is the subject of [2]. In [12], which focuses on linearly-separating classifiers (thresholded logistic regression and SVM), the rotation of a separating hyperplane to capture more true- *and false*-positives from one group concisely illustrates the trade-off between fairness and accuracy.

As noted in [12], legal attacks have succeeded against algorithmic decision-making aiming for proportional outcomes between groups, on the grounds that it encouraged "reverse-discrimination". They cite Ricci vs. Destefano, a 2009 case in which promoting Black firefighters was deemed unconstitutional due to their having scored lower on a standardized test than other candidates; the ruling enforced a policy that promotions go to those scoring in the top three of applicants, essentially outlawing any affirmative action effort. This type of resistance to equity-seeking algorithmic decision-making may make a fairness *constraint* hard to justify in a setting in which algorithmic decisions are immediately operationalized.

In some cases, enforcing a fairness *constraint* may result in an infeasible problem, or may degrade outcomes over the population in an unacceptable way, as explored in section 3. In these cases, and for a flexible and interpretable menu of solutions, we propose a simple re-weighting of individual utilities u_i in section 4.

3 Principles of Adverse Decisionmaking

The topic of fair algorithmic decision making has garnered recent attention in relation to advances in (and scrutiny of) machine learning. A common concern is the tradeoff between predictive accuracy and fairness [8, 1, 8, 4, 12, 14, 2, 7, 13]. In any context, parties may be loathe to sacrifice a notion of "accuracy", "efficiency" or "total utility". Still, we feel that when operational decisions, rather than predictors, are being made algorithmically, the principles of fair decision-making governing this tradeoff merit increased attention.

Before defining and requiring our own notion of Pareto efficiency in Section 4, we present several types of bad decisions that can be made by applying strategies for fair algorithmic decision-making from machine-learning literature. Our goal is to show that strategies proposed in the context of prediction may not be appropriate in the context of operations.

In this section we cover the striking pitfalls that can result from exclusively using a fairness-inducing objective. While this is certainly not proposed in any context in the literature, it explores the types of decisions encouraged by avoiding unfairness via the methods in the works above. These measures of fairness might be a term in an objective, an objective subjected to an accuracy constraint, one of multiple objectives including another measuring accuracy, or a constraint; these examples aim to help us understand the implications of any of those methods.

The central problem of this work is how to quantify *social welfare* from a collection of individual utilities and features. Utility $f(\theta, r)$ can be any function in (exogenous) individual features θ and (endogenous) resource allocation r. We do not rely on concavity, supermodularity, or the lack of either, in our problematization of fairness-seeking strategies. A wide range of conditions can result in "optimally" allocating disproportionate resources to *better-off* individuals, and "optimally" depriving others of zero-cost utility. Some of the examples do not rely on features at all, and only include one "type" of individual.

To illustrate these examples, however, we choose a specific family of utility functions, where individual utility is equated with "probability of success," and is given by the logistic model as in (1) with only one exogenous feature θ_i, representing type, as well as the endogenous r_i:

$$\begin{aligned} P(\text{success for individual } i) &= f(\theta_i, r_i) \\ &= \frac{1}{1+e^{-\beta_0-\beta_\theta\theta_i-\beta_r r_i}} \end{aligned} \tag{14}$$

where $\beta_\theta > 0$, meaning a higher type results in a higher success rate. Usually r_i represents a "distance assignment" and $\beta_r < 0$, meaning a lower distance to the nearest selected facility results in a higher success rate; in other examples, r_i is a generalized "resource allocation", and $\beta_r > 0$. The values of β_θ, β_r are changed slightly for each example, while $\beta_0 = 0$ is used in all; in the spatial examples, this is equivalent to the examples occurring over regions of different scales.

3.1 Taxonomy of Adverse Decisions

We define the following types of bad decision-making:

1. *Feature Ignorance.* Like all work addressing between-group inequality, we assert that "fair" decisions in the context of different subpopula-

tions requires acknowledgement of those subpopulations. Further, we consider that any available features, in addition to group affiliation, merit inclusion in a fair optimization model. The standard of "anti-classification" in machine learning seeks for subgroup-conditioned expected outcomes to be comparable, and our proposed methods for promoting fair between-group outcomes in Section 4 centrally rely on individuals' group identities to do so.

2. *Adverse Triage.* Almost all optimization problems seek to make decisions with high marginal utility (triage). We make a distinction between *feature-triage* and *instance-triage*. *Feature triage* refers to systematically prioritizing resources to individuals with a given protected feature. *Instance triage* refers to prioritizing resources based on unprotected features, possibly including or comprising aspects of the problem instance such as location and graph connectivity. Each type of triage has a possible adverse outcome.

    (a) *Unintended Triage.* Being "feature-aware" always creates the possibility of feature triage, which can either help or hurt a lower-utility subpopulation. It is possible for subgroups to have simultaneously higher utility and higher marginal utility than others due to the role of their features in the utility function. In this case, the feature-aware utility function ends up prioritizing the higher-utility subpopulation! Some social welfare functions avoid this by prioritizing low-utility individuals or promoting equality in the distribution of utilities. We believe the prioritization of one or another subpopulation should be a separate, socially-motivated decision by a social planner, rather than a by-product of systematic differences in the marginal utilities of different subgroups. In Section 4 we propose a standard of commitment to serving a given subpopulation.

    (b) *Anti-Triage.* We consider between-group fairness to be a first-order concern, and otherwise generally advocate for a utilitarian objective as a measure of social welfare. Within-group *equality* is not the main goal of the methods proposed in Section 4. Part of the reason why this is not our focus is that by overly prioritizing low-utility individuals, natural opportunities to serve individuals with high marginal utilities are lost, and we term this phenomenon *anti-triage*.

        For examples, if two individuals have utilities e^r and $100e^r$, the Nash bargaining solution would be indifferent to allocating resources to either individual; an appeal to max-min equality would favor the former and is literally insensitive to changes in the other individuals' utilities; and an appeal to utilitarianism would favor the latter. In contrast to *unintended triage*, which disproportionately allocates resources to one group due to the effect of their protected features on their utility, *anti-triage* can occur with or without feature-awareness due to a failure to apply *instance triage*. Rather than prioritizing the lowest-utility individuals, we advocate instead to prioritize systematically low-utility subpopulations, but otherwise to respect utilitarian decisions, and in Section 4 we provide tools to do so.

3. *Self-Sabotage* can result from some of the between-group equality-inducing objectives focused on in the machine-learning literature. For

example, in a discrete resource-allocation problem, it may be possible to meet the needs of both high-utility and low-utility individuals, but, in an effort to reduce the between-group inequality, the high-utility group will be denied available resources or resources will be wasted, depending on the problem formulation.

We strive for solutions on the efficient frontier of *population-wide social welfare* and *subpopulation social welfare*, measured by the same social welfare function, but restricted to a subpopulation with systematically or historically lower utility. There are two identifiable and distinct *self-sabotage* outcomes, both defined by deviations from this frontier.

(a) *Missed Opportunities.* Denying high-utility individuals available resources in order to reduce the disparity between high- and low-type individuals amounts to a *missed opportunity.* This would unnecessarily reduce population-wide welfare in the name of reducing inequality, moving adversely along the corresponding axis of the efficient frontier. This can characterize decision-making even in a *feature-ignorant* setting.

(b) *Spiteful Allocation.* Possibly more concerning is if many high-utility individuals *and* some low-utility individuals could benefit from a decision that is avoided in the name of reducing inequality: this we term *spiteful allocation* or *cutting off the nose to spite the face.* This moves adversely along *both* axes of the efficient frontier.

These examples may seem to argue against using these measures of fairness in general, but our intention is to highlight the differences between fair algorithmic decision-making in machine-learning and that in operations.

3.2 Adverse Continuous Resource Allocation

Consider a variant of (4) with the social welfare function (5):

$$\begin{array}{lll} \max & \sum_{i=1}^{N} u_i & \\ \text{s.t.} & u_i & = f(\theta_i, r_i) \\ & \sum_{i=1}^{N} r_i & = 1 \\ & r_i & \geq 0. \end{array} \tag{15}$$

3.2.1 *Unintended Triage* in Continuous Resource Allocation

The formulation (15) can encourage the exogenously lower-utility individuals to optimally receive no resources, as in the first pane of Figure 1. This is an example of *unintended triage*: individuals with a high marginal utility due to a protected feature are favored, and when those individuals also have high overall utility, feature-awareness widens inequality.

The KKT optimality condition for an interior point is $\frac{\partial}{\partial r_i} f(\theta_i, r_i) = \frac{\partial}{\partial r_j} f(\theta_j, r_j)$ for all i and j. However, the role of protected features in the utility function results in globally higher marginal utility one subgroup, and so the solution is on the boundary where those individuals receive all the resources.

