# Fairness without Regret*

**Marcus Hutter**

DeepMind & ANU

http://www.hutter1.net/

11 July 2019

**Abstract**

A popular approach of achieving fairness in optimization problems is by constraining the solution space to "fair" solutions, which unfortunately typically reduces solution quality. In practice, the ultimate goal is often an aggregate of sub-goals without a unique or best way of combining them or which is otherwise only partially known. I turn this problem into a feature and suggest to use a parametrized objective and vary the parameters within reasonable ranges to get a *set* of optimal solutions, which can then be optimized using secondary criteria such as fairness without compromising the primary objective, i.e. without regret (societal cost).

**Contents**



**Keywords**

utility, objective, optimal, fair/equitable/just, cost/regret, uncertainty.

*Regret is meant here in a mathematical sense, and fairness could be replaced by various other criteria.

# 1 Introduction

We consider the problem of optimizing a primary objective while also caring about a second criterion. Before introducing our model, we need to clarify terminology: The words "(sub)optimal", "best", "solution quality", "regret", "(ir)relevant", "(in)comparable" will always refer to the *primary* objective, henceforth called "objective". On the other hand, "fair", "just", "equitable" will always refer to the *secondary* criterion. If some of the latter aspects are relevant to the primary objective, they should be incorporated there. While in practice there are differences in meaning of fair/just/equitable and possibly treatment, the difference does not matter in formalizing our basic idea, so we use the terms interchangeably.

We consider the problem of (automated) decision-making based on some (primary) objective $U : S \to \mathbb{R}$. Optimal solutions $s^* := \arg\max_s U(s)$ sometimes appear to be unfair or unjust or not equitable. A popular approach of achieving fairness or equality is by constraining the solution space $S$ [ZVRG17, ABD+18]. Sometimes (primary-objective) irrelevant attributes such as gender are used (e.g. admit the best students, but constrained by selecting at least 30% women). Diversity arguments have more force, if based on objective-relevant attributes, e.g. diversity in thinking or skills, rather than diversity in looks or genes. In the former case, diversity is (only) an instrumental goal: If diversity indeed improves whatever the ultimate goal is, then it could in principle (already) be accounted for in the to-be-optimized objective, although operationally it may be easier to treat it as a constraint. If diversity does not positively correlate with the ultimate goal, but is desirable for other reasons, it can be modeled as a secondary objective or constraint. This constraining-of-solution-space by (esp. irrelevant) factors is a popular approach, which unfortunately reduces solution quality [MW18].[1]

In practice, the ultimate goal is often a (possibly non-linear) aggregate of subgoals, e.g. "life goals" include food, shelter, family, education, entertainment, health, wealth, ... Few would argue there is a unique or best way of combining the different sub-goals into one objective.

If we allow for a parametrized[2] objective $U_\theta$ and vary the parameters $\theta$ within reasonable ranges $\Theta$, we get a *set* of (incomparable) optimal solutions $\{s^*_\theta : \theta \in \Theta\}$, one $s^*_\theta := \arg\max_s U_\theta(s)$ for each $\theta$, and can optimize within this set for secondary criteria such as fairness $F : S \to \mathbb{R}$ without compromising the primary objective, i.e. without (societal) cost. The optimal *fair* solution is $s^*_{\theta^*}$, where $\theta^* := \arg\max_\theta F(s^*_\theta)$. The next few pages discuss and illustrate this idea a bit more, but hardly go beyond this basic idea.

I kept this note deliberately simple and focus on the basic idea. No probabilities, no machine learning, no fancy optimization algorithm – yet. Besides an example, which is purely for illustration purpose only, I also don't discuss how objectives or

[1] In machine learning classification this is known as the Accuracy↔Fairness tradeoff.

[2] This formulation also covers partially specified, partially observed, and imprecise objectives, but not stochastic uncertainty.

fairness criteria could or should be chosen. Whatever practitioners/society/ethicists deem appropriate, can be plugged in. This work is also not about bias in the data; it assumes data is unbiased or has been debiased by other means [BS16, CWV+17].

The focus is on how to improve a (given) fairness criterion without compromising solution quality with respect to some (given) primary objectives, given unbiased data.

