Balancing Competing Objectives with Noisy Data: Score-Based Classifiers for Welfare-Aware Machine Learning

Esther Rolf [1] **Max Simchowitz** [1] **Sarah Dean** [1] **Lydia T. Liu** [1] **Daniel Björkegren** [2] **Moritz Hardt** [1]
Joshua Blumenstock [3]

Abstract

While real-world decisions involve many competing objectives, algorithmic decisions are often evaluated with a single objective function. In this paper, we study algorithmic policies which explicitly trade off between a private objective (such as profit) and a public objective (such as social welfare). We analyze a natural class of policies which trace an empirical Pareto frontier based on learned scores, and focus on how such decisions can be made in noisy or data-limited regimes. Our theoretical results characterize the optimal strategies in this class, bound the Pareto errors due to inaccuracies in the scores, and show an equivalence between optimal strategies and a rich class of fairness-constrained profit-maximizing policies. We then present empirical results in two different contexts — online content recommendation and sustainable abalone fisheries — to underscore the applicability of our approach to a wide range of practical decisions. Taken together, these results shed light on inherent trade-offs in using machine learning for decisions that impact social welfare.

1. Introduction

From medical diagnosis and criminal justice to financial loans and humanitarian aid, consequential decisions increasingly rely on data-driven algorithms. Machine learning algorithms used in these contexts are mostly trained to optimize a single metric of performance. As a result, the decisions made by such algorithms can have unintended adverse side effects: profit-maximizing loans can have detrimental effects on borrowers (Skiba & Tobacman, 2009) and fake news can undermine democratic institutions (Persily, 2017).

The field of fair machine learning proposes algorithmic approaches that mitigate the adverse effects of single objective maximization. Thus far it has predominantly done so by defining various fairness criteria that an algorithm ought to satisfy (see e.g., Barocas et al., 2019, and references therein). However, a growing literature highlights the inability of any one fairness definition to solve more general concerns of social equity (Corbett-Davies & Goel, 2018). The impossibility of satisfying all desirable criteria (Kleinberg et al., 2017) and the unintended consequences of enforcing parity constraints based on sensitive attributes (Kearns et al., 2018) indicate that existing fairness solutions are not a panacea for these adverse effects. Recent work (Liu et al., 2018; Hu & Chen, 2020) contend that while social welfare is of primary concern in many applications, common fairness constraints may be at odds with the relevant notion of welfare.

In this paper, we consider *welfare-aware machine learning* as an inherently multi-objective problem that requires explicitly balancing multiple objectives and outcomes. A central challenge is that certain objectives, like welfare, may be harder to measure than others. Building on the traditional notion of Pareto optimality, which provides a characterization of optimal policies under complete information, we develop methods to balance multiple objectives when those objectives are measured or predicted with error.

We study a natural class of selection policies that balance multiple objectives (e.g., private profit and public welfare) when each individual has predicted *scores* for each objective (e.g., their predicted contribution to total welfare and profit). We show that this class of score-based policies has a natural connection to statistical parity constrained classifiers and their ϵ-fair analogs. In the likely case where scores are imperfect predictors, we bound the sub-optimality of the multi-objective utility as a function of the estimator errors. Simulation experiments highlight characteristics of problem settings (e.g. correlation of the true scores) that affect the extent to which we can jointly maximize multiple objectives.

[1]Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, California, USA [2]Department of Economics, Brown University, Providence, USA [3]School of Information, University of California at Berkeley, Berkeley, USA. Correspondence to: Esther Rolf <esther_rolf@berkeley.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

We apply the multi-objective framework to data from two diverse decision-making settings. We first consider an ecological setting of sustainable fishing, where we study score degradation to mimic certain dimensions being costly or impossible to measure. Our second empirical study uses existing data on the popularity and 'social health' of roughly 40,000 videos promoted by YouTube's recommendation algorithm, and shows that multi-objective optimization could produce substantial increases in average video quality for almost negligible reductions in user engagement.

In summary, we provide a characterization, theoretical analysis, and empirical study of a score-based multi-objective optimization framework for learning welfare-aware policies. We hope that our framework may help decouple the complex problem of defining and measuring welfare, which has been studied at length in the social sciences, e.g. (Deaton, 2016), from a machine toolkit geared towards optimizing it.

2. Related Work

2.1. Fair and Welfare-Aware Machine Learning

The growing subfield of *fairness in machine learning* has investigated the implementation and implications of machine learning algorithms that satisfy definitions of fairness (Dwork et al., 2012; Barocas & Selbst, 2016; Barocas et al., 2019). Machine learning systems in general cannot satisfy multiple definitions of group fairness (Chouldechova, 2017; Kleinberg et al., 2017), and there are inherent limitations to using observational criteria (Kilbertus et al., 2017). Alternative notions of fairness more directly encode specific trade-offs between separate objectives, such as per-group accuracies (Kim et al., 2019) and overall accuracy versus a continuous fairness score (Zliobaite, 2015). These fairness strategies represent trade-offs with domain specific implications, for example in tax policy (Fleurbaey & Maniquet, 2018) or targeted poverty prediction (Noriega et al., 2018).

An emerging line of work is concerned with the long-term impact of algorithmic decisions on societal welfare and fairness (Ensign et al., 2018; Hu & Chen, 2018; Mouzannar et al., 2019; Liu et al., 2020). Liu et al. (2018) investigated the potentially harmful delayed impact that a fairness-satisfying decision policy has on the well-being of different subpopulations. In a similar spirit, Hu & Chen (2020) showed that always preferring "more fair" classifiers does not abide by the Pareto Principle (the principle that a policy must be preferable for at least one of multiple groups) in terms of welfare. Motivated by these findings, our work acknowledges that algorithmic policies affect individuals and institutions in many dimensions, and explicitly encodes these dimensions in policy optimization.

