Fairness without Regret[*]

Marcus Hutter

DeepMind & ANU

Latest version & more @
http://www.hutter1.net/official/bib.htm#fair

3 March 2020

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. We 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, solution quality

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

1 Introduction

Terminology. 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.

Fairness by constraint. 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. a *quota-based* system that admits 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 primary 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]

Uncertainty in objective. In practice, the ultimate goal is often a (possibly non-linear) aggregate of sub-goals, 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.

Fairness without regret. If we allow for a parametrized[2] objective U_θ and vary the parameters θ within reasonable ranges Θ, we get a *set* of (incomparable) optimal solutions $\{s^*_\theta : \theta \in \Theta\}$, one solution $s^*_\theta := \arg\max_s U_\theta(s)$ for each θ, 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.

[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.

What this work is NOT about. We 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, we also don't discuss how objectives or fairness criteria or attributes 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 (sufficiently) unbiased or has been (sufficiently) debiased by other means, a non-trivial [BS99, BS16] but not impossible endeavor [CWV+17]. We also do not provide a ready algorithm nor an integrated practical system nor treat specific applications.

What this work IS about. The focus is on introducing and discussing *a novel way to improve a (given) fairness criterion without compromising solution quality with respect to some (given) primary objectives*, given unbiased data.

Background and (un)related work. The literature on algorithmic fairness is vast, but we are not aware of any idea to solving the fairness problem similar to ours. [Zho18] is a lean tutorial introducing and comparing the most prominent notions of bias and fairness in machine learning, and [MMS+19] contains a more comprehensive survey and list of references. An empirical study comparing fairness-enhancing interventions in machine learning can be found in [FSV+19]. The literature is scattered due to the interdisciplinarity of the subject and its broad relevance for and diverse applications in society, and possibly due to the plethora of contradictory [Zho18, Sec.4.7] and contentious [VR18] and controversial [Har16, ZVRG17] notions of fairness. How to deal with implicit bias in the data is subject to ongoing research [BS16, CWV+17] and not addressed by our work. For instance, [KC09] "learn[] unbiased models on biased training data [by] massaging the dataset by making the least intrusive modifications which lead to an unbiased dataset." Our work is closer to work that tries to improve fairness by constraints [ZVRG17] in the sense that we reject this notation and provide and argue for an alternative solution.

Content. Section 2 introduces the setup and the classical approach to fairness as a constraint. In Section 3 we present the main idea of uncertain objectives, which allows simultaneously for optimal *and* fair(er) solutions. Section 4 discusses the resulting mathematical coupled bilevel optimization problem and, for linearly parametrized objectives, its similarities and differences to multi-objective optimization. Section 5 contains (more/informal) discussion of fairness, biased data, non-unique objectives, and uncertainty in data. Section 6 is a brief outlook.

2 Fairness as a Constraint

In this section we describe the classical approach to fairness by confining the solution space to solutions deemed fair. We illustrate its effect in reducing the solution quality w.r.t. the primary objective on a small student admission example. This sets the stage for our novel proposal in the subsequent section which circumvents this problem.

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]

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

Example. Consider a simple example of student admissions based on IQ and grade. Assume there is a pool of 6 potential students $P \equiv \{\mathsf{student}_i : 1 \leq i \leq 6\}$ 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 IQ and grade are deemed equally important for admission to University. IQ is in the range 80–150 or maybe 50–200 in general, while grades 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 IQ by 10 may be adopted. We thus arrive at a performance measure

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

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

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

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 Bob and Zac have the highest score, i.e. $A^* = \{\mathsf{Bob}, \mathsf{Zac}\}$.[4]

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

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

[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.

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

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

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

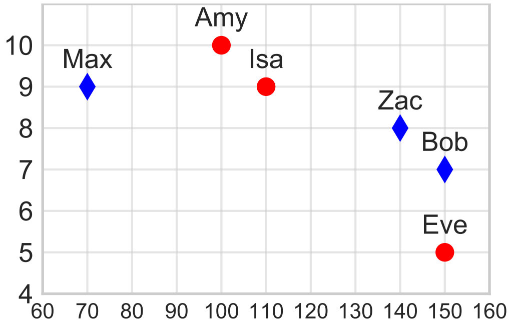

Figure 1: (**Student example**) 6 students from P with $\mathsf{IQ}/\mathsf{grade}$ on horizontal/vertical axis.