In the second pane of Figure 1, the Nash bargaining solution (6) allocates more resources to the low-type individual, which could be seen as a remedy. In later examples, however, the Nash bargaining solution will be seen to result in *anti-triage.*

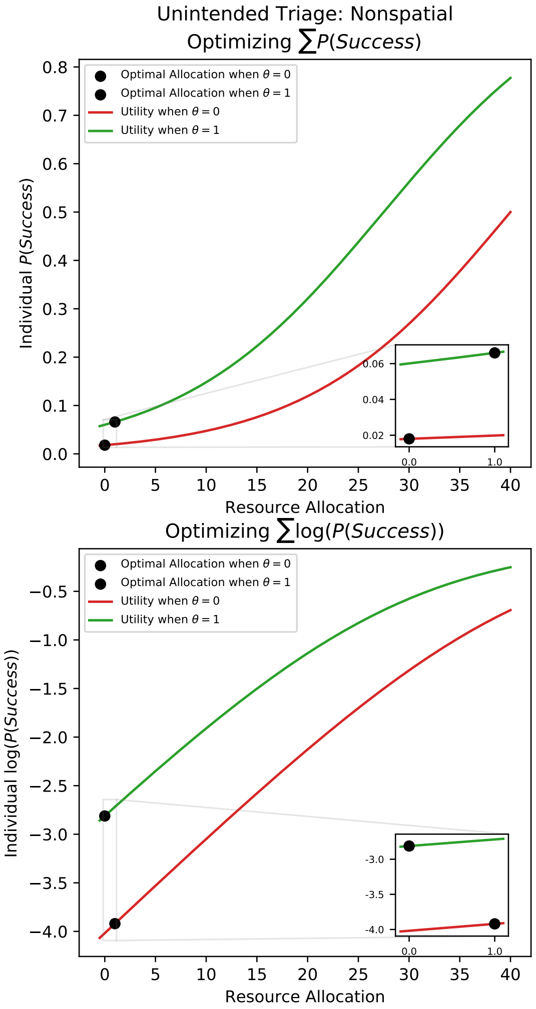


Figure 1: Maximizing the sum of utilities can result in *unintended triage.*

3.3 Adverse Discrete Resource Allocation

Unlike the continuous optimization problem in Section 3.2, discrete optimization problems do not have continuous domains, and thus do not have optimality conditions in terms of marginal utilities. Nevertheless, by visualizing discrete solutions along utility curves, the same phenomena of bad decisionmaking can be observed.

Several discrete optimization problems fall under the general *capacitated facility location* framework.

$$
\begin{array}{llll}
\max & \sum_{i=1}^{N}\sum_{j=1}^{M} f_{ij}x_{ij} & +\sum_{j=1}^{M} c_j y_j & \\
\text{s.t.} & \sum_{j=1}^{M} x_{ij} & = 1, & i = 1,\dots,N \\
 & \sum_{i=1}^{N} x_{ij} & \leq K_j y_j, & j = 1,\dots,M \\
 & \sum_{j=1}^{M} y_j & \leq m & \\
 & x_{ij} & \geq 0, & i = 1,\dots,N, \\
 & & & j = 1,\dots,M \\
 & y_j & \in \{0,1\}, & j = 1,\dots,M.
\end{array}
\tag{16}
$$

While a number of discrete problems are special cases of the capacitated facility location problem (16), certainly not all discrete optimization problems are part of this family. We will give examples from two special cases:

- *Maximum weighted bipartite matching*, which is a form of one-to-one matching, can also be written in the form of (16), where $K_j = 1$ and $c_j = 0$ for all j (facilities can only be matched to one individual and there is no facility-choice cost).

- The *uncapacitated facility location problem* (3), the main setting of this paper, can also be written in the form of (16), where f_{ij} represents the utility to individual i when assigned to location j (instead of $u_i = f(\theta_i, r_i)$), the facility capacity $K_j = N$ or $= \infty$ for all $j = 1, \dots, M$ (all facilities can accommodate arbitrarily many individuals), and facility cost $c_j = 0$. In the facility location problem, utilities f_{ij} depend on "distance to closest facility", and so also satisfy some correspondence to the triangle inequality, which could be leveraged via the resulting set submodularity (see Section 4.4.1) or other solution methods; whereas in the general form of (16), the f_{ij} need not satisfy any correspondence to the triangle inequality. Facility location is a one-to-many matching.

3.3.1 *Unintended Triage* in Maximum Weighted Bipartite Matching

We consider a variant of bipartite matching in which individual utility is equated with their "probability of success" and is measured as:

$$P(Success) = f(\theta, r) = \frac{1}{1 + e^{-\beta_0 - \beta_\theta \theta - \beta_r r}}$$

where θ is the individual type and r represents the *value* of the resource with which they are matched, and is equal to zero if they are not matched. A higher type and a higher-value resource match result in higher utility. Importantly, utility is not *entirely* determined by the matching decision, and is strictly positive, allowing an easy application of the Nash social welfare function (7), which is illustrated as a strategy that avoids *unintended triage*.

The first pane of Figure 2 demonstrates that the utilitarian objective (5) can result in high-type individuals being allocated more higher-value resources, and for lower-type individuals to be more likely to be un-matched, in an example of *unintended triage*. In constrast, the second pane demonstrates that social welfare function (7) favors exogenously low-utility individuals, the effect of which dominates the (unintended) *feature-triage*.

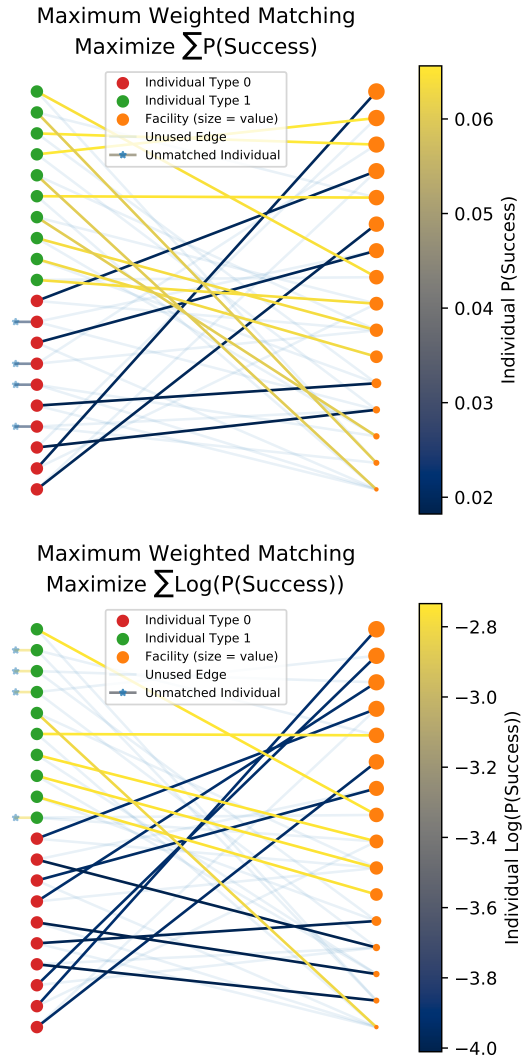


Figure 2: Utilitarian objective (5) results in exogenously higher-utility individuals being allocated disproportionately more resources, in an example of *unintended triage*. The equality-seeking social welfare function (7), by contrast, matches more low-type individuals to the most valuable resources, and leaves fewer of them un-matched.

3.3.2 *Unintended Triage* in Uncapacitated Facility Location

Figure 3 issues an important warning regarding feature-aware optimization. Maximizing the sum of individuals' probabilities of success, taking into account their features, may *seem* to promote fairness, by in fact avoiding *feature ignorance*. In this model, however the low type individual gains less utility from a close facility than the high type individual, and the sum of utilities is highest when benefiting the high type individual, in an example of *unintended triage*.

As in the continuous example, this issue is avoided, favoring the individual whose protected features yield a systematically lower utility, when maximizing the sum of the logarithms of the utilities (maximizing the Nash Standard of Comparison, as in [1, 8], and [2], which is equivalent to maximizing their product).

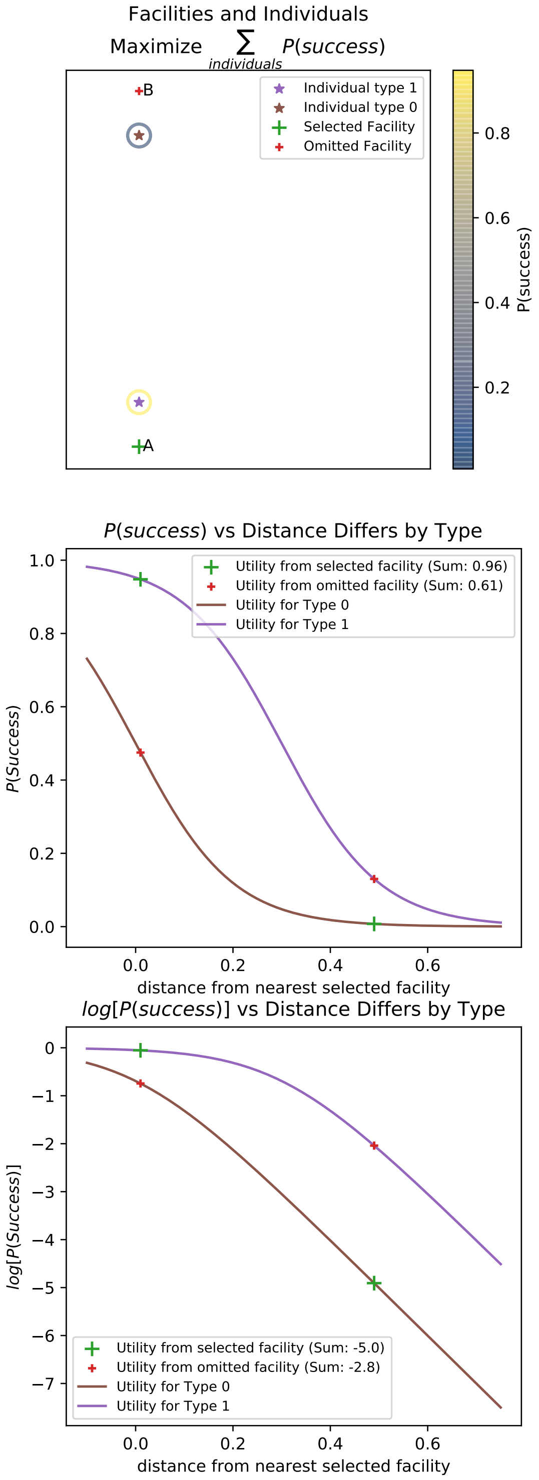


Figure 3: Incorporating individual features into decision-making is essential to fair decision-making, but can have exactly the opposite effect.