# 2 Fairness as a Constraint

**Optimal unconstrained solution.** Consider an optimization problem with *solutions space* $S$, and some *objective* quantified by an *utility function* $U : S \to \mathbb{R}$. The[3]

$$\textit{optimal solution (by definition) is} \qquad s^* := \arg\max_{s \in S} U(s) \tag{1}$$

**Example.** Consider a simple example of student admissions based on IQ and grade. Assume there is a pool of 6 potential students $P \equiv \{\mathsf{student}_i : 1 \le i \le 6\}$ applying with information $\mathsf{student} = (\mathsf{ID}, \mathsf{name}, \mathsf{IQ}, \mathsf{grade}, \mathsf{gender}\}$ as displayed in Table 1 and Figure 1.

Assume that high $\mathsf{IQ}$ and $\mathsf{grade}$ are deemed equally important for admission to University. $\mathsf{IQ}$ is in the range 80–150 or maybe 50–200 in general, while $\mathsf{grade}$s are in the range 5–10 or in general 0–10, so they are not directly commensurable. Administrators typically rescale factors to make them commensurable, so dividing $\mathsf{IQ}$ by 10 may be adopted. We thus arrive at a performance measure

$$U(\mathsf{student}) \;:=\; \tfrac{1}{2}\mathsf{IQ}(\mathsf{student})/10 + \tfrac{1}{2}\mathsf{grade}(\mathsf{student}) \quad \in \; [0; 15]$$

Assume the University can admit 2 students and $A \subseteq P$ is the set of potentially admitted students. The goal then is to maximize objective

$$U(A) := \sum_{\mathsf{student} \in A} U(\mathsf{student})$$

which is the same as selecting the two students with highest $U$. The by-definition optimal selection is

$$A^* \;:=\; \arg\max_{A \subseteq P : |A| = 2} U(A) \;=\; \{\mathsf{student} \in P : U(\mathsf{student}) \ge u\}$$

for some suitable choice of $u$ such that the condition holds for exactly 2 students. From the $U = U_{1/2}$-column in Table 1 one can see that $\mathsf{Bob}$ and $\mathsf{Zac}$ have the highest score, i.e. $A^* = \{\mathsf{Bob}, \mathsf{Zac}\}$[4]

[3] We assume finite $S$ and bounded $U$ to avoid distracting math subtleties.

[4] To connect the notation back to (1), set $s := A$ and $S = \{A \subseteq P : |A| = 2\}$, then $s^* = A^*$. See also Figure 2.

Table 1: (**Student data&score)** There are 6 students in our running example $P$, together with their $\theta$-weighted score $U_\theta := \theta \cdot \mathsf{IQ}/10 + (1-\theta) \cdot \mathsf{grade}$ for various $\theta$.

| ID | name | IQ | grade | gender | $U = U_{1/2}$ | $U_{0.35}$ | $U_{0.2}$ |
|---|---|---|---|---|---|---|---|
| A | Amy | 100 | 10 | f | 10 | **10** | **10** |
| B | Bob | 150 | 7 | m | **11** | 9.8 | 8.6 |
| E | Eve | 150 | 5 | f | 10 | 8.5 | 7.0 |
| I | Isa | 110 | 9 | f | 10 | 9.7 | **9.4** |
| M | Max | 70 | 9 | m | 8 | 8.3 | 8.6 |
| Z | Zac | 140 | 8 | m | **11** | **10.1** | 9.2 |

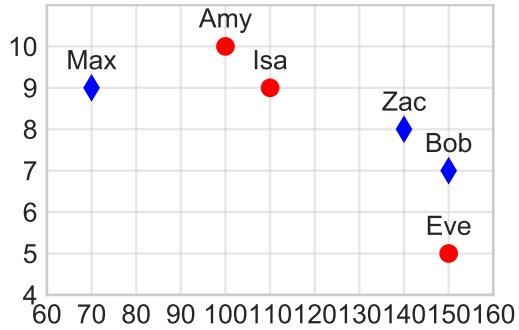


Figure 1: (**Student example)** 6 students from $P$ with IQ/grade on horizontal/vertical axis.