We will show that fairness constrained policies that result in per-group score thresholds and their ϵ-fair equivalent soft-constrained analogs (Elzayn et al., 2019) can be cast as specific instances of the Pareto framework that we study. Analyzing the limitations of this optimization regime with imperfect scores therefore connects to a recent literature on achieving group fairness with noisy or missing group class labels (Lamy et al., 2019; Awasthi et al., 2019), including using proxies of group status (Gupta et al., 2018; Chen et al., 2019). The explicit welfare effects of selection in our model also complement the notion of utilization in fair allocation problems (Elzayn et al., 2019; Donahue & Kleinberg, 2020).

2.2. Multi-objective Machine Learning

We consider two simultaneous goals of a learned classifier: achieving high profit value of the classification policy, while improving a measure of social welfare. This relates to an existing literature on multi-objective optimization in machine learning (Jin & Sendhoff, 2008; Jin, 2006), where many algorithms exist for finding or approximating global optima (Deb & Kalyanmoy, 2001; Knowles, 2006; Désidéri, 2012) under different problem formulations.

Our work studies the Pareto solutions that arise from learned score functions, and is therefore related to, but distinct from a large literature on learning Pareto frontiers directly. Evolutionary strategies are a popular class of approaches to estimating a Pareto frontier from empirical data, as they refine a class of several policies at once (Deb & Kalyanmoy, 2001; Kim & de Weck, 2005). Many of these strategies use surrogate convex loss functions to afford better convergence to solutions. Surrogate functions can be defined over each dimension independently (Knowles, 2006), or as a single function over both objective dimensions (Loshchilov et al., 2010). While surrogate loss functions play an important role in a direct optimization of non-convex utility functions, our framework provides an alternative approach, so long as scores functions can be reliably estimated.

Another class of methods explicitly incorporates models of uncertainty in dual-objective optimization (Peitz & Dellnitz, 2018; Paria et al., 2019). For sequential decision-making, there has been recent work on finding Pareto-optimal policies for reinforcement learning settings (Van Moffaert & Nowé, 2014; Liu et al., 2014; Roijers & Whiteson, 2017). To promote applicability of our work to a variety of real-world domains where noise sources are diverse, and the effects of single policy enactments complex, we first develop a methodology under a noise-free setting, then extend to reasonable forms of error in provided estimates.

2.3. Measures of Social Welfare

The definition and measurement of welfare is an important and complex problem that has received considerable attention in the social science literature (cf. Deaton, 1980; 2016; Stiglitz et al., 2009). There, a standard approach is to sum

up individual measures of welfare, to obtain an aggregate measure of societal welfare. The separability assumption (independent individual scores) is a standard simplifying assumption (e.g. Florio, 2014) that appears in the foundational work of (Pigou, 1920), as well as (Burk, 1938), (Samuelson, 1947), (Arrow, 1963) and (Sen, 1973). Future work may explore alternative social welfare function (e.g. Clark & Oswald, 1996). Our focus is on bringing machine learning to the most common notion of welfare.

3. Problem Setting: Pareto-optimal Policies

We consider a setting in which a centralized policymaker has two simultaneous objectives: to maximize some private return (such as revenue or user engagement), which we generically refer to as *profit*; and to improve a public objective (such as social welfare or user health), which we refer to as *welfare*. The policymaker makes decisions about *individuals*, who are specified by feature vectors $x \in \mathbb{R}^d$. Decision policies are functions that output a randomized decision $\pi(x) \in [0, 1]$ corresponding to the *probability* that an individual with features x is selected. To each individual we associate a value p representing the expected profit to be garnered from approving this individual and w encoding the change in welfare. The profit and welfare objectives are thus expectations over the joint distribution of (w, p, x):

$$\mathcal{U}_{\mathsf{W}}(\pi) = \mathbb{E}[w \cdot \pi(x)] \quad \text{and} \quad \mathcal{U}_{\mathsf{P}}(\pi) = \mathbb{E}[p \cdot \pi(x)]\,. \quad (1)$$

Notice that this aggregate measure of societal welfare is defined as a sum of individual measures of welfare; this is a standard approach in the social science literature (see Section 2.3). While this induces limitations on the form of the welfare function, it affords flexibility when focusing instead on the resulting binary decision, a point we expand on in Section 6.

Given two objectives, one can no longer define a unique optimal policy π. Instead, we focus on policies π which are *Pareto-optimal* (Pareto, 1906), in the sense that they are not strictly dominated by any alternative policy, i.e. there is no π' such that both $\mathcal{U}_{\mathsf{P}}$ and $\mathcal{U}_{\mathsf{W}}$ are strictly larger under π'.

For a general set of policy classes (defined in Proposition A.1[1]), it is equivalent to consider policies that maximize a weighted combination of both objectives. We can thus parametrize the Pareto-optimal policies by $\alpha \in [0, 1]$:

Definition 3.1 (Pareto-optimal policies)**.** *An α-Pareto-optimal policy (for $\alpha \in [0, 1]$) satisfies:*

$$\pi^\star_\alpha \in \operatorname{argmax} \mathcal{U}_\alpha(\pi),$$
$$\mathcal{U}_\alpha(\pi) := (1 - \alpha)\mathcal{U}_{\mathsf{P}}(\pi) + \alpha\mathcal{U}_{\mathsf{W}}(\pi).$$

In the definition above, the maximization of π is taken over the class of randomized policies $\pi(x) \to [0, 1]$. In Section 3.1 we show that when features x can exactly encode scores the optimal policy is a threshold of the scores.