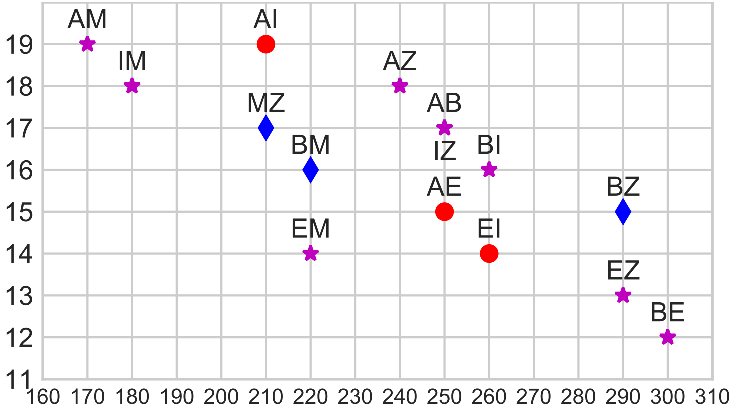

Figure 2: (**Student pairs**) All 15 pairs of students (all potential admissions $A \subseteq P : |A| = 2$) with summed $\mathsf{IQ}/\mathsf{grade}$ on horizontal/vertical axis, labeled with name initials.

Example. In the student admission example, one could set $F(A) := \mathsf{fair}$ *iff* A contains more than 30% women, in which case A^*_{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^*_{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

We now introduce and discuss our novel idea, namely to improve a (given) fairness criterion *without* compromising solution quality with respect to some (given) primary objective. The main idea is based on the concept of uncertain objectives, which allows simultaneously for optimal *and* fair(er) solutions. We illustrate the idea on the running student example of the previous section.

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 analyses 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 (FAWOR). 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 (primary) 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

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

Since (by assumption) no objective among $\{U_\theta\}$ is more justified than another, U_θ-optimal solutions s^*_θ are incomparable. We hence can use some (secondary) fairness criterion $F : S \to \mathbb{R}$ to choose among the U_θ-optimal solutions $\boxed{S^*_\Theta := \{s^*_\theta : \theta \in \Theta\}}$ without regret, e.g. the fairest solution:

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

is (the parameter corresponding to) the maximally fair solution among optimal solutions s^*_θ.

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

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

We definitely want to take IQ *and* grade into account, so θ should not be close to 0 or 1. A range $\frac{1}{3} \leq \theta \leq \frac{2}{3}$ 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_θ-optimal admissions then are

$$A^*_\theta \;:=\; \underset{A\subseteq P:|A|=2}{\arg\max}\, U_\theta(A) \;=\; \{\mathsf{student} \in P : U_\theta(\mathsf{student}) \geq u_\theta\}$$

for a suitable threshold $u_\theta \in \mathbb{R}$ so that $|A^*_\theta| = 2$. As a (soft/approximate) fairness criterion we could measure the male-female number mismatch

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

Table 1 shows U_θ 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 Θ 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 θ 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_θ to U_θ, e.g. using $\tilde{U}_\theta := (1-\theta)\cdot U_\theta$.

Generality. We want to emphasize the generality of this idea/approach and formulation (3) and (4). It does not rely on any specific functional shape of utility U_θ nor fairness criterion F. The parameter space Θ as well as underlying (student) attributes can be continuous or ordinal (IQ, grades, age) or boolean (award-nominee) or categorical (school, ethnicity). Also, the application domain and solution space S are completely general. In the above toy example we considered student admission based on IQ and grades and gender, and S consisted of subsets of students of size two.

As a side remark, while studies have shown that IQ positively correlates with (alternative measures of) intelligence, with academic success, with job performance, and with social status [HM96, NAB$^+$12], others have disputed or at least cautioned about their validity [RN15]. Similarly for (school) grades. We want to stress that these controversies are irrelevant to our proposal in the following sense: This paper neither endorses nor rejects the suitability of *any* attributes or categories (be it IQ or grades or else) for assessing students, not even whether gender or ethnic balance is desirable or not. If an attribute or category is deemed (ir)relevant or controversial, this has to be sorted out, whether using FAWOR or quotas or any other methodology. Indeed, if anything, FAWOR ameliorates the problem, by allowing to *flexibly* incorporate the attributes, without having to completely resolve the issue. Anyone concerned about the automation of ethical decisions should be pleased that 4 of the 5 steps in the protocol in the box have to be done by humans, with Step 4 reserved for ethicists. Only the last step is purely algorithmic. Further discussion is deferred to Section 5.