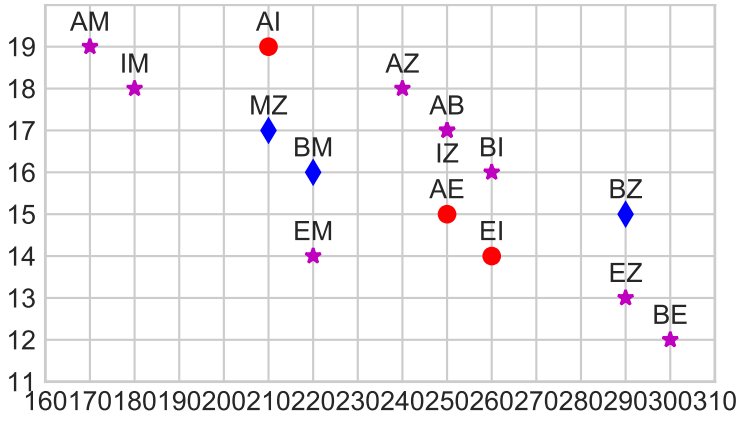


Figure 2: (**Student pairs)** All 15 pairs of students (all potential admissions $A \subseteq P : |A| = 2$) with summed IQ/grade on horizontal/vertical axis, labeled with name initials.

**Classical fairness constraint.** In the example, the average IQ and grade of men is the same as for women, namely $\langle \mathsf{IQ}|m\rangle = 120 = \langle \mathsf{IQ}|f\rangle$ and $\langle \mathsf{grade}|m\rangle = 8 = \langle \mathsf{grade}|f\rangle$. Arguably admitting two men in this situation is unfair[5]. Quotas have been argued to increase fairness, e.g. admit at least 30% women. Formally one restricts the solution space $S$ to fair solutions $S_{\mathsf{fair}} := \{s \in S : F(s) = \mathsf{fair}\}$, where $F : S \to \{\mathsf{unfair}, \mathsf{fair}\}$ is some (exact/hard) fairness constraint, leading to the

$$\textit{optimal fair solution} \qquad s^*_{\mathsf{fair}} := \arg\max_{s \in S_{\mathsf{fair}}} U(s) \tag{2}$$

**Example.** In the student admission example, one could set $F(A) := \mathsf{fair}$ *iff* $A$ contains more than 30% women, in which case $A^*_{\mathsf{fair}}$ consists of Bob or Zac and Amy or Eve or Isa. All 6 solutions score the same $U(A^*_{\mathsf{fair}}) = 21$ but less than $U(A^*) = 22$. In general, the constrained optimum $s^*_{\mathsf{fair}}$ is sub-optimal compared to the unconstrained optimal solution $s^*$. Fairness comes with a cost or regret of $U(s^*) - U(s^*_{\mathsf{fair}}) > 0$ (unless $s^*_{\mathsf{fair}} = s^*$). In the example, 10 IQ-points or 1 in grade is sacrificed.

# 3 Fairness without Regret

**Uncertain objective.** The considered admission protocol involved a number of not-so-well justified steps. For instance, IQ and grade were weighed equally, but what if overall student grade is refined to STEM grade and HASS grade, then weighing IQ:STEM:HASS as 1:1:1 may be more natural, effectively weighing IQ by $\frac{1}{3}$ and grade by $\frac{2}{3}$. Equally concerning is the adopted rescaling to make IQ and grade commensurable. The chosen scaling was plausible but by far unique. Dividing IQ by 20 to get IQ and grade into the same range 0 – 10 seems equally justified. While sophisticated analyzes or deliberations may narrow down the choices, in many social real-world problems, a considerable degree of freedom or uncertainty in the objective remains.

**Core Idea: Fairness without Regret.** The main idea of this note is to actually turn this problem into a feature, enabling fairer decision making without regret: If a unique objective is not achievable, consider the class of reasonable objective functions, or at least a sub-class thereof, say $\{U_\theta : \theta \in \Theta\}$. Each choice $\theta \in \Theta$ leads to a potentially different

$$\theta\textit{-optimal solution} \qquad s^*_\theta := \arg\max_{s \in S} U_\theta(s) \tag{3}$$

Since (by assumption) no objective among $\{U_\theta\}$ is more justified than another, $U_\theta$-optimal solutions $s^*_\theta$ are incomparable. We hence can use some secondary criterion

[5]This is neither the place for such argument, nor what term, fair↔just↔equitable, is most appropriate.