[1] All references starting with letters appear in the appendices.

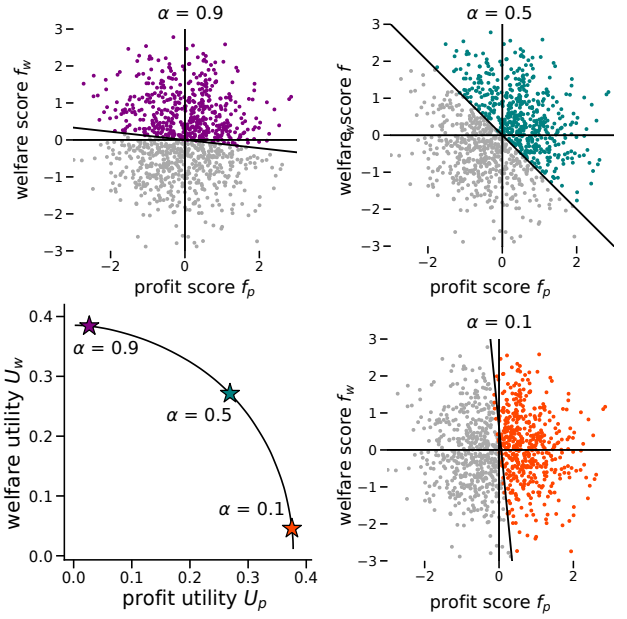


Figure 1. Illustration of a Pareto curve (bottom left) and the decision boundaries induced by three different trade-off parameters α. Colored (darker in gray scale) points indicate selected individuals.

3.1. Optimal Policies with Exact Scores

We briefly consider an idealized setting, where the welfare and profit contributions w and p can be directly determined from the features x via exact *score functions*, $f_{\mathsf{W}}(x) = w$, $f_{\mathsf{P}}(x) = p$. These exact score functions can be thought of as sufficient statistics for the decision: the expected weighted contribution from accepted individuals is described by $((1 - \alpha)p + \alpha w)$. Therefore, one can show (Proposition A.2) that the optimal policy is given by thresholding this composite:

$$\pi^\star_\alpha(p, w) = \mathbb{I}((1 - \alpha)p + \alpha w \geq 0). \quad (2)$$

Though they are all Pareto-optimal, the policies $\pi^\star_\alpha$ induce different trade-offs between the two objectives. The parameter α determines this trade-off, tracing the *Pareto frontier*:

$$\mathcal{P}_{\text{exact}} := \{(\mathcal{U}_{\mathsf{P}}(\pi^\star_\alpha), \mathcal{U}_{\mathsf{W}}(\pi^\star_\alpha)) : \alpha \in [0, 1]\}$$

Figure 1 plots an example of this curve (bottom-left panel) and the corresponding decision rules for three points along it. We note the concave shape of this curve, a manifestation of *diminishing marginal returns*: as a decision policy forgoes profit to increase total welfare, less welfare is gained for the same amount of profit forgone. The notion of diminishing return is formalized in Theorem A.5.

4. Pareto Frontiers with Inexact Scores

In many settings, we typically do not know the profit score p or welfare score w — or the score functions f_{P} and f_{W} — for all individuals a priori. Instead, we might estimate score functions $\widehat{f}_{\mathsf{P}}(x)$ and $\widehat{f}_{\mathsf{W}}(x)$ from data in the hope that these models can provide good predictions on future examples.

We study the class of *score-based policies* that act on the predicted scores:

Definition 4.1 (Score-based policy class)**.**

$$\Pi_{\text{emp}} := \{\, \pi : (\widehat{f}_{\mathsf{P}}(X), \widehat{f}_{\mathsf{W}}(X)) \mapsto [0,1] \,\}$$

Focusing on this class of policies allows us to characterize optimal policies within this class (Section 4.1), derive diagnosable bounds the utility of suboptimal policies (Section 4.3), and relate our results to common fairness criteria (Section. 6). We summarize additional benefits as well as potential limitations of restricting our study to this policy class in Section 7.

4.1. Pareto-optimality for Learned Scores

To characterize Pareto-optimal policies over Π_{emp}, we define the following conditional expectations over the distribution $\mathcal{D}$ of (x, p, w):

$$\overline{\mu}_{\mathsf{P}}(\widehat{f}_{\mathsf{P}}(x), \widehat{f}_{\mathsf{W}}(x)) := \mathbb{E}_{\mathcal{D}}[p \mid \widehat{f}_{\mathsf{P}}(x), \widehat{f}_{\mathsf{W}}(x)],$$
$$\overline{\mu}_{\mathsf{W}}(\widehat{f}_{\mathsf{P}}(x), \widehat{f}_{\mathsf{W}}(x)) := \mathbb{E}_{\mathcal{D}}[w \mid \widehat{f}_{\mathsf{P}}(x), \widehat{f}_{\mathsf{W}}(x)]\,.$$

Intuitively, these values represent our best guesses of p and w, given the predicted scores. We define π_α^{opt} as the threshold policy on the composite of these predictions:

$$\pi_\alpha^{\text{opt}} := \mathbb{I}((1-\alpha)\cdot\overline{\mu}_{\mathsf{P}} + \alpha\cdot\overline{\mu}_{\mathsf{W}} \geq 0).$$

Theorem 4.1 (Pareto frontier in inexact knowledge case)**.** *Given any population distribution $\mathcal{D}$ over (x, p, w) and empirical score functions $\widehat{f}_{\mathsf{W}}$ and $\widehat{f}_{\mathsf{P}}$,*

(i) The policies π_α^{opt} are Pareto optimal over the class Π_{emp}, with $\pi_\alpha^{\text{opt}} \in \operatorname{argmax}_{\pi\in\Pi_{\text{emp}}} \mathcal{U}_\alpha(\pi)$.