FAWOR Protocol

1. *Choose S:* A committee (say) searches for attributes that potentially affect the *primary objective*, e.g. are potentially indicative of the suitability and performance of students, such as prior grades and other achievements.

2. *Choose U_θ:* Based on these attributes s, the committee develops "utility" functions $U_\theta(s)$ that *potentially* capture the a-priori intuitive objective, e.g. of student's success. In the simplest case this could just be a θ-weighted average of attributes.

3. *Choose Θ:* The next step is to select a reasonable range for θ, e.g. reasonable ranges for each weight. The range of utility functions $\{U_\theta(s) : \theta \in \Theta\}$ should be chosen as broad as reasonable. An important feature that distinguishes FAWOR from constraint-based quota systems is that the committee members do not need to agree. They could just take the union (or convex hull) of their individual choices, as long as everyone is acting in good faith.

4. *Choose F:* The committee determines the *fairness criterion $F(s)$* of interest, which can depend on additional attributes such as gender or race.

5. *Compute $s^*_{\theta^*}$:* In theory, they then calculate the *set* of all optimal solutions S^*_Θ via (3) and from that the *optimal fair solution* $s^*_{\theta^*}$ via (4). Practical approaches to finding this solution are suggested in Section 4.

Step 1 is the same as for any methodology optimizing some (social) objective. It neither is nor needs to be specific to FAWOR, and for better comparison with other methods indeed should not be tailored to FAWOR. Step 2 is an exploratory phase which should also be very similar between different methods.

Unlike in the pedagogical example, where we first considered a single fixed objective $U = U_{1/2}$, rejected the solution, and then reconsidered how to achieve fair(er) solutions, it is important that Step 3 is actually done before Step 5, i.e. before looking at the result and ideally also before looking at the data. If the committee would start modifying/enlarging Θ, maybe because they feel the outcome is not "fair enough", they face the risk of post-hoc rationalization and inadvertently or deliberately rigging the system. But note that the classical fairness-as-a-constraint approach described in Section 2 is (arguably) rigged from the outset, i.e. carefully iterating the Protocol above, is still less problematic than quota-based approaches.

4 The Optimization Problem

Unfortunately our novel fairness criterion (3) and (4) involve a nasty double optimization problem. Structurally it is a (special case of a) so-called bilevel optimization problem. We first suggest a naive gradient ascent algorithm, which can work for some objectives, but unfortunately not for the most likely linear ones appearing in practice. Fortunately for linearly parametrized objectives, our problem reduces to a multi-objective optimization problem, albeit with the crucial difference that the fairness criterion restricts the Pareto front. Again, we illustrate the different solutions on our running student admission example. Finally, a machine learning approach to unknown objective U_θ would be to somehow learn θ from data itself, which can reduce the uncertainty Θ (somewhat) but rarely if ever completely.

Bilevel optimization. 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 bilevel optimization problem over discrete choices ($s \in S$ or $A \subseteq P$) and continuous parameters ($\theta \in \Theta$). Principled [Bar98] or heuristic [Tal13] off-the-shelf general-purpose optimization algorithms may work for some choices of $U_\theta()$, $F()$, Θ, and data. Possibly special-purpose optimizers have to be developed for large-scale real-world problems. Our problem is actually not a fully general bilevel optimization problem, but exhibits special properties which may be exploitable.

A naive gradient ascent algorithm. 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,

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

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 α is the learning rate and Π_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 β is a learning rate and Π_Θ a projection back into Θ. 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 naively. If we add the convex constraints on s, this becomes (a special case of) a BiLevel optimization problem [GFC$^+$16].

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

[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 $\nabla_{\boldsymbol{\theta}}F(\boldsymbol{s}^*_{\boldsymbol{\theta}})=\nabla_{\boldsymbol{\theta}}\boldsymbol{s}^{*\top}_{\boldsymbol{\theta}}\cdot\nabla_{\boldsymbol{s}}F(\boldsymbol{s})_{|\boldsymbol{s}=\boldsymbol{s}^*_{\boldsymbol{\theta}}}$.

[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 Isa is *not* $U_{\boldsymbol{\theta}}$-optimal for any $\boldsymbol{\theta}>0$.

[10] 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 UPF lie on the boundary of the convex hull of UPF.

is called weakly Pareto optimal (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 Θ 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):

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

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^*_Θ and ultimately $s^*_{\theta^*}$. Note though that this approach is limited to linear mixtures of objectives, but not generally parametrized objectives U_θ, and even Θ does not need to be a (subset of a) vector space, so the connection to Pareto optimality is somewhat weak.