3.3.3 *Missed Opportunity* in Uncapacitated Facility Location with Covariance Minimization

In Figure 4, placing a facility at location B would yield improvements for all individuals compared to location A. However, the objective is *minimizing covariance between type θ and $f(\theta, r)$*, adapted from [13, 12] and [10]. As the higher-type individual has a much better outcome even at an inferior distance, the more "fair" solution is to hurt this individual. This is an example of *missed opportunity*, a variant of *self-sabotage*.

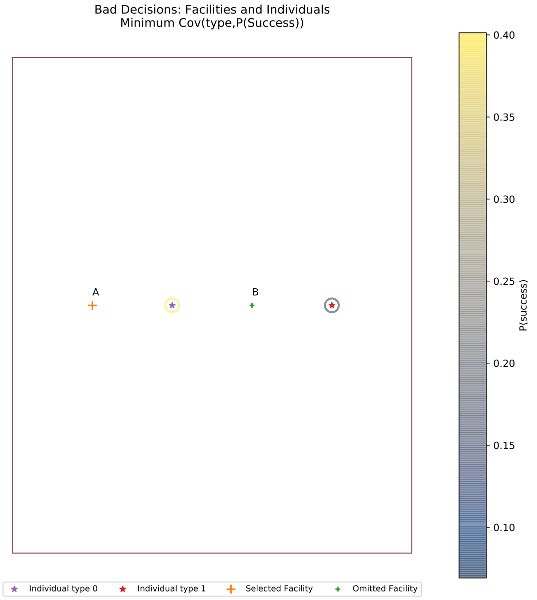


Figure 4: Minimizing covariance encourages a decision whose utility for the two individuals is dominated by the other option.

3.3.4 *Spiteful Allocation* in Uncapacitated Facility Location with Covariance Minimization

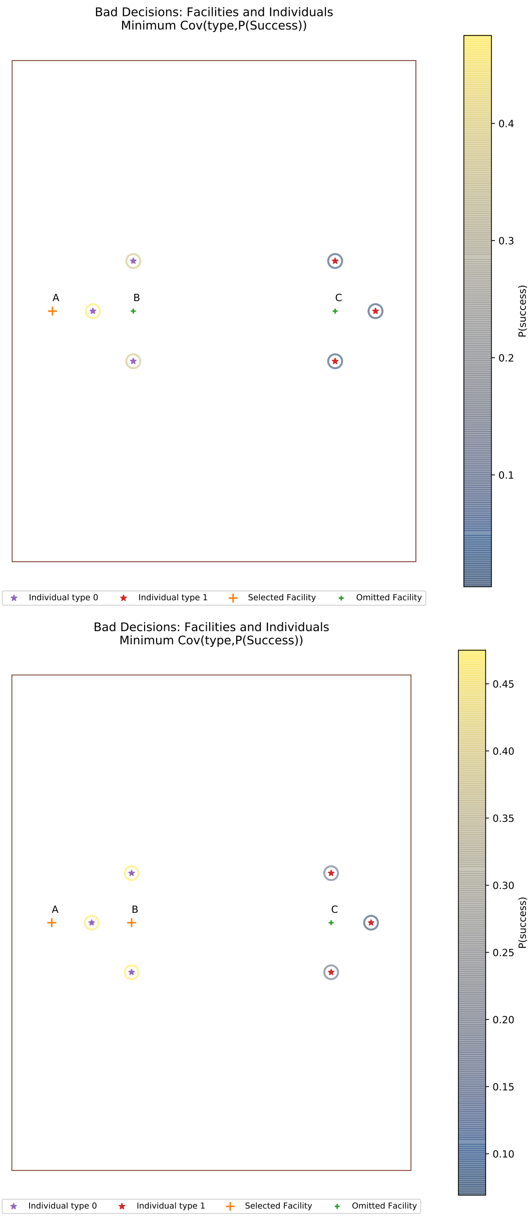


Figure 5: Minimizing covariance encourages decisions whose utility among individuals in both groups is dominated by other options.

In figure 5, again minimizing covariance results in *self-sabotage*, but in such a way that individuals from both groups are worse-off. When *one* facility is selected, it is facility A, which yields worse outcomes for all individuals than location B (including two low-type individuals). When *two* facilities are selected, location A is included even though it provides *no additional utility to any individuals* simply because the outcomes of the high-type individuals are already much better than those of the low-type individuals, and improving their outcomes further would increase inequality.

3.3.5 *Self-Sabotage* in Uncapacitated Facility Location with Generalized Entropy Minimization

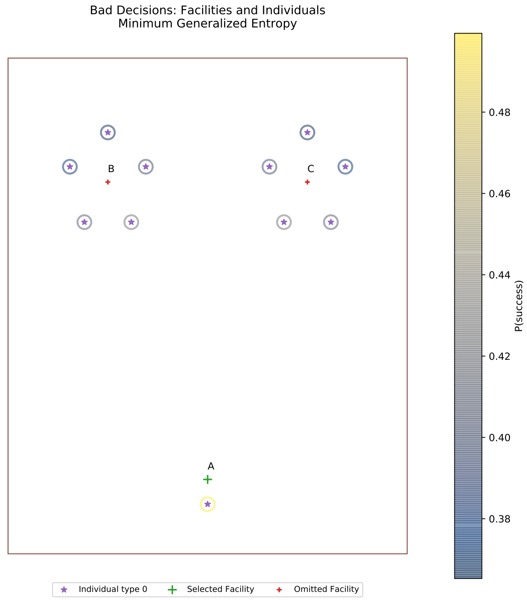


Figure 6: Minimizing generalized entropy encourages choosing a facility that results in the least possible average utility. This example has only one type of individual.

Figure 6 illustrates that individual features and/or group affiliation are not essential for encouraging *self-sabotage*. The objective is to minimize the *generalized entropy* function from [10]:

$$\mathcal{E}^{\alpha}(\mathbf{u}) = \frac{1}{N\alpha(\alpha-1)} \sum_{i=1}^{N} \left[\left(\frac{u_i}{\overline{u}} \right)^{\alpha} - 1 \right] \tag{17}$$

While the solution selecting the facility at location A is not *dominated* by those at B or C, the identical individuals near location B would outnumber the individual near location A and receive the same utility were the facility at location B chosen. While generalized entropy was used to produce this plot, minimizes any index that measures "inequality" would have this effect. That is, a Gini coefficient or McCloone index could produce this same phenomenon.

This example of *self-sabotage* is classified as a *missed opportunity*, but it could also be viewed as *spiteful allocation* in that the (majority of) individuals receiving lower-than-possible utility in the name of "equality" are of the same group as the individual near the facility at location A whose low utility is being avoided. This could be described as *cutting off the nose to spite the face.*

3.3.6 *Anti-Triage* in Uncapacitated Facility Location with Nash Bargaining Solution

In several examples in this section, the Nash bargaining solution, via objective (7) (sum of log-utilities), is presented as a remedy to *unintentional triage*, by favoring low-utility individuals when they are the victims of *feature-triage*. Unfortunately, this can prevent efficient *instance triage*, resulting in many fairly low-utility individuals instead of just one individual with utility that is even lower, but only by a small amount.