(ii) The Pareto frontier $\mathcal{P}(\Pi_{\text{emp}})$ is given by $\{(\mathcal{U}_{\mathsf{P}}(\pi_\alpha^{\text{opt}}), \mathcal{U}_{\mathsf{W}}(\pi_\alpha^{\text{opt}}) : \alpha \in [0,1]\}$. The associated function mapping $\sup_{\pi\in\Pi_{\text{emp}}}\{\mathcal{U}_{\mathsf{W}}(\pi) : \mathcal{U}_{\mathsf{P}}(\pi) = p\}$ is concave and non-increasing in p.

(iii) The empirical frontier $\mathcal{P}_{\text{emp}}$ is dominated by the exact frontier $\mathcal{P}_{\text{exact}}$. That is, if $(p, w_{\text{exact}}) \in \mathcal{P}_{\text{exact}}$ and $(p, w_{\text{emp}}) \in \mathcal{P}_{\text{emp}}$, then $w_{\text{emp}} \leq w_{\text{exact}}$.

Thus an optimal empirical-score based policy can also be realized as a threshold policy (this time of the conditional expectations), and it obeys the same diminishing-returns phenomenon as in the exact score case. One example of score predictors that achieves this optimality is the *Bayes optimal estimators* i.e., $\widehat{f}_{\mathsf{P}}(x) = \mathbb{E}[p \mid x]$ and $\widehat{f}_{\mathsf{W}}(x) = \mathbb{E}[w \mid x]$. We present a proof of Theorem 4.1 in Appendix A.4.

4.2. Plug-in Policies

In general, we may have access to score predictions or the ability to learn them from data, but not a guarantee that the predictions are Bayes' optimal. In the hopes that the predicted scores will suffice, we can define a natural selection rule based on α-defined *plug-in threshold policies*.

Definition 4.2 (Plug-in policy)**.** *For $\alpha \in [0,1]$ and score predictions $\widehat{f}_{\mathsf{P}}(x), \widehat{f}_{\mathsf{W}}(x)$, the α-plug-in policy is:*

$$\pi_\alpha^{\text{plug}}(x) = \mathbb{I}((1-\alpha)\widehat{f}_{\mathsf{P}}(x) + \alpha\widehat{f}_{\mathsf{W}}(x) \geq 0)\,. \quad (3)$$

Since π_α^{opt} requires computing conditional expectations over the distribution $\mathcal{D}$, it will in general will differ from the plug-in policy (3). The following corollary of Theorem 4.1 gives a condition in which π_α^{opt} and π_α^{plug} coincide.

Corollary 4.2. *The plug-in policies π_α^{plug} are optimal in the class Π_{emp} as long as the predicted score functions are* well-calibrated*, in the sense that $\mathbb{E}[p \mid \widehat{f}_{\mathsf{P}}(x), \widehat{f}_{\mathsf{W}}(x)] = \widehat{f}_{\mathsf{P}}(x)$ and $\mathbb{E}[w \mid \widehat{f}_{\mathsf{P}}(x), \widehat{f}_{\mathsf{W}}(x)] = \widehat{f}_{\mathsf{W}}(x)$.*

Proof. In this case, $\bar{\mu}_p = \widehat{f}_{\mathsf{P}}(x)$ and $\bar{\mu}_w = \widehat{f}_{\mathsf{W}}(x)$, so we may invoke Theorem 4.1. □

Under typical conditions (Liu et al., 2019), this form of calibration can be achieved by empirical risk minimization.

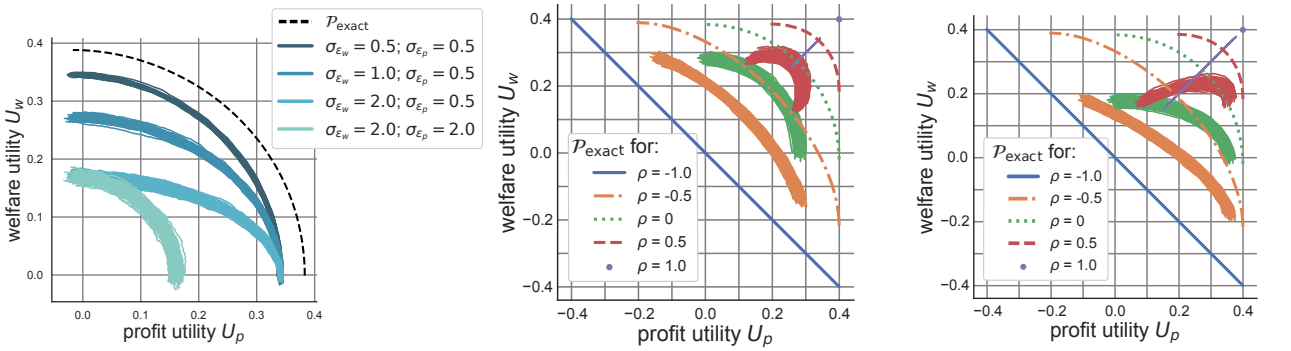
In Section 4.3, we bound the error in the plug-in policies by the error by the individual errors in each score. Simulation experiments in Section 5 detail the use of the plug in policy under controlled degradations of learned score accuracy. Real-data experiments provide further insight into using the plug-in policy for welfare-aware optimization in practice.