$F : S \to \mathbb{R}$ to choose among the $U_\theta$-optimal solutions $S^*_\Theta := \{s^*_\theta : \theta \in \Theta\}$ without regret, e.g. the fairest solution:

$$s^*_{\theta^*} = \arg\max_{s \in S^*_\Theta} F(s) \qquad (\text{with } \theta^* := \arg\max_{\theta \in \Theta} F(s^*_\theta)\ ) \tag{4}$$

is (the parameter corresponding to) the maximally fair solution among optimal solutions $s^*_\theta$.

**Example.** In our example, we could introduce a parameter weighing IQ versus grade:

$$U_\theta(\mathsf{student}) \ := \ \theta{\cdot}\mathsf{IQ}(\mathsf{student})/10 \ + \ (1{-}\theta){\cdot}\mathsf{grade}(\mathsf{student}) \quad \text{with} \quad \tfrac{1}{3} \le \theta \le \tfrac{2}{3}$$

We definitely want to take IQ *and* grade into account, so $\theta$ should not be close to 0 or 1. A range $\frac{1}{3} \le \theta \le \frac{2}{3}$ may be deemed plausible. A smaller range seems too dogmatic while a much larger range risks to focus too much on one attribute. The $U_\theta$-optimal admissions then are

$$A^*_\theta \ := \ \arg\max_{A \subseteq P : |A|=2} U_\theta(A) \ = \ \{\mathsf{student} \in P : U_\theta(\mathsf{student}) \ge u_\theta\}$$

As a (soft/approximate) fairness criterion we could measure the male-female number mismatch

$$\begin{aligned} -F(A) \ := \ \Big| & \#\{\mathsf{student} \in A : \mathsf{gender}(\mathsf{student}) = m\} \\ & - \#\{\mathsf{student} \in A : \mathsf{gender}(\mathsf{student}) = f\} \Big| \end{aligned}$$

Table 1 shows $U_\theta$ for $\theta = 1/2$ and $\theta = 0.35$ and the out-of-range $\theta = 0.2$. $-F(A)$ is minimized if the number of male and female admissions is the same. For $\theta = 0.35$, $A_\theta = \{\mathsf{Amy}, \mathsf{Zac}\}$ achieves this, while our original objective $U = U_{1/2}$ does not. Hence the optimal fair solution is $A^*_{\theta^*} = \{\mathsf{Amy}, \mathsf{Zac}\}$ achieved by reducing the weight of IQ a bit to e.g. $0.35 = \theta^* \in \arg\max_\theta F(A^*_\theta)$.

More generally one can show (most conveniently by inspecting Figure 2) that $3/8 < \theta < 3/4$ admits two men, $1/4 < \theta^* < 3/8$ admits one male and one female, and $\theta < 1/4$ would admit two women, but this has been deemed out-of-range, so no fairness criterion could achieve this, unless $\Theta$ is enlarged.

Note that $U_{0.35}(A^*_{0.35}) = 20.1$ while $U_{1/2}(A^*_{1/2}) = U(A^*) = 22$. This does *not* imply that the fair solution is inferior to the original unconstrained solution. $U_\theta(A)$ for different $\theta$ are incomparable (even on the same $A$). Indeed, in general, the fair utility $U_{\theta^*}(s^*_{\theta^*})$ may even be higher than the original $U(s^*)$ (assuming $\exists\theta : U = U_\theta$). For instance, this would happen if we added an (otherwise irrelevant) large positive constant to all grades, or if we apply an (otherwise irrelevant) transformation $f_\theta$ to $U_\theta$, e.g. using $\tilde{U}_\theta := (1{-}\theta){\cdot}U_\theta$.