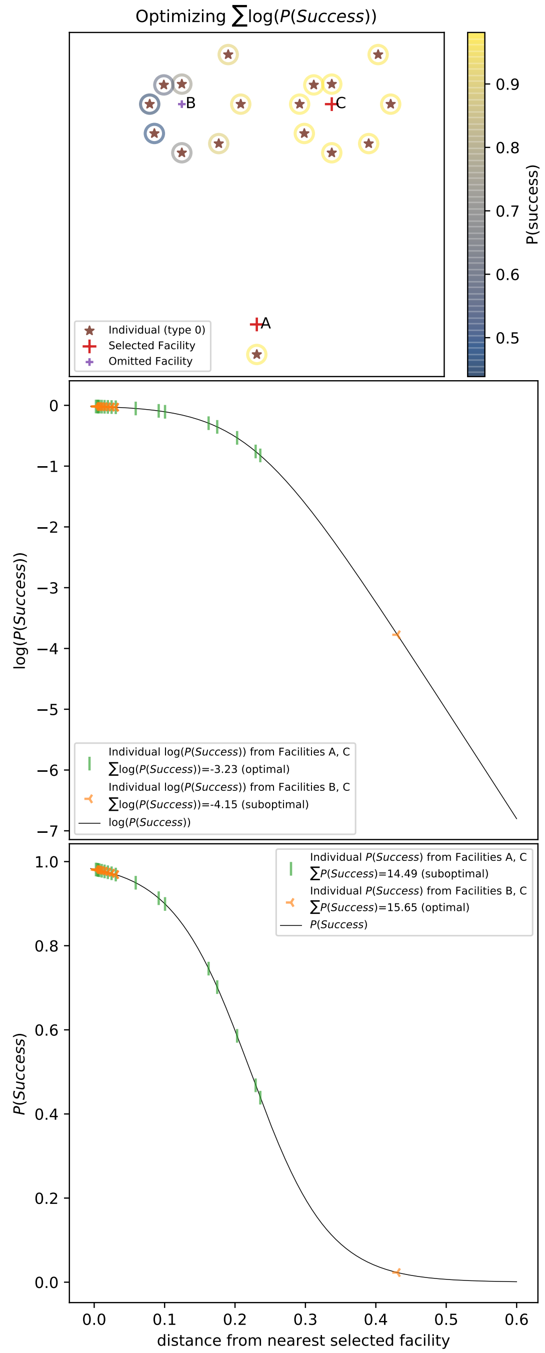


Figure 7: The Nash bargaining solution, which prioritizes the lowest-utility individuals, places facilities at locations A and C, rather than B and C, which would result in a greater utilitarian objective. We consider the over-investment in avoiding a low-utility individual to be *anti-triage*, which is in this case a failure to apply *instance triage*.

Failing to maximize the utilitarian objective is not automatically an example of *anti-triage*: indeed, decreasing the utilitarian objective in order to benefit a systematically lower-utility protected group is the central focus of this paper. In contrast, devoting undue resources to an individual who has a low utility in the utilitarian solution due to *unprotected features*, such as their location within a problem instance, is an example of *anti-triage*. We consider between-group equality to be a first-order concern, but otherwise advocate for a utilitarian objective rather than an equality-promoting one.

4 Commitment to α-Fairness

If one group is systematically disadvantaged, regardless of equality-seeking methods applied, the outcome of social planning will still likely fail to create equity. We propose that inequity, whether or not it can be equalized through decisionmaking, should motivate efforts to serve lower-utility individuals and/or subpopulations.

Instead of attempting to quantify and control a numerical index of inequality between subpopulations, we propose instead to focus on the *fraction of effort* devoted to serving the various subpopulations, and describe how to incorporate that into an objective. We describe the sense in which this produces "efficient" solutions with respect to utility in the population overall and within the subgroup.

Finally, we propose a heuristic in the same spirit, which finds approximate solutions while devoting a specified fraction of effort to different subpopulations.

4.1 Efficient Solutions

In Section 3, we criticize the application of objectives that:

- fail to usefully allocate available resources to high-utility individuals to promote equality, which can hurt either strictly the high-utility individuals (*missed opportunity*) or both high- and low-utility individuals (*spiteful allocation*);

- systematically triage resources away from a subpopulation identified by a protected feature (*unintended triage*); and
- fail to triage resources when appropriate (*anti-triage*).

These adverse outcomes result from:

- maximizing feature-aware utility without prioritizing lower-utility subpopulations identified by a protected feature
- optimizing a fairness index, such as
  - minimizing covariance between protected features and utility
  - minimizing an inequality index such as generalized entropy or a Gini coefficient
- optimizing a low-utility-averse transformation of individual utility, such as those in (11) for $\xi > 0$:

  $$SW_\xi(\boldsymbol{u}) = \begin{cases} \frac{1}{1-\xi} \sum_i u_i^{1-\xi} : \xi \geq 0, \xi \neq 1 \\ \sum_i \log(u_i) : \xi = 1, \end{cases}$$

  which include

  - the Nash standard of comparison (7) when $\xi = 1$
  - the max-min social welfare objective (9) when $\xi \to \infty$

We propose that "fair and efficient" solutions lie on the frontier of utility allocated to the entire population vs. utility allocated to a subpopulation identified by a protected feature.

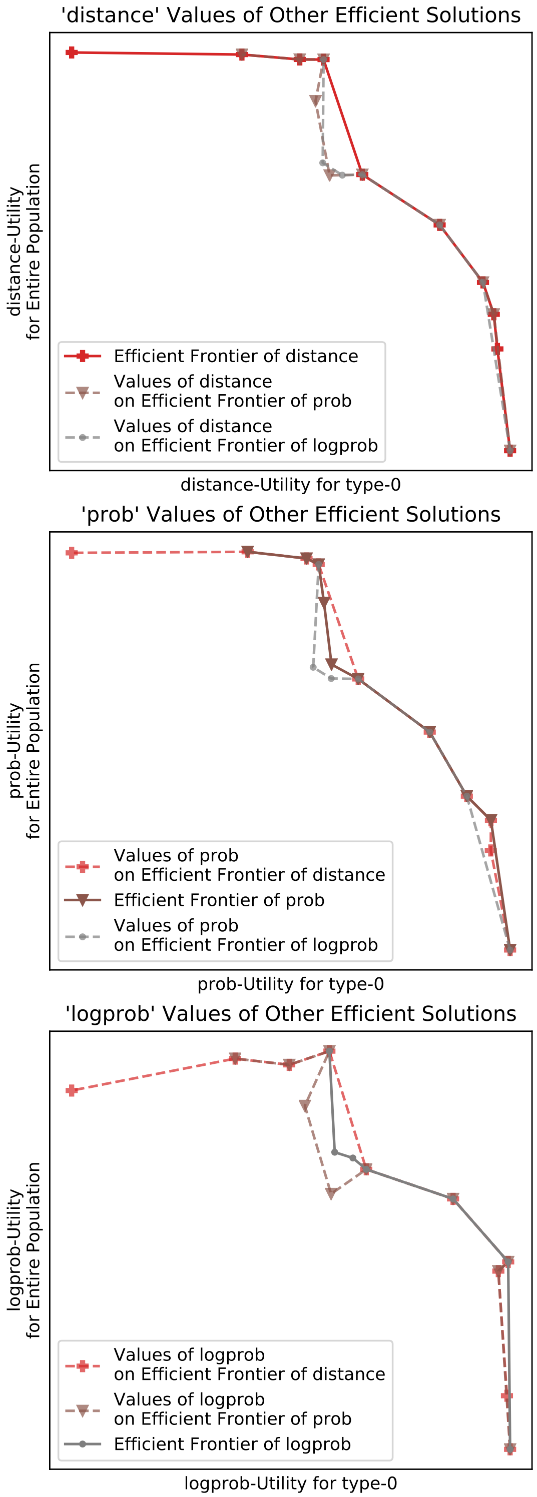


Figure 8: The frontier of optimal social welfare allocated to the entire population vs. social welfare allocated to the systematically lower-utility subpopulation.

The three panes of Figure 5 can be thought of as different strategies to promoting fairness. The first pane ('distance') is the efficient frontier of the feature-ignorant objective of minimizing the sum of distances between facilities and individuals. The second pane ('prob') is the efficient frontier of maximizing the sum of individuals' utilities ($u_i = f(\theta_i, r_i)P(Success|\theta_i, r_i)$), which are a function of their distance to the nearest facility as well as their features (social welfare function (5)). The third pane ('logprob') is the efficient frontier of maximizing the sum of the logarithm of individuals' utilities (social welfare function (7)).

Section 3 advocates against striving for the frontier in the first pane, as it is based on a *feature-ignorant* objective. Section 3 also takes issue with the low-utility-averse transformation applied to utility yielding the frontier in the third pane, which may focus too heavily on avoiding individuals with very low utility, regardless of their protected features. Instead, we advocate for seeking solutions on the efficient frontier in the second pane.

As should be clear from Figure 8, the efficient solutions with respect to one social welfare function may not be efficient with respect to another. Some efficient solutions may coincide across objectives, but this figure should illustrate that, for example, an efficient Nash bargaining solution may be "dominated" in the sense that some other solution yields a utilitarian improvement to both low-type individuals and the population as a whole.

4.2 α-Fairness in a Social Welfare Function

There is no a priori reason to prefer any of the solutions along the efficient frontier illustrated in the second pane of Figure 8, except perhaps the top-left solution, which is the utilititarian objective, which maximizes the sum of utilities across the entire population.

In the context of operations contracting, it is common for a parameter to denote an exogenous "negotiating power" between parties when a continuum of efficient contracts exists. Similarly, we consider the choice of which efficient solution is most appropriate to be an exogenous negotiation reflecting the urgency of prioritizing a subpopulation's welfare. In Section 4.3, we develop tools to inform this type of decisions.