4.3. Bounding Pareto Inefficiencies

Even when plug-in policies are not optimal, the sub-optimality of the resulting classifier in terms of the utility function $\mathcal{U}_\alpha$ is bounded by the α-weighted sum of ℓ_1 errors in the profit and welfare scores.

Proposition 4.3 (Sub-optimality Bound)**.** *For any score prediction functions $\widehat{f}_{\mathsf{P}}(x), \widehat{f}_{\mathsf{W}}(x)$ and $\alpha \in [0,1]$, the gap in α-utility from applying the plug-in policy (3) with $\widehat{f}_{\mathsf{P}}(x), \widehat{f}_{\mathsf{W}}(x)$ versus applying the optimal policy (2) with true scores $f_{\mathsf{P}}, f_{\mathsf{W}}$, is bounded above as*

$$\mathcal{U}_\alpha(\pi_\alpha^\star) - \mathcal{U}_\alpha(\pi_\alpha^{\text{plug}}) \leq$$
$$(1-\alpha)\mathbb{E}[|\widehat{f}_{\mathsf{P}}(x) - f_{\mathsf{P}}(x)|] + \alpha\mathbb{E}[|\widehat{f}_{\mathsf{W}}(x) - f_{\mathsf{W}}(x)|]. \quad (4)$$



(a) No correlation, varying noise.

(b) Varying correlation, fixed and equal noise $\sigma_{\varepsilon_p} = \sigma_{\varepsilon_w} = 1$.

(c) Varying correlation, fixed and unequal noise $\sigma_{\varepsilon_p} = 0.5$, $\sigma_{\varepsilon_w} = 2.0$.

Figure 2. Simulated experiments corresponding to the setting in Example 1 (fixing $\sigma_w = \sigma_p = 1$). Empirical frontiers $\mathcal{P}(\Pi_{\text{emp}})$ for 100 random trials with $n = 5,000$ each are shown as overlaid translucent curves. Exact frontiers $\mathcal{P}_{\text{exact}}$ are shown as dashed curves.

Note that by definition of $\pi^\star_\alpha$, $\mathcal{U}_\alpha(\pi^\star_\alpha) - \mathcal{U}_\alpha(\pi^{\text{plug}}_\alpha) \geq 0$. The proof of Proposition 4.3 is given in Appendix A.5.

Proposition 4.3 provides a general bound on the α-performance of the plug-in policy which holds for any distribution on scores and estimator errors. To provide further insight, we consider a specific distributional setting.

Example 1. *Suppose that individuals' true scores are distributed as:*

$$(w_i, p_i) \sim_{i.i.d.} \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\sigma_w^2 & \rho\sigma_w\sigma_p\\ \rho\sigma_w\sigma_p & \sigma_p^2\end{bmatrix}\right) \quad (5)$$

Let the prediction errors $\varepsilon_{p_i} := \hat{p}_i - p_i$ and $\varepsilon_{w_i} := \hat{w}_i - w_i$ be independent of the true scores p_i, w_i, zero-mean, and sub-Gaussian with parameters σ_{ε_p} and σ_{ε_w}, respectively.

This example elucidates how correlation between profit and welfare scores affects the empirical Pareto frontier.

Proposition 4.4. *In the setting of Example 1 with $-1 \leq \rho \leq 1$, $\mathbb{E}[\mathcal{U}_\alpha(\pi^\star_\alpha)] = \frac{\sigma_y}{\sqrt{2\pi}}$ and the expected α-utility of the plug in policy is at least:* [2]

$$\mathbb{E}[\mathcal{U}_\alpha(\pi^{\text{plug}}_\alpha)] \geq \mathbb{E}[\mathcal{U}_\alpha(\pi^\star_\alpha)]\left(1 - \frac{2\cdot\tilde{\sigma}^2}{\tilde{\sigma}^2 + \sigma_y^2}\right) \quad (6)$$

where $\sigma_y^2 = \alpha^2\sigma_w^2 + (1-\alpha)^2\sigma_p^2 + 2\rho\alpha(1-\alpha)\sigma_w\sigma_p$ and $\tilde{\sigma}^2 = 4(\alpha^2\sigma^2_{\varepsilon_w} + (1-\alpha)^2\sigma^2_{\varepsilon_p})$.

The proof of Proposition 4.4 is given in Appendix A.5. This lower bound is in terms of both the optimal α-utility and a discount factor. Because σ_y^2 is increasing in ρ for any $\alpha \in (0,1)$, both of these terms are increasing in ρ. Thus, the expected α-utility of the plug in policy is higher for correlated scores, not only because the optimal α-utility is higher, but also because the discount factor is closer to 1.

[2] The constant on $\tilde{\sigma}^2$ can be reduced to 1 when $\varepsilon_w \perp\!\!\!\perp \varepsilon_p$.

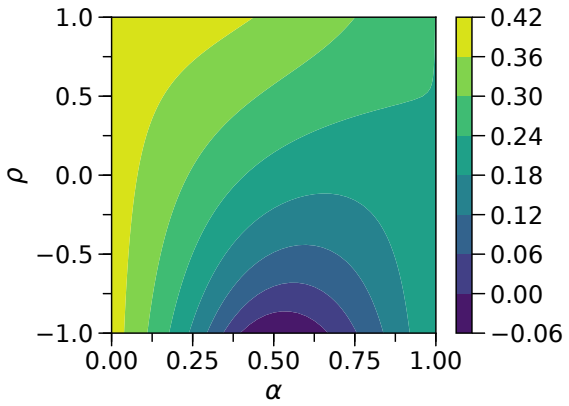


Figure 3. Lower bound (right hand side of Eq. (6)) on expected α-utility as a function of α and correlation in the true scores, from Proposition 4.4, with $\sigma_w = \sigma_p = 1$; $\sigma_{\varepsilon_w} = .5$; $\sigma_{\varepsilon_p} = 0.1$.