# 4 The Optimization Problem

**A naive gradient ascent algorithm.** In order to obtain an optimal fair solution $s^*_{\theta^*}$ or $A^*_{\theta^*}$, one has to solve the coupled optimization problems (3) and (4). In general, this is a nasty non-convex and non-continuous double-optimization problem over discrete choices ($s \in S$ or $A \subseteq P$) and continuous parameters ($\theta \in \Theta$). Off-the-shelf general-purpose optimization algorithms may work sometimes. Possibly special-purpose optimizers have to be developed for large-scale real-world problems.

In case of a continuous solution space $S \subseteq \mathbb{R}^{d'}$ and continuous parameter space $\Theta \subseteq \mathbb{R}^d$ and (twice) continuously differentiable $U_{\boldsymbol{\theta}}(\boldsymbol{s})$ and $F(\boldsymbol{s})$, we could try to incrementally improve both by gradient ascent: Assume first, we solve (3) exactly, and want to improve fairness $F(\boldsymbol{s}^*_{\boldsymbol{\theta}})$ by updating $\boldsymbol{\theta}$ in direction of[6]

$$\nabla_{\boldsymbol{\theta}} F(\arg\max_{\boldsymbol{s}} U_{\boldsymbol{\theta}}(\boldsymbol{s})) \;\equiv\; \nabla_{\boldsymbol{\theta}} F(\boldsymbol{s}^*_{\boldsymbol{\theta}}) \;=:\; \mathbf{G}_{\boldsymbol{\theta}}(\boldsymbol{s}^*_{\boldsymbol{\theta}})$$

An explicit expression for $\mathbf{G}$ can be obtained by implicit differentiation [FAHG16, Lem.1&2][7]

$$\mathbf{G}_{\boldsymbol{\theta}}(\boldsymbol{s}) \;=\; -\nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{s}}^\top U_{\boldsymbol{\theta}}(\boldsymbol{s}) \cdot [\nabla_{\boldsymbol{s}} \nabla_{\boldsymbol{s}}^\top U_{\boldsymbol{\theta}}(\boldsymbol{s})]^{-1} \cdot \nabla_{\boldsymbol{s}} F(\boldsymbol{s})$$

Starting with some $(\boldsymbol{s}, \boldsymbol{\theta})$, for this fixed $\boldsymbol{\theta}$, we could now improve $\boldsymbol{s}$ by either solving maximization (3) exactly for $\boldsymbol{s} \leftarrow \boldsymbol{s}^*_{\boldsymbol{\theta}}$ or incrementally by gradient ascent

$$\boldsymbol{s} \;\leftarrow\; \Pi_S[\boldsymbol{s} + \alpha \nabla_{\boldsymbol{s}} U_\theta(\boldsymbol{s})]$$

where $\alpha$ is the learning rate and $\Pi_S$ a projection back into $S$. We then update $\boldsymbol{\theta}$ to increase fairness by

$$\boldsymbol{\theta} \;\leftarrow\; \Pi_\Theta[\boldsymbol{\theta} + \beta \mathbf{G}_{\boldsymbol{\theta}}(\boldsymbol{s})]$$

where $\beta$ is a learning rate and $\Pi_\Theta$ a projection back into $\Theta$. We then repeat and alternate between the two gradient steps. This is just one (naive) suggestion how the optimization problem could be solved. This naive algorithm may give satisfactory approximate solutions on some problems.

In our student example, $S$ is discrete, but we could try some integer relaxation. For instance, we could represent selected students $A$ as a binary vector $\boldsymbol{s} \in S := \{0,1\}^6$ with $s_i = 1$ iff $\mathsf{student}_i \in A$, then $U(A) \equiv U(\boldsymbol{s}) = \sum_{i=1}^6 s_i U(\mathsf{student}_i)$. We could then expand $S$ to the simplex $\{\boldsymbol{s} \in \mathbb{R}^6 : s_i \geq 0 \,\forall i \;\wedge\; \sum_{i=1}^6 s_i = 2\}$. Unfortunately $\nabla_{\boldsymbol{s}} \nabla_{\boldsymbol{s}}^\top U_{\boldsymbol{\theta}}(\boldsymbol{s}) \equiv 0$, since $U_{\boldsymbol{\theta}}(\boldsymbol{s})$ is linear in $\boldsymbol{s}$, so the double gradient algorithm above cannot be applied.