Machine Learning [MB12]. The machine learning approach to unknown objective U_θ would be to somehow learn θ. This requires relevant labeled examples, e.g. a data base of now-graduated students containing their original application profile side-by-side with their grades at graduation. Even better would be to follow-up and record their future success in life, which would presumably be closer to the true objective of student admission selection.

Absent of such data we are left with the problem of hand-designed objectives, with a range of reasonable θ as in our example. Even if we have training data, this still leaves at least two sources of indeterminate θ.

(i) Finite data can never determine sharp values for θ. Some parameters may be learnable exactly-for-all-practical-purposes, others may have large error bars. In the latter case, a confidence interval/set could be chosen for θ.

(ii) Even infinite data can only determine θ uniquely and correctly if we uniquely and exactly know the ultimate success criterion (the labels) we are aiming at. Since we often don't, this just lifts the problem one level up. For instance, is it really grades we care to optimize for when admitting students? Or is future lifetime income more relevant, which would still be measurable in a longitudinal study? Or is maximizing for positive contribution to society a better criterion? But is that measured by a student's direct+indirect increase of GDP or GNH [UAZW12] or otherwise [SFD18]? Such uncertainties lead to a set of possible training labels $\mathsf{admit}^\theta_i = 1 \in \{0,1\}$ if a student turned out to be successful according to criterion C_θ. Assume we have a parametrized probabilistic classifier $\hat{C}^w_\theta : P \to [0;1]$. The predictive log-loss then is $L_\theta(w) := \log|\mathsf{admit}^\theta_i - \hat{C}^w_\theta(\mathsf{student}_i)|$. The minimizer $w^* := \arg\min_w L_\theta(w)$ (or some regularized version or confidence interval estimation) then leads to objective $U_\theta := \hat{C}^{w^*}_\theta$. Note that w^* depends on θ, so we now have a coupled *triple*-optimization problem.

So as long as the ultimate objective is debatable or data not perfect, we are left with a set $\{U_\theta : \theta \in \Theta\}$ of possible objectives, and the approach in this paper remains relevant.

5 Discussion

This section contains a more qualitative discussion of issues related to notions of fairness, bias in data, non-unique objectives, and uncertainty in data.

Perfect fairness. We 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? Body height? Eye color? Should any imbalance in the pool of applicants be taken into account (not a problem in our example)? Is representational/demographic parity fair? If so, w.r.t. which attributes? Gender? Political spectrum? IQ? Detention rates? 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. Our running student example should neither be construed as endorsing any of the chosen attributes or categories or numbers, nor choice of fairness criterion.

Biased data. We 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 [KC09, 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 Θ to be the convex hull of expert choices $\theta_1, \theta_2, \theta_3, ...$. One may lean towards a smaller range Θ if the fairness criterion is controversial, or a larger range Θ 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 (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.

Ethical concerns. Many ethicists disapprove of (semi)automating ethical decisions. They believe this approach is misguided or dangerous or inhuman, but this attitude is unhelpful. Society unlikely will forgo decision algorithms with societal impact and hence ethical consequences. The realistic choice is between algorithms that make more or less ethical decisions. In any case, 4 out of 5 steps in our FAVOR protocol in Section 3 are under human control.

We are also fully aware of the (five) abstraction errors criticized in [SBF$^+$19], but progress on fundamental questions requires over-simplifying abstractions initially. Scientists who present neat and clean ideas (including us) are (usually) neither naive nor ignorant about the intricacies of real societal challenges.

A short response to such accusations is: Quotas are a fairly crude instrument to improve equality, but are nevertheless often used in practice if/since other instruments are not realistically available or effective. The (only) claim this paper makes is that FAWOR is a *superior* alternative to such constraint-based approaches, and

[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.

its usage is nearly as easy as quota-based systems (apart from algorithmic complexities).

A longer response is: Rome was not built in a single day by a single person, so neither does every paper need to produce a complete ready-to-employ ethical system approved by society. Many philosophical, mathematical, ethical, computational, machine learning, political, and engineering ideas need to be combined to produces such systems. Science works by a division of labor, some develop fundamental novel ideas and concepts in the abstract, others (jointly) try to integrate and test them and adapt them to make them fit for practice, still others decide whether the resulting system will in the end be deployed or not. But without over-simplifying abstractions of real-world problems in a first instance, progress on difficult problems would be impossible.