Figure 3 shows the lower bound on expected α-utility with noisy scores as a function of possible score correlations ρ and trade-off parameters α, for a fixed setting of predictor noise in Example 1. For comparatively small error in profit scores and moderate welfare error, the lower bound on the α-utility increases as the correlation (ρ) between the scores increases. This captures how the low-noise profit score indirectly improves decisions about the high-noise welfare. The lower bound is decreasing in α for positive ρ, which reflects the higher variance introduced by placing more weight on the noisier welfare score.

5. Experiments

This section presents three sets of empirical results. In Section 5.1 we corroborate our theoretical results under different simulated distributions on scores and prediction errors. Our second experiment studies empirical Pareto frontiers

from learned scores with realistic degradation of training data, in the context of sustainable abalone collection in Section 5.2. Our third experiment in Section 5.3 shows how our methods facilitate trading off between user engagement with predicted quality of content in a corpus of YouTube videos, using pre-learned scores.

5.1. Simulation Experiments

Our first set of simulations shows the performance of the plug-in policy when scores are perturbed by additive noise of varying degrees in each dimension (Fig. 2a). We instantiate true scores w_i and p_i as in Eq. (5) with $\rho = 0$ and $\sigma_w^2 = \sigma_p^2 = 1$, and instantiate predicted scores as:

$$\widehat{f}_{\mathsf{W}}(x_i) = w_i + \varepsilon_{w_i} \quad \varepsilon_{w_i} \sim \mathcal{N}(0, \sigma^2_{\varepsilon_w}), \qquad (7)$$
$$\widehat{f}_{\mathsf{P}}(x_i) = p_i + \varepsilon_{p_i} \quad \varepsilon_{p_i} \sim \mathcal{N}(0, \sigma^2_{\varepsilon_p})$$

These score predictions satisfy the *well-calibrated* condition of Corollary 4.2. The results for different pairs $(\sigma_{\varepsilon_w}, \sigma^2_{\varepsilon_p}$ are shown in Figure 2a. As the noise in scores increases, the empirical Pareto frontiers recede from the exact frontier $\mathcal{P}_{\text{exact}}$. Additionally, higher noise in the predicted scores imposes a wider distribution of empirical Pareto frontiers.

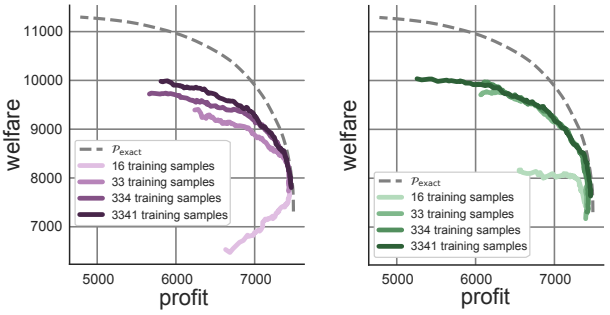
Next, we study the effect of noise in predictions when scores are correlated (Fig. 2b). We draw w_i and p_i according to Eq. (5) with $\sigma_w = \sigma_p = 1$ and correlation parameter ρ. We then add random noise as in Eq. (7) with parameters $\sigma_{\varepsilon_w} = \sigma_{\varepsilon_p} = 1.0$. Note that in this setting, scores are in general not calibrated due to the correlation between w_i and p_i. For positive values of ρ, the exact and empirical utilities are greatest at $\alpha = 0.5$, since the correlation in the scores allows us to overcome some of the noise in each individual parameter, as predicted by Proposition 4.4.

Lastly, we study the space of empirical and exact frontiers with degraded noise when scores are correlated and prediction error is higher in the welfare the score, with $\sigma_{\varepsilon_p} = 0.5$ whereas $\sigma_{\varepsilon_w} = 2.0$ (Fig. 2c). While the optimal Pareto frontiers are the same as in Fig. 2b, we see a stark change in the empirical Pareto frontiers. Compared to the case of no correlation, the empirical Pareto frontier is expanded when $\rho > 0$ and when $\rho < 0$ the frontier recedes. Additionally, we see evidence that due to the correlation, π^{plug}_α is no longer guaranteed to be optimal, as welfare utility decreases for large enough α when $\rho = 0.5$.

5.2. Learned Scores with Imperfect Data: Abalone

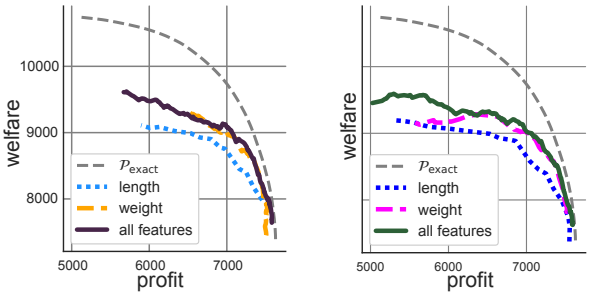
Our next example is motivated by the domain of ecologically sustainable selection, where the goal is to select profitable mollusks to catch and keep, while having minimal impact on the natural development of the mollusks' ecosystem. We learn scores for the age and profitability of each abalone from data, and perform experiments to test the degradation of the empirical Pareto frontiers under realistic degradations of the data. While our characterization of the problem is highly simplified, the main focus of this experiment is to demonstrate the instantiation of Pareto curves for different predictor function classes and different regimes of data availability.