**Multi-objective optimization and Pareto optimality [Mie08].** For linearly parametrized objectives (and only for those), there is the following relation to

[6] All vectors are taken to be column vectors, including the gradient $\nabla$, unless transposed by $\top$.

[7] Differentiate $\nabla_{\boldsymbol{s}} U_{\boldsymbol{\theta}}(\boldsymbol{s})_{|\boldsymbol{s}=\boldsymbol{s}^*_{\boldsymbol{\theta}}} \equiv 0$ w.r.t. $\boldsymbol{\theta}$ and solve for $\nabla_{\boldsymbol{\theta}} \boldsymbol{s}^*_{\boldsymbol{\theta}}$, and plug this into $\nabla_{\boldsymbol{\theta}} F(\boldsymbol{s}^*_{\boldsymbol{\theta}}) = \nabla_{\boldsymbol{\theta}} \boldsymbol{s}^{*\top}_{\boldsymbol{\theta}} \cdot \nabla_{\boldsymbol{s}} F(\boldsymbol{s})_{|\boldsymbol{s}=\boldsymbol{s}^*_{\boldsymbol{\theta}}}$.

Pareto optimality: In multi-objective optimization one considers $m > 1$ objectives $U_1, ..., U_m : S \to \mathbb{R}$ over solution space $S$. A solution $s \in S$ is called Pareto optimal *iff* it is not dominated by any other $s' \in S$ in the sense of $\neg\exists s' \in S : [\forall j : U_j(s') \geq U_j(s) \wedge \exists j : U_j(s') > U_j(s)]$. The Pareto front $\mathsf{PF} \subseteq S$ is the set of all Pareto optimal $s \in S$. All other $s \notin \mathsf{PF}$ are clearly sub-optimal. Consider now the weighted sum of utilities $U_{\boldsymbol{\theta}}(s) := \sum_{j=1}^m \theta_j U_j(s)$ with $\theta_j > 0 \, \forall j$ (strict inequality is important here). It is easy to see that for any $\boldsymbol{\theta} > 0$, $s^*_{\boldsymbol{\theta}} := \arg\max_{s \in S} U_{\boldsymbol{\theta}}(s)$ is Pareto optimal. The converse, that any $s \in \mathsf{PF}$ is $U_{\boldsymbol{\theta}}$-optimal for some $\boldsymbol{\theta} > 0$ however is in general not true. It holds true if $\{(U_1(s), ..., U_m(s)) : s \in S\}$ is a convex set[8], but for a finite data sets $S$ (e.g. $S = \{A \subseteq P : |A| = k = 2\}$) in the student example) this is *never* convex[9]. Lacking a better term, let us call $\mathsf{CPF} := \{s^*_{\boldsymbol{\theta}} : \boldsymbol{\theta} > 0\}$ the "convex" Pareto front[10] An $s \in S$ is called weakly Pareto optimal ($\mathsf{WPF}$) *iff* $\neg\exists s' \in S : [\forall j : U_j(s') > U_j(s)]$. For fair decision making, we are only interested in reasonable mixtures $\boldsymbol{\theta} \in \Theta \subsetneq (0; \infty)^m$, and the set $\Theta$ may not even be an axis-aligned hypercube. Therefore in general

$$\{s^*_{\boldsymbol{\theta}^*}\} \subsetneq S^*_\Theta \subsetneq \mathsf{CPF} \subsetneq \mathsf{PF} \subsetneq \mathsf{WPF} \subsetneq S$$