Where along the frontier of "efficient" solutions a social planner decides is most appropriate can be described succinctly by a single parameter, which we call α, corresponding to a "fraction of effort" allocated to the subpopulation.

Consider a subpopulation defined by a binary protected feature $\theta = 0$, known as "type-0" individuals; for the rest of the population, known as "type-1", $\theta = 1$. The *α-fair utilitarian objective* is defined as follows:

DEFINITION 1. *α-Fair Utilitarian Objective*
For $\alpha \in [0, 1)$ and a utility distribution $\boldsymbol{u}$, the α-fair utilitarian objective is

$$SW_\alpha(\boldsymbol{u}) = \frac{1+\alpha}{2} \sum_{i:\theta_i=0} u_i + \frac{(1-\alpha)}{2} \sum_{i:\theta_i \neq 0} u_i. \quad (18)$$

DEFINITION 2. *α-Fair Efficient Solution*
For $\alpha \in [0, 1)$ and a utility distribution $\boldsymbol{u}$, the α-fair efficient solution is

$$\boldsymbol{u}^*_\alpha = \arg\max SW_\alpha(\boldsymbol{u}). \quad (19)$$

Note that the objective (18) is a form of the social welfare function (10) defined in Section 1.4. A higher value of $\alpha \in [0, 1)$ allocates an arbitrarily high priority to individuals in the type-0 subpopulation. At $\alpha = 1$, only the utility of type-0 individuals is included, and the objective becomes insensitive to changes in the utility of the rest of the population, which is no longer guaranteed to yield a solution on the efficient frontier of solutions in Figure 9, as are other values of α.

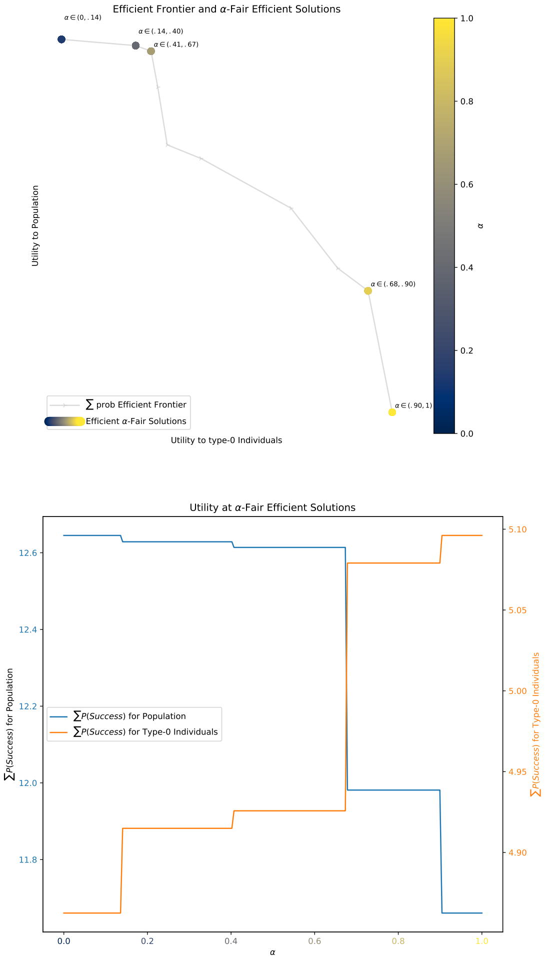


Figure 9: The first pane depicts the frontier of efficient solutions with respect to the social welfare of the entire population and the type-0 subpopulation, as well as the α-fair efficient solutions comprising the convex hull of the region bounded by the frontier and the origin. The second pane depicts the utility to the two subpopulations at solutions optimizing the α-Fair Utilitarian objective.

Theorem 1. *Solutions maximizing the α-Fair Utilitarian Objective are on the efficient frontier of subpopulation social welfare for all $\alpha \in [0, 1)$.*

Proof. Consider the alternative, a solution $\boldsymbol{u}$ maximizing (18) for some $\alpha \in [0, 1)$ for which another solution $\tilde{\boldsymbol{u}}$ exists that weakly improves both the social welfare of the type-0 subpopulation as well as to the type-1 subpopulation (or equivalently the population overall). Substituting $\tilde{\boldsymbol{u}}$ then weakly improves both terms $\frac{1+\alpha}{2}\sum_{i:\theta_i=0} u_i$ and $\frac{(1-\alpha)}{2}\sum_{i:\theta_i\neq 0} u_i$ in (18), contradicting the optimality of $\boldsymbol{u}$. □

Furthermore, many of the solutions on the efficient frontier of subpopulation social welfare can be obtained by maximizing the α-Fair Utilitarian Objective. In particular, all solutions on the *convex hull* of the region bounded by the origin and the efficient frontier can be found in this way. Note that the efficient frontier is not convex, and indeed not all efficient solutions can be found in this way, as can be seen in Figure 9. This convex hull limits the severity of the trade-off between subpopulation utilities to some extent, and we believe that this technique's simplicity outweighs its lack of comprehensiveness with respect to attaining every possible efficient solution.

Theorem 2. *All solutions on the convex hull of the efficient frontier of subpopulation social welfare are obtained by maximizing the α-Fair Utilitarian Objective SW_α for some $\alpha \in [0, 1)$.*

Proof. Consider the solutions on the convex hull of the region bounded by the efficient frontier and the origin, and refer to the summed utility to the "low-utility" subpopulation as L, and the summed population-wide utility as A ("all others"), denoted by $\boldsymbol{u} = (L, A)$, with A^* and L^* being the extreme values attainable via a utilitarian objective and an objective that fully prioritizes the utility of the low-utility subpopulation. For simplicity, consider a slight variant of SW_α, called $\tilde{SW}_\alpha$ that simply maximizes a convex combination of L and A (the result still holds for SW_α). This variant of the α-fair utilitarian objective is then simply written

$$\tilde{SW}_\alpha(\boldsymbol{u}) = \alpha L + (1-\alpha)A.$$

We prove the theorem by induction. To start, $\tilde{SW}_0$

and a sequence of $\tilde{SW}_\alpha$ with $\alpha \to 1$ will clearly attain the extreme solutions A^* and L^*, respectively. We then show that any solution that is "between" two solutions on this convex hull that are both attainable for values of $\alpha \in [0, 1)$ is also attainable for some value of $\alpha \in [0, 1)$.

Consider three consecutive points on the convex hull of the efficient frontier of subpopulation-wide and population-wide social welfare:

$$u_1 = \begin{bmatrix} L_1 \\ A_1 \end{bmatrix},$$
$$u_2 = \begin{bmatrix} L_2 \\ A_2 \end{bmatrix},$$
$$u_3 = \begin{bmatrix} A_3 \\ L_3 \end{bmatrix}.$$

Say these solutions run from "top-left" to "bottom-right" of the frontier (visualized in the first pane of Figure 9), and as they are *efficient*, none dominates any other:

$$L_1 \leq L_2 \leq L_3$$
$$A_1 \geq A_2 \geq A_3.$$

Furthermore, as these solutions are on the *convex hull* of the region bounded by this frontier and the origin, it must be the case that u_2 is on the "top-right" side of the line segment connecting u_1 and u_3.

Without loss of generality, assume $L_1 = 0 = A_3$ (or consider applying the orientation-preserving translation $L \mapsto L - L_1, A \mapsto A - A_3$). In this case, u_2 being to the "top-right" of the segment connecting u_1 and u_3 simply means:

$$A_2 \geq \frac{A_1}{L_3} L_2 + A_1$$

Now, for $\alpha = \frac{A_1}{A_1+L_3}$ (and $(1-\alpha) = \frac{L_3}{A_1+L_3}$), we evaluate the variant of the α-fair utilitarian objective at u_1, u_2, and u_3 to yield

$$\begin{aligned}
\tilde{SW}_\alpha(\boldsymbol{u}_1) &= \alpha \underbrace{L_1}_{=0} + (1-\alpha)A_1 \\
&= \frac{A_1 L_3}{A_1 + L_3} \\
\tilde{SW}_\alpha(\boldsymbol{u}_3) &= \alpha L_3 + (1-\alpha) \underbrace{A_3}_{=0} \\
&= \frac{A_1 L_3}{A_1 + L_3} \\
\tilde{SW}_\alpha(\boldsymbol{u}_2) &= \alpha L_2 + (1-\alpha) A_2 \\
&= \frac{A_1 L_2}{A_1 + L_3} + \frac{L_3}{A_1 + L_3} A_2 \\
&\geq \frac{A_1 L_2}{A_1 + L_3} + \frac{L_3}{A_1 + L_3}\left[\frac{A_1}{L_3} L_2 + A_1\right] \\
&= \underbrace{\frac{2A_1 L_2}{A_1 + L_3}}_{\geq 0} + \underbrace{\frac{A_1 L_3}{A_1 + L_3}}_{=\tilde{SW}_\alpha(\boldsymbol{u}_1)=\tilde{SW}_\alpha(\boldsymbol{u}_3)}
\end{aligned}$$

Since $\tilde{SW}_\alpha$ (for $\alpha = \frac{A_1}{A_1+L_3}$) attains a greater value at u_2 than u_1 or u_3 , $\boldsymbol{u}_2$ is the maximizer of SW_α.