The welfare measure we use is an increasing function of age (see Appendix B for full experimental details), encoding that it is more sustainable to harvest older abalones. We define the profit score of each abalone as a linear function of meat weight and shell area. We use the features (sex, total weight, height, width, and diameter) to train score predictors. We derive these measures from physical data collected by Nash et al. (1994) (accessed via the UCI data repository (Dua & Graff, 2017)). The correlation of the profit and welfare scores is 0.56.



(a) Ridge regression model. (b) Random forest model.

Figure 4. Abalone empirical frontiers as training set size increases.

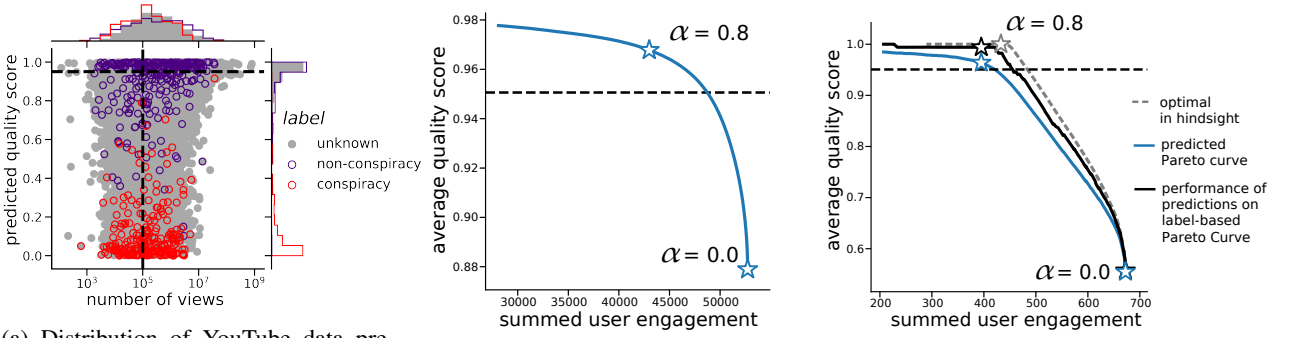


(a) Ridge regression model. (b) Random forest model.

Figure 5. Abalone empirical frontiers for different feature sets.

In this setting, we study the effectiveness of two models — ridge regression and random forests — to learn scores with which to instantiate the plug-in policy. To assess how the empirical Pareto frontiers degrade under realistic notions of imperfect data, we subsample training instances to reflect a hypothetical regime were data is sparse and we subsample features to reflect a hypothetical regime where entire measurements were not recorded in the original dataset.

Figure 4 shows the empirical Pareto frontiers reached as we



(a) Distribution of YouTube data predicted quality scores unlabeled videos (gray), and hand labeled conspiracy (red) and non-conspiracy (purple) videos.

(b) Estimated Pareto curve using precomputed predictions. Stars indicate specific α trade-offs.

(c) Estimated Pareto curve on labeled data subset (blue). Optimal-in-hindsight curve (dashed gray) and performance of predictions on label set (black).

Figure 6. Balancing user engagement and health of hosted YouTube videos.

change the size of the training data set from which learn the profit and welfare scores. Even with 33 training samples (1% of the original training set), the set of plug-in policies traces a meaningful trade-off over α. For severely degraded scores (16 training samples - just 0.5% of the original training sets), the error on the welfare score predictions is so high that instantiating a plug-in policy with $\alpha > 0$ actually decreases welfare overall.

Figure 5 shows the empirical Pareto frontiers reached as we change the features learned to train the model, using just length, just weight, or all seven features as in Fig. 4.

The trends to increasing the data set size and feature set are consistent with four replications done on separate training and evaluation splits; we find that Pareto frontiers dominate each other roughly in accordance with the mean average error of the score predictions (Figs. 9 and 10 in Appendix B.1). The mean average error of welfare scores is substantially greater than the average error of profit scores for most prediction instances (Figures 9 and 10 in Appendix B.1), thus the empirical frontiers are farther from $\mathcal{P}_{\text{exact}}$ in the welfare dimension than the profit dimension.

Altogether, the empirical Pareto frontiers are relatively robust to small data regimes, as well as to missing predictors. However, when predictions have very high error (diagnosable by cross-validation or holdout set error), empirical Pareto frontiers degrade quickly.

5.3. Balancing User Engagement and Health

We now illustrate how the multi-objective framework can be used to balance the desire to promote high quality content with the need for profit. We work with a dataset that contains measures of content quality and content engagement for 39,817 YouTube videos, which was constructed as part of an independent effort to automatically ascertain the quality and truthfulness of YouTube videos (Faddoul et al., 2020).

The measure of quality $\widehat{f}_{\mathsf{W}}$ we use is a function of the 'conspiracy score' developed by Faddoul et al. (2020), which estimates the probability that the video promotes a debunked conspiracy theory. From this score $s_{\text{conspiracy}} \in [0, 1]$ we derive a predicted 'quality score' as $(0.95 - s_{\text{conspiracy}})$ (see Appendix B.2 for details).

We instantiate the profit score $f_{\mathsf{P}}[i]$ for video i as $\log((1 + \#\text{ views}[i])/100,000)$. Dividing by a large constant represents that videos with low view counts may not be profitable due to storage and hosting costs. The resulting distribution over f_{P} and $\widehat{f}_{\mathsf{W}}$ is shown in Figure 6a (gray dots), where dotted lines denote 0-utility thresholds in each score.