For instance, for our example one can show that all inclusions are strict (see Figure 2):

$$\begin{aligned}
\text{optimal Fair solution:} \quad s^*_{\theta^*} &= \{\mathsf{Amy}, \mathsf{Zac}\}, \\
(\Theta)\text{-optimal solution set:} \quad S^*_\Theta &= \{s^*_{\theta^*}\} \cup \{\{\mathsf{Bob}, \mathsf{Zac}\}\}, \\
\text{convex Pareto front:} \quad \mathsf{CPF} &= S^*_\Theta \cup \{\{\mathsf{Amy}, \mathsf{Isa}\}, \{\mathsf{Bob}, \mathsf{Eve}\}\}, \\
\text{Pareto front:} \quad \mathsf{PF} &= \mathsf{CPF} \cup \{\{\mathsf{Amy}, \mathsf{Bob}\}, \{\mathsf{Bob}, \mathsf{Isa}\}, \{\mathsf{Isa}, \mathsf{Zac}\}\}, \\
\text{weak Pareto front:} \quad \mathsf{WPF} &= \mathsf{PF} \cup \{\{\mathsf{Eve}, \mathsf{Zac}\}\}, \\
\text{solution space:} \quad S &= \mathsf{WPF} \cup \{\{\mathsf{Amy}, \mathsf{Eve}\}, \{\mathsf{Bob}, \mathsf{Eve}\}, \{\mathsf{Max}, *\}\}
\end{aligned}$$

Nevertheless, for linear mixtures one may use ideas from multi-objective optimization and Pareto frontiers to narrow down the solution space to aid finding $S^*_\Theta$ and ultimately $s^*_{\theta^*}$. Note though that this approach is limited to linear mixtures of objectives, but not generally parametrized objectives $U_\theta$, and even $\Theta$ does not need to be a (subset of a) vector space, so the connection to Pareto optimality is somewhat weak.

[8] Or more generally if all points in this set lie on the boundary of its convex hull, which may or may not be true for finite $S$.

[9] For instance if we admit $k = 1$ student in our example, then $\mathsf{PF} = \{\mathsf{Amy}, \mathsf{Bob}, \mathsf{Isa}, \mathsf{Zac}\} \subsetneq S$ are Pareto optimal, but $\mathsf{Isa}$ is *not* $U_{\boldsymbol{\theta}}$-optimal for any $\boldsymbol{\theta} > 0$.

[10] $\mathsf{CPF}$ itself can of course not be convex, since $S$ is not a vector space, but even $\mathsf{UPF} := \{(U_1(s), ..., U_m(s)) : s \in \mathsf{CPF}\}$ is usually not convex, but all points in $\mathsf{UPF}$ lie on the boundary of the convex hull of $\mathsf{UPF}$.

# 5 Discussion

**Perfect fairness.** I have demonstrated how to incorporate fairness as a secondary optimization criterion without compromising solution quality by exploiting that many real-life objectives cannot unambiguously be defined. It is important to note that if there is a binary notion of perfect fairness, it may not be achievable with this procedure (unlike in the simple example).

**Controversial fairness.** On the other hand, fairness is a notoriously contentious notion [VR18]. In our example, should irrelevant birth factors be even taken into account, i.e. included in the data? If so, then which ones and why? Gender? Skin color? Eye color? Body height? Should any imbalance in the pool of applicants be taken into account (not a problem in our example)? Given there are many contradictory notions of fairness [Zho18, Sec.4.7], *improving* (presumed) fairness is probably wiser than aiming for perfect fairness. Our approach does the former without harming solution quality; even optimizing for controversial fairness notions (e.g. demographic parity [Har16, ZVRG17]) becomes unproblematic.

I also assumed that there is no bias in the data, or at least this work did not address this issue. While removing explicit attributes in the data regarded as irrelevant is easy, how to deal with implicit bias in the data is subject to ongoing research [BS16, CWV$^+$17]. One may argue that once data is debiased, there is no need for secondary fairness criteria, but the former seems difficult to achieve or even know, and further diversity arguments will probably always remain.

**Non-unique objectives.** Coming up with an appropriate parametrized objective can itself be a challenge, but arguably this is a better/easier problem than to specify a unique objective. Being forced to agree on a relative weighing of factors can be arduous and the result may easily be determined by authority or whoever shouts loudest rather than rationally by reason and deliberation. A range of objectives seems easier to converge to. In the simplest case one could pool the proposed utility functions of different experts, or better, start with a large parametrized class $\{U_\theta\}$, e.g. *any* (non)linear combination of attributes, then choose $\Theta$ to be the convex hull of expert choices $\theta_1, \theta_2, \theta_3, ....$ One may lean towards a smaller range $\Theta$ if the fairness criterion is controversial, or a larger range $\Theta$ if fairness is deemed crucial.