By induction, all solutions along the convex hull of the efficient frontier, ranging between the solution that achieves L^* and the one that achieves A^*, are attained by maximizing $\tilde{SW}_\alpha$ for some $\alpha \in [0, 1)$.

As noted earlier, $\tilde{SW}_\alpha$ is simply a relabeling of SW_α to simplify notation during the proof, and the result holds for SW_α as well. □

Finally, we define α-fairness.

Definition 3. *α-Fairness*
For $\alpha \in [0, 1)$, let L_α be the utility to the (low-utility) protected subpopulation in the α-fair efficient solution *(which maximizes SW_α). An* α-fair *solution is any solution that achieves at least L_α utility for the protected subpopulation.*

For all experiments presented in this paper, solutions to discrete optimization problems were either found through exhaustive search or using COIN-OR's *BONMIN* mixed-integer nonlinear programming (MINLP) solver, neither of which addresses the NP-hardness of these problems in general.

We do not propose any algorithm for attaining α-fair solutions, which is in general as computationally hard as finding α-fair efficient solutions. In Section 4.4.1, however, we present an algorithm that is guaranteed to attain a fraction of L_α, and which shares the spirit of α-fairness, in that it devotes a *fraction of effort* to the protected subpopulation, subject to the social planner's social or ethical prioritization of benefiting that subpopulation.

4.3 Choosing α

We propose an extension of this work dedicated to automating the process of choosing an α that reduces between-group inequality without inverting and then exacerbating the between-group inequality. For now, we explore the effect of changing α on the population- and subpopulation-wide distribution of utilities in a facility location setting.

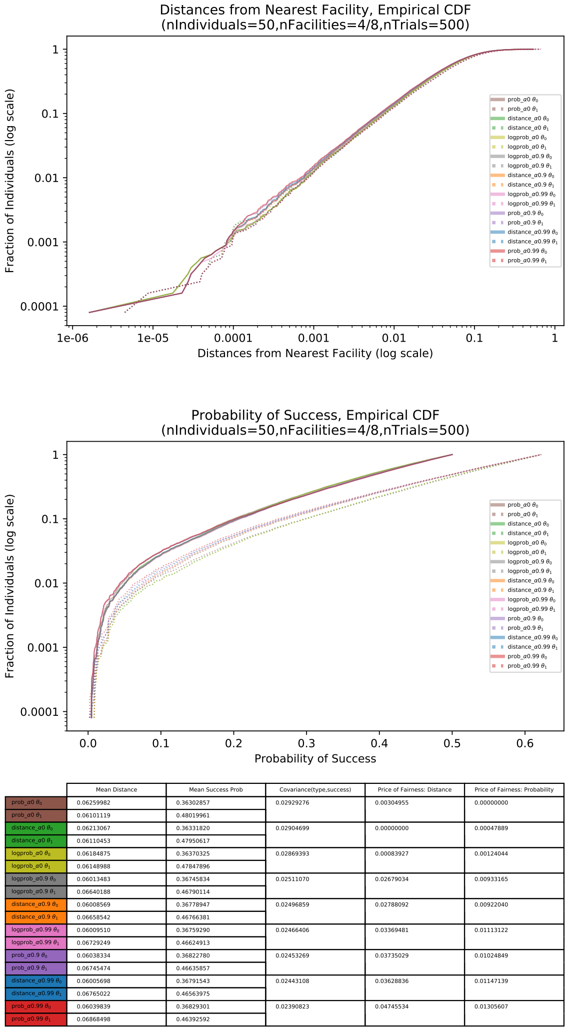
In the social context we wish to bear in mind, there are subpopulations whose systematically lower utility is likely to persist regardless of the effort of the social planner. The lower utility could be caused by a learned coefficient in a stochastic utility function (such as (1)) that predicts worse outcomes for members of that group, or due to some unprotected features that result in systematically worse outcomes (such as location).

Whatever the reason, one subpopulation is identified as being likely to suffer lower utilities than another, or than the population as a whole. A social planner elects to maximize the α-fair utilitarian objective SW_α to find an α-fair efficient solution.

What effect is this likely to have on the distribution of utilities among the population and among the protected subpopulation? How do the outcomes compare to a strictly utilitarian objective, a Nash bargaining solution, a feature-ignorant solution, or other α-fair efficient solutions?

We conducted numerical experiments to explore this question, maximizing several objectives over many trials of an uncapacitated facility location problem. In Figure 10, we plot the empirical CDF of outcomes for the low-utility subpopulation (marked θ_0) and the high-utility subpopulation (marked θ_1).

We also observed the result of objectives similar to the α-fair utilitarian objective, with α and $(1-\alpha)$ weights, but replacing utility with distance to closest facility (as in the feature-ignorant objective) and log utility (as in the Nash bargaining solution).



	Mean Distance	Mean Success Prob	Covariance(type,success)	Price of Fairness: Distance	Price of Fairness: Probability
prob_α0 θ_0	0.06259982	0.36302857	0.02929276	0.00304955	0.00000000
prob_α0 θ_1	0.06101119	0.48019961			
distance_α0 θ_0	0.06213067	0.36331820	0.02904699	0.00000000	0.00047889
distance_α0 θ_1	0.06110453	0.47950617			
logprob_α0 θ_0	0.06184875	0.36370325	0.02869393	0.00083927	0.00124044
logprob_α0 θ_1	0.06148988	0.47847896			
logprob_α0.9 θ_0	0.06013483	0.36745834	0.02511070	0.02679034	0.00933165
logprob_α0.9 θ_1	0.06640188	0.46790114			
distance_α0.9 θ_0	0.06008569	0.36778947	0.02496859	0.02788092	0.00922040
distance_α0.9 θ_1	0.06658542	0.46766381			
logprob_α0.99 θ_0	0.06009510	0.36759290	0.02466406	0.03369481	0.01113122
logprob_α0.99 θ_1	0.06729249	0.46624913			
prob_α0.9 θ_0	0.06038334	0.36822780	0.02453269	0.03735029	0.01024849
prob_α0.9 θ_1	0.06745474	0.46635857			
distance_α0.99 θ_0	0.06005698	0.36791543	0.02443108	0.03628836	0.01147139
distance_α0.99 θ_1	0.06765022	0.46563975			
prob_α0.99 θ_0	0.06039839	0.36829301	0.02390823	0.04745534	0.01305607
prob_α0.99 θ_1	0.06868498	0.46392592			

Figure 10: Empirical CDF of utility and feature-ignorant distance from nearest facility across the population

The motivating setting described in the beginning of this section is when the welfare of the low-utility subpopulation is so much lower than that of the high-utility subpopulation that no amount of *fairness* can "overshoot" between-group equality, and this experiment was tuned as such. As can be seen in Figure 10, the outcomes for the subpopulation with $\theta = 1$ are significantly better than those of the subpopulation with $\theta = 0$ even after optimizing *all* of the fairness-seeking objectives: the empirical distributions of utility nearly stochastically dominate.

What is notable in Figure 10 is that every one of the "fairness-" or "equality-seeking" objectives ends up inducing a *worse* distance distribution for the high-utility subpopulation than for the low-utility subpopulation. That is, while utility is determined by both *resource allocation* (r_i) and *type* (θ_i), any feature-aware equality-seeking objective will identify the need to allocate more resources to the population with systematically lower utilities; any fairness-seeking objective, like SW_α or a similarly weighted version of the Nash bargaining solution objective (7), will also prioritize allocating resources for this subpopulation.

As should be expected, a high value of α results in utility distributions that are *closer* together. The objectives are ordered in the legend and the table according their performance with respect to minimizing the covariance between type and outcome, and $SW_{0.99}$ was more effective than an equally-weighted *log-utility* objective. This is because the log-utility prevents effective triaging of resources in general (*anti-triage*), and in particular to the low-utility subpopulation, as discussed in Section 3.

For each objective, we computed the "price of fairness" as in [1], which is the relative loss in the utilitarian objective due to instead maximizing a fairness-seeking objective. We also computed this relative loss with respect to the solution that minimizes the feature-ignorant sum of distances, which is a common objective in operations. As should be expected, more "fair" solutions (with respect to covariance between type and outcome) tended to have a higher price of fairness, however this was not always the case, as $SW_{0.9}$ has better fairness but a lower cost of fairness than the 0.9-weighted log-probability function. This is because the cost of fairness was computed with respect to a utilitarian objective, which the log-probability function's optimum was not well-suited to maximize.

The 0.99-weighted feature-ignorant distance-minimizing objective performed exceptionally well. By adding subpopulation-specific weights, the objective can no longer be called *feature-ignorant*, but it's still interesting that it outperformed several other solutions that maximized a (weighted) sum of utilities *without considering utilities.* This is promising, as it may not always be feasible consider individuals' features for reasons or privacy or data availability, nor is it likely that a feature-aware utility function will be justified or reliable. Furthermore, while off-the-shelf mixed-integer linear programming (MILP) solvers are able to handle extremely large problems, mixed-integer *nonlinear* programming (MINLP) solvers are often not; the sum-of-weighted-distances objective is a MILP, while every other objective is a nonlinear (and not always convex) MINLP. Simply knowing (or guessing) individuals' group affiliation and solving a classically-studied, thoughtfully weighted MILP is likely a good strategy for achieving many of the goals of this paper.