Using these scores and predictions, we estimate a Pareto frontier using the optimal policies $\pi_{\alpha}^{\text{plug}}$ for learned scores from Eq. (3). The resulting estimated Pareto curve is shown in Figure 6b. The curve is concave, demonstrating the phenomenon of diminishing returns in the trade-off between total user engagement and average video quality. While there is always some quality to gain by sacrificing some total engagement, these relative gains are greatest when the starting point is close to an engagement-maximizing policy. Specifically, at the maximum-engagement end of the spectrum (lower right star), we can gain a 1.1% increase in average video quality for a 0.1% loss in total engagement. However, for a policy with trade-off rate $\alpha = 0.8$ (upper left star), to obtain an increase of 0.3% in welfare, a larger loss of 5.2% in user engagement is required.

Next, we assess the validity of this estimated Pareto curve using the small set of 541 hand-labeled training set instances from which $s_{\text{conspiracy}}$ was learned. This assessment is likely optimistic due to the fact that the score predictor functions were trained on this same data; nonetheless, this is an important check to perform on the estimated Pareto frontier.

In Figure 6c we plot the optimal-in-hindsight Pareto frontier (dashed gray line) had we known the labels a priori

and applied thresholds according to (2). We also plot the performance of our estimated policy π_α^{plug} on the labeled instances (black line). The stars on each curve correspond to decision thresholds with $\alpha = 0$ and $\alpha = 0.8$, and illustrate the alignment of the curves.

Relating back to Theorem 4.1, we see that performance of the learned scores (black line) is dominated by that of the optimal classifier, as is the predicted Pareto curve (thick blue line). Here the predicted Pareto curve under-predicts the actual performance; in general it is possible for the opposite to be true. Encouragingly, we observe that the curves representing the predicted and actual performance show similar qualitative trade-offs.

6. Connections to Fairness Constraints

Having shown our main results on learning Pareto-optimal policies with limited data, we now illustrate connections between our framework and approaches based on fair machine learning that constrain classification decisions to satisfy certain criteria. For example, in the setting of hiring or admissions, one might require that the same proportion of male and female candidates are admitted, i.e. *demographic parity*. We demonstrate that profit maximization with group fairness constraints corresponds to multi-objective optimization over profit and welfare for an induced definition of welfare. This connection illustrates that even though we consider a welfare function defined from individual welfare scores, our framework can encode more collective conceptions of welfare, like those arising from group fairness constraints.

Consider the setting of requiring demographic parity between two subgroups A and B of a larger population (more general results are presented in Appendix C). In this case, we decompose policies over groups such that $\pi = (\pi_{\mathsf{A}}, \pi_{\mathsf{B}})$. Policies are chosen to maximize the following ϵ-demographic parity constrained problem:

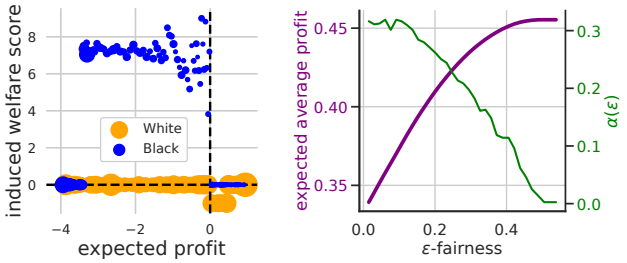
$$\max_{\pi,\beta} \ \mathcal{U}_{\mathsf{P}}(\pi) \ \text{ s.t. } \mathbb{E}[\pi_{\mathsf{j}}(x) \,|x \text{ in group j}] = \beta_{\mathsf{j}}, \quad |\beta_{\mathsf{A}} - \beta_{\mathsf{B}}| \leq \epsilon \tag{8}$$

We can restrict our attention to threshold policies $\pi_{\mathsf{j}}(p) = \mathbb{I}(p \geq t_{\mathsf{j}})$ where t_{j} are group-dependent thresholds (Liu et al., 2018). Notice that the unconstrained solution would simply be $\pi^{\texttt{MaxUtil}}(p) = \mathbb{I}(p \geq 0)$ for all groups. For this reason, we consider groups with $t_{\mathsf{j}} < 0$ as comparatively *disadvantaged* (since their threshold increases in the absence of fairness constraints) and $t_{\mathsf{j}} > 0$ as *advantaged*. Then, the multi-objective framework provides an additional perspective on the trade-offs between ϵ-fairness and profit.

Corollary 6.1. *It is possible to define fixed welfare scores such that the family of inexact fair policies parametrized by any $\epsilon \geq 0$ in (8) corresponds to a family of Pareto-optimal policies parametrized by $\alpha(\epsilon)$. The group-dependent welfare scores are such that $w \geq 0$ for all individuals in the disadvantaged group and $w \leq 0$ in the advantaged group. Furthermore, the induced trade-off parameter $\alpha(\epsilon)$ increases as ϵ decreases.*

Corollary 6.1 follows from Theorem C.3. Fairness constraints can be seen as encoding implicit group-dependent welfare scores for individuals, where members of disadvantaged groups are assigned positive welfare weights and members of advantaged groups assigned negative weights. Figure 7 illustrates this result applied to data from a credit lending scenario from Barocas et al. (2019), where welfare scores are induced for individuals depending on their race and likelihood of repayment. Further details on the generation of these weights are presented in Appendix C. This correspondence is related to the analysis of welfare weights in Hu & Chen (2018), however, our perspective focuses on trade-offs between welfare and profit objectives, in contrast to pure welfare maximization.

In the case that group membership is believed to correspond to the welfare impact of selection, Corollary 6.1 connects our results in Section 4 with a body of work on achieving fairness when group labels are approximate or estimated (Kallus et