**Uncertainty in data.** Consider a selection problem of $k$ items from a large(r) population $P = \{x_1, ..., x_n\}$ as in the example, where $x_i \in X$ was a student record, $n = 6$ and $k = 2$. Assume some attributes such as IQ are missing or not precisely known, which can be modeled as interval-valued or more generally set-valued attributes. In this case, a student record becomes a set $X_i \subseteq X$, the data set becomes $\mathcal{P} = X_1 \times ... \times X_n$, and $P \in \mathcal{P}$ is one (arbitrary) completion or choice or imputation of attributes.[11] For each choice we can find the optimal solution and then the

[11] While in this notation $P$ strictly speaking is an $n$-tupel, we will interpret $P$ also as a set of size $n$, so that $A \subseteq P$ is well-defined.

(supposedly) fairest choice:

$$A_P^* := \arg\max_{|A|=k} U_P(A) \quad \text{and} \quad P^* := \arg\max_{P\in\mathcal{P}} F(A_P^*) \tag{5}$$

Despite the similarity in mathematical structure to the uncertain objective case ($\Theta\widehat{=}\mathcal{P}$ and $\theta\widehat{=}P$), there is a crucial difference which renders $A_{P^*}^*$ actually very biased or *un*fair. Assume that naively using mean values for uncertain attributes leads to a high proportion of male admissions. Using (5) instead may indeed lead to more women being admitted, but inspecting $P^*$ would reveal that this has been achieved by imputing IQ and grades at the low interval boundary for males and at the high interval end for women, which is difficult to justify as fair. To summarize: Uncertainty in data is fundamentally different from uncertainty in the objective, and procedure (5) does *not* lead to fair decisions.

# 6 Outlook

The basic proposed idea (possibly) can and needs to be extended in various ways: For instance, I have not discussed stochastic uncertainty: The data could be stochastic, and/or the evaluation of the objective may be stochastic.

Many problems involve a machine learning component to solve, so there could be bias and uncertainty in the learned model.

Possibly the most important question is how much can fairness be increased by expanding a single objective to a parametrized class, or more generally, how does $F(s_\theta^*)$ depend on $\Theta$. This will heavily depend on the problem domain, primary objective, the fairness criterion, the data, and how large a $\Theta$ can be well-justified before it becomes an opportunity for rigging rather than fairness. To make theoretical progress on this question, some structural assumptions on $U_\theta$, $\Theta$, and $F$ have to be made.

Finally, in order to obtain optimal fair solutions one has to solve a challenging non-convex and non-continuous double-optimization problem over discrete choices and continuous parameters.

**Acknowledgements.** I thank Iason Gabriel for valuable feedback on earlier drafts.

[ABD+18] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning (ICML'18) and ACM Conference on Fairness, Accountability and Transparency (ACM-FAT'18)*, pages 60–69, 2018.

[BS16] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

[CWV+17] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems (NIPS 2017)*, pages 3992–4001, 2017.

[FAHG16] B. Fernando, P. Anderson, M. Hutter, and S. Gould. Discriminative hierarchical rank pooling for activity recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*, pages 1924–1932, Las Vegas, NV, USA, 2016. IEEE.

[Har16] Moritz Hardt. Approaching fairness in machine learning. *Moody Rd*, September 2016. http://blog.mrtz.org/2016/09/06/approaching-fairness.html.

[Mie08] Kaisa Miettinen. Introduction to multiobjective optimization: Noninteractive approaches. In *Multiobjective optimization*, pages 1–26. Springer, 2008.

[MW18] Aditya Krishna Menon and Robert C Williamson. The cost of fairness in binary classification. In *ACM Conference on Fairness, Accountability and Transparency (ACM FAT)*, pages 107–118, 2018.

[VR18] Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pages 1–7. IEEE, 2018.

[Zho18] Z. Zhong. A tutorial on fairness in machine learning. *Towards Data Science*, October 2018. https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb.

[ZVRG17] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics (AISTATS 2017)*, pages 962–970, 2017.