4.4 Between-Group *Approximate* α-Fairness Heuristic

As discussed in Section 4.2 (fairness criteria), in the group-fairness setting we aim to promote solutions on the efficient frontier of *population-wide* and *subpopulation-wide* social welfare, while controlling the *fraction of effort* devoted to different subpopulations. We defined a family of α-fair utilitarian social welfare functions (SW_α) whose fairness is attenuated by a parameter $\alpha \in [0, 1)$.

Here we propose a heuristic approach that shares the spirit of the α-fair utilitarian social welfare functions, in that a parameter, which we choose to also call α, attenuates the tradeoff of *effort* for the protected subpopulation and the rest of the population. Unlike in the social welfare function, where "effort" represents a weight, in this case effort truly refers to a fraction of greedy decisions dedicated to serving a subpopulation as effectively as possible.

4.4.1 An α-Fair Heuristic

A broad class of combinatorial optimization problems admit a greedy approximation performance guarantee by the classification of their objectives as *monotone* and *submodular*, and this facility location problem is no different.

Consider a finite set U and a set function $f : 2^U \to \mathbb{R}$, where 2^U represents the power set of U. Then f is submodular if for any $S \subseteq T \subseteq U$ and $x \in U \setminus T$,

$$f(S \cup \{x\}) - f(S) \geq f(T \cup \{x\}) - f(T), \tag{20}$$

meaning the *marginal benefit of adding elements is decreasing*. In addition, f is *monotone* (monotone-increasing) if $f(S) \leq f(T)$.

In the facility-location setting, let $f(\theta_i, r_i)$ denote the benefit to individual i of being located distance r_i from the nearest chosen facility (which is decreasing in r_i). Consider the set of possible facilities U and two subsets of selected facilities $S \subseteq T \subseteq U$, along with d_{ij}, the distance between individual i and facility j. Then the corresponding set function $\tilde{f}_i : 2^U \to \mathbb{R}$

$$\tilde{f}_i(S) = f(\overbrace{(\min_{j \in S} d_{ij})}^{=r_i(S)}, \theta_i)$$

is submodular.

THEOREM 3. *The set function $\tilde{f}_i(S)$ is submodular.*

Proof. Since $\tilde{f}_i$ is clearly monotone, consider two cases:

1. $\tilde{f}_i(S) = \tilde{f}_i(T)$

   In this case, $\tilde{f}_i(S \cup \{x\}) = \tilde{f}_i(T \cup \{x\})$ because the closest facility to individual i in T is no closer than the closest facility in S; considering a potentially closer facility will either improve both or neither. Then the two sides of (20) are equal, and the condition holds.

2. $\tilde{f}_i(S) < \tilde{f}_i(T)$

   Here there are three sub-cases.

   (a) $\tilde{f}_i(S \cup \{x\}) = \tilde{f}_i(S)$

   This implies there is a facility in S closer to individual i than x. Since $S \subseteq T$, this implies $\tilde{f}_i(T \cup \{x\}) = \tilde{f}_i(T)$ as well. In this case, condition (20) reads $0 \geq 0$, and thus holds.

   (b) $\tilde{f}_i(S \cup \{x\}) > \tilde{f}_i(S)$ and $\tilde{f}_i(T \cup \{x\}) = \tilde{f}_i(T)$

   This implies there is a facility in T closer to individual i than facility x, but none of the facilities in S are closer to individual i than facility x. In this case, the left-hand side of condition (20) is $\tilde{f}_i(S \cup \{x\}) - \tilde{f}_i(S) > 0$ and the right-hand side is $\tilde{f}_i(T \cup \{x\}) - \tilde{f}_i(T) = 0$, so the condition holds.

   (c) $\tilde{f}_i(T \cup \{x\}) > \tilde{f}_i(T)$

   This implies that facility x is closer to individual i than any of those in T. Since $S \subseteq T$, x is also closer than any in S, and $\tilde{f}_i(S \cup \{x\}) > \tilde{f}_i(S)$. In particular, however, it means that

$$\tilde{f}_i(T \cup \{x\}) = \tilde{f}_i(\{x\}) = \tilde{f}_i(S \cup \{x\}).$$

Then, condition (20) holds by the following:

$$\begin{gathered}
\tilde{f}_i(T) \geq \tilde{f}_i(S) \quad \text{(monotonicity)} \\
\Downarrow \\
\tilde{f}_i(\{x\}) - \tilde{f}_i(S) \geq \tilde{f}_i(\{x\}) - \tilde{f}_i(T) \\
\Downarrow \\
\tilde{f}_i(S \cup \{x\}) - \tilde{f}_i(S) \geq \tilde{f}_i(T \cup \{x\}) - \tilde{f}_i(T).
\end{gathered}$$

So, in all cases, condition (20) holds, and so the set function $\tilde{f}_i$ is submodular. □

Note that the sum of submodular set functions is also a submodular set function, so the objective of a facility location problem of the form $\sum_{i=1}^{N} \tilde{f}_i(S) = \sum_{i=1}^{N} f(\theta_i, r_i(S))$ is also submodular.

All monotone submodular functions admit a $(1 - \frac{1}{e})(\approx 0.632)$ optimality guarantee for a greedy solution. As noted in [9], the $(1 - \frac{1}{e})$ guarantee follows from the definition of submodularity and greediness.

Theorem 4. *A nonnegative, monotone, submodular set function permits a $(1 - \frac{1}{e})$ greedy guarantee.*

Proof. Consider any monotone submodular function f and a greedy algorithm that creates sets $S_0 = \{\}$ and iteratively adds elements $S_{i+1} = S_i \cup \{x_{i+1}\}$ where $x_{i+1} = \arg\max_x f(S_i \cup \{x\})$. Let $S^* = \arg\max_{S:|S|=m} f(S)$ denote the cardinality-m maximizer of f.

Let $S^* = \{y_1, \ldots, y_m\}$ and let x_i be the greedy choices. Now note

$$\begin{aligned}
& f(S^*) \leq f(S_i \cup S^*) \\
& = f(S_i) + (f(S_i \cup \{y_1\}) - f(S_i)) \\
& \;\; + (f(S_i \cup \{y_1, y_2\}) - f(S_i \cup \{y_1\})) \\
& \;\; + \ldots \\
& \;\; + (f(\underbrace{S_i \cup \{y_1, \ldots, y_m\}}_{=S_i \cup S^*}) - f(S_i \cup \{y_1, \ldots, y_{m-1}\})) \\
& \leq f(S_i) + (f(S_i \cup \{y_1\}) - f(S_i)) \\
& + \;(f(S_i \cup \{y_2\}) - f(S_i)) \\
& + \ldots \\
& \;\; + (f(S_i \cup \{y_m\}) - f(S_i)) \;\; \text{(by submodularity)} \\
& \leq f(S_i) + m(f(S_i \cup \{x_1\}) - f(S_i)) \;\; (x_i \text{ are greedy}) \\
& = f(S_i) + m(f(S_{i+1}) - f(S_i)),
\end{aligned}$$

which shows

$$f(S_{i+1}) - f(S_i) \geq \frac{1}{m}(f(S^*) - f(S_i)).$$

Now, this implies

$$f(S_i) \geq (1 - (1 - \frac{1}{m})^i) f(S^*),$$

which is demonstrated inductively. First, $f(S_0) = f(\{\}) = 0 = (1 - (1 - \frac{1}{m})^0) f(S^*)$. Then,

$$\begin{aligned}
& f(S_{i+1}) \geq f(S_i) + \frac{1}{m}(f(S^*) - f(S_i)) \\
& \;\; = (1 - \frac{1}{m}) f(S_i) + \frac{1}{m} f(S^*) \\
& \;\; \geq (1 - \frac{1}{m})(1 - (1 - \frac{1}{m})^i) f(S^*) + \frac{1}{m} f(S^*) \text{(induction)} \\
& \;\; = (1 - (1 - \frac{1}{m})^{i+1}) f(S^*).
\end{aligned}$$

This implies that $f(S_m) \geq (1 - (1 - \frac{1}{m})^m) f(S^*)$, and since the sequence $(1 - \frac{1}{m})^m$ is increasing and approaching $\frac{1}{e}$, this implies

$$f(S_m) \geq (1 - \frac{1}{e}) f(S^*) \approx 0.63 f(S^*).$$

□

The corollarly simply follows from $\tilde{f}_i$ being nonnegative and monotone-increasing.

Corollary 1. *The facility location set function $\tilde{f}_i(S)$ permits a $(1 - \frac{1}{e})$ greedy guarantee.*

Now, this result implies a greedy heuristic for the proposed α-fair efficient solution to SW_α defined in (18), for any α! Being a nonnegative linear combination of nonnegative, monotone, submodular functions, SW_α is also nonnegative, monotone, and submodular.