For instance, the difference between fair/just/equitable didn't matter for this work. Of course this does not mean that we believe there is no difference between these concepts or that they don't matter. They matter when concretely instantiating U_Θ and F, and possibly for refining the core idea itself, and when integrated it into an overall system.

Also, the mentioning of certain attributes (gender, race, IQ, ...) was for purely illustrative purposes, and in any case "the use of racial categories in algorithmic fairness research (i.e. the research community which has emerged around venues like FAT* and AIES) has largely gone unquestioned." [HDSS20], and as explained above, it would be off-topic for this work to engage in these questions.

6 Outlook

The basic proposed idea (possibly) can and needs to be extended in various ways: For instance, we 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 theoretical 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 Θ. This will heavily depend on the problem domain, primary objective, the fairness criterion, the data, and how large a Θ 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_θ, Θ, and F have to be made. In practice one should probably refrain from iterating the FAWOR protocol.

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

Acknowledgements. We thank Iason Gabriel for valuable feedback on earlier

drafts.

References

[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.

[Bar98] Jonathan F. Bard. *Practical Bilevel Optimization: Algorithms and Applications*. Number v. 30 in Nonconvex Optimization and Its Applications. Kluwer Academic Publishers, Dordrecht ; Boston, 1998.

[BS99] Geoffrey C. Bowker and Susan Leigh Star. *Sorting Things out: Classification and Its Consequences*. Inside Technology. MIT Press, Cambridge, Mass, 1999.

[BS16] Solon Barocas and Andrew D. Selbst. Big Data's Disparate Impact. *SSRN Electronic Journal*, 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] Basura Fernando, Peter Anderson, Marcus Hutter, and Stephen 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.

[FSV$^+$19] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*, pages 329–338, Atlanta, GA, USA, 2019. ACM Press.

[GFC$^+$16] Stephen Gould, Basura Fernando, Anoop Cherian, Peter Anderson, Rodrigo Santa Cruz, and Edison Guo. On Differentiating Parameterized Argmin and Argmax Problems with Application to Bi-level Optimization. *arXiv:1607.05447 [cs, math]*, July 2016.

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

[HDSS20] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 501–512, Barcelona Spain, January 2020. ACM.

[HM96] Richard J. Herrnstein and Charles A. Murray. *The Bell Curve: Intelligence and Class Structure in American Life.* Simon & Schuster, New York, 1st free press pbk. ed edition, 1996.

[KC09] Faisal Kamiran and Toon Calders. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*, pages 1–6, Karachi, Pakistan, February 2009. IEEE.

[MB12] Kevin P. Murphy and Francis Bach. *Machine Learning: A Probabilistic Perspective.* MIT Press, Cambridge, MA, September 2012.

[Mie08] Kaisa Miettinen. Introduction to Multiobjective Optimization: Noninteractive Approaches. In Jürgen Branke, Kalyanmoy Deb, Kaisa Miettinen, and Roman Słowiński, editors, *Multiobjective Optimization*, volume 5252, pages 1–26. Springer, Berlin, Heidelberg, 2008.

[MMS$^+$19] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A Survey on Bias and Fairness in Machine Learning. *arXiv:1908.09635 [cs]*, September 2019.

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

[NAB$^+$12] Richard E. Nisbett, Joshua Aronson, Clancy Blair, William Dickens, James Flynn, Diane F. Halpern, and Eric Turkheimer. Intelligence: New findings and theoretical developments. *American Psychologist*, 67(2):130–159, February 2012.

[RN15] Ken Richardson and Sarah H. Norgate. Does IQ Really Predict Job Performance? *Applied Developmental Science*, 19(3):153–169, July 2015.

[SBF$^+$19] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 59–68, Atlanta GA USA, January 2019. ACM.

[SFD18] Joseph E. Stiglitz, Jean-Paul Fitoussi, and Martine Durand, editors. *For Good Measure: Advancing Research on Well-Being Metrics beyond GDP.* OECD Publishing, 2018.

[Tal13] El-Ghazali Talbi. *Metaheuristics for Bi-Level Optimization.* Number 482 in Studies in Computational Intelligence. Springer, New York, 1st ed edition, 2013.

[UAZW12] Karma Ura, Sabina Alkire, Tshoki Zangmo, and Karma Wangdi. *An Extensive Analysis of GNH Index.* OECD, 2012.

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

[Zho18] Ziyuan Zhong. A Tutorial on Fairness in Machine Learning, August 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.