But the objective value of SW_α is not as meaningful as the two objectives it balances, attenuated by "fairness-commitment" or "effort-level" α. In the same interest of balancing "commitment" towards multiple subpopulations, we propose a heuristic that does exactly that through greedy choices.

Definition 4. *α-Fair Greedy Solution.*
The α-fair greedy solution *is that attained by making the* first $\lfloor (1-\alpha)m \rfloor$ *greedy choices that maximize population-wide social welfare, and the* last $\lceil \alpha m \rceil$ *greedy choices that seek to maximize social welfare of the protected subpopulation.*

Suppose the maximum achievable social welfare for the entire population is A^*, and the maximum achievable social welfare for the protected subpopulation is L^*. Without much loss in generality, suppose α is some fraction of the number of selected facilities m, so αm and $(1-\alpha)m$ are integers. Then the α-fair greedy solution, with subpopulation utilities $u_G^\alpha = (L_G^\alpha, A_G^\alpha)$, simultaneously guarantees $\alpha(1-\frac{1}{e})L^*$ and $(1-\alpha)(1-\frac{1}{e})A^*$.

This is true because, on their own, any αm greedy choices dedicated to serving the protected subpopulation would achieve at least $\alpha(1-\frac{1}{e}L^*$, as each remaining greedy choice would add diminishing social welfare returns to the eventual $(1-\frac{1}{e})$ guarantee when $\alpha = 1$, following from the submodularity discussed above. Making these greedy choices *in addition to* the other $(1-\alpha)m$ choices, which may also improve the welfare of this subpopulation, assures the guarantee. The proof for A_G^α is the same. We define the solution as making the protected subpopulation's decisions *last* as this can only improve the outcome for that subpopulation, which is our central focus.

As should be apparent, these two guarantees will be far from tight in most problem instances, as can be seen in Figure 11. The efficient frontier discussed in the previous section is far from the guarantees for the α-fair greedy solutions.

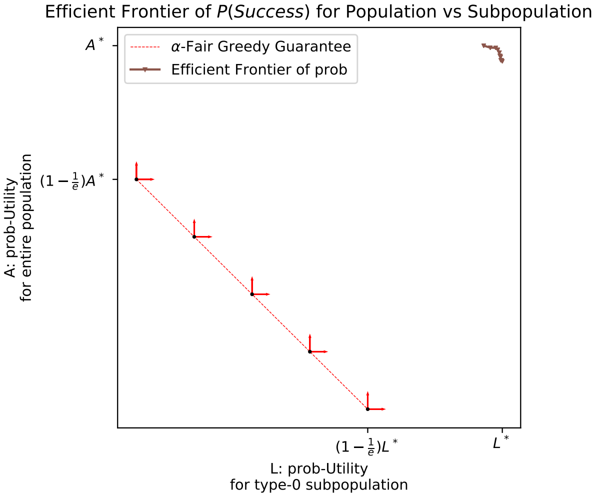


Figure 11: An α-fair greedy solution is guaranteed to lie to the upper-right of the red dashed line. In particular, a solution is guaranteed in one of the quadrants bounded below by red arrows, attenuated by α.

One notable feature of the guaranteed frontier visualized in Figure 11 is that the relative loss to the two interested subpopulations along the efficient frontier is extremely small compared to the suboptimality ratio $(1-\frac{1}{e})$. This means that the tradeoffs necessitated by choosing among α-fair *efficient* solutions are far smaller than those typically associated with approximate solutions to computationally hard problems.

5 Conclusion

Much of the literature on fairness in machine learning focuses on either between-group equality or within-group inequality. When a systematic disparity in utilities between subpopulations exists, the two approaches largely coincide. We point out ways in which objectives that reduce within-group inequality can fail to efficiently serve one or both segments of a two-group population, and can promote seemingly pathological decisions. In particular, we build a case against low-utility-averse objectives, such as the Nash bargaining solution (sum of log-utilities), as they generally fail to triage resources in favor of serving individuals who would otherwise experience the lowest possible utility. Unfortunately, when aiming to reduce inequality throughout an entire population, it's possible to under-serve members of both high-utility and low-utility subpopulations.

We visit a classical facility location problem, incorporate feature-aware individual utilities, and point out that the typical utilitarian objective (the sum of individual utilities) can on its own exacerbate between-group inequality by triaging resources to individuals in the subpopulation that already experiences systematically high utility. In general, while considering individuals' features is the only way to address systemic inequality between groups defined by protected features, there is no assurance that doing so will reduce between-group inequality.

In a two subpopulation case, we define *efficient and fair* solutions as those on the Pareto frontier of optimal social welfare for a low-utility subpopulation and that of everyone else. We provide an extremely simple modification of the typical utilitarian objective that is guaranteed to attain an efficient and fair solution, and is attenuated by a single parameter $\alpha \in [0, 1)$ via which many of these solutions can be attained. We define α-fairness as serving the low-utility subpopulation to at least the level of the α-fair efficient solution. This provides a standard for applying affirmative action, and we think of α as a "level of effort" devoted to fairness. We argue that when between-group disparities in welfare are considerable, it may be justified to commit to a high level of effort for fairness, especially when between-group inequality will persist even after allocating resources in this way. We provide a heuristic for applying a given level of effort to fairness in approximate solutions to the facility location problem.

There are many natural ways to automate the process of choosing α that we leave to future work. This parameter should reflect differences in welfare between groups: when inequality is larger, a larger α is justified. There are two cases to consider: when between-group inequality is very large and when it is moderate. When between-group inequality is large, a natural way to select α might be to measure the disparity between groups by calculating individuals' utilities based on some "worst-case average" for each group, by assigning dummy values for resource allocations to individuals from the two groups. This technique is used to evaluate heuristic performance in facility location problems in [5]. In the other case, where there is only moderate between-group inequality that might be eliminated through resource allocation, it would be natural to consider an iterative process that finds a *fair and efficient* solution that minimizes this disparity using binary search over the domain of α.

We also leave to future work the prospect of extending the study of *fair and efficient* solutions and *α-fairness* to more than two subpopulations. The *α-fair utilitarian objective* simply re-weights individuals' utilities. In a more general framework, α-fairness could be referred to as $(\frac{1+\alpha}{2}, \frac{1-\alpha}{2})$-fairness (as those are the weights applied to the two subpopulations in our scheme, and an extension to k groups could be $\boldsymbol{\alpha}$-fairness, for some $\boldsymbol{\alpha} \in \mathbb{R}^k_{++}$ that denotes the coefficients applied to the objective terms for individuals in the different subpopulations. What remains to be explored is treatment of individuals in the intersections of multiple subpopulations whose equal welfare is being sought. In *critical race theory*, a framework developed in law, sociology, and ethnic studies, the term *intersectionality* describes the reality that the intersection of political and social identities can produce unique - rather than additive - experiences of marginalization or privilege. Directives to improve outcomes for several systematically lower-utility subpopulations would need to be transformed to separately consider any and all intersections of those subpopulations in non-obvious ways.

References

[1] Dimitris Bertsimas, Vivek F. Farias, and Nikolaos Trichakis. The price of fairness. *Operations Research*, 59(1):17–31, 2011.

[2] Violet Xinying Chen and J. N. Hooker. Balancing fairness and efficiency in an optimization model, 2020.

[3] Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *CoRR*, abs/1808.00023, 2018.

[4] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 797–806, New York, NY, USA, 2017. Association for Computing Machinery.

[5] Gerard Cornuejols, Marshall Fisher, and George L. Nemhauser. On the uncapacitated location problem**this research was supported by nsf grants eng75-00568 and soc-7402516. sections 1–4 of this paper include a technical summary of some results given in [2]. some proofs are omitted and may be obtained in [2]. In P.L. Hammer, E.L. Johnson, B.H. Korte, and G.L. Nemhauser, editors, *Studies in Integer Programming*, volume 1 of *Annals of Discrete Mathematics*, pages 163 – 177. Elsevier, 1977.

[6] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, page 214–226, New York, NY, USA, 2012. Association for Computing Machinery.

[7] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. Fairness-aware learning through regularization approach. pages 643–650, 12 2011.

[8] Fabbri Marco and GC Britto Diogo. Distributive Justice, Public Policies and the Comparison of Legal Rules: Quantify the "Price of Equity". *Review of Law & Economics*, 14(3):1–23, November 2018.

[9] Sewoong Oh. Submodular function optimization. University Lecture, IE 512: Graphs, Networks, and Algorithms, 2013.

[10] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P. Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. A unified approach to quantifying algorithmic unfairness: Measuring individual and group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, KDD '18, page 2239–2248, New York, NY, USA, 2018. Association for Computing Machinery.

[11] A. Tang, J. Wang, and S. H. Low. Is fair allocation always inefficient. In *IEEE INFOCOM 2004*, volume 1, page 45, 2004.

[12] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness beyond disparate treatment and disparate impact. *Proceedings of the 26th International Conference on World Wide Web*, Apr 2017.

[13] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness constraints: Mechanisms for fair classification, 2015.

[14] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 325–333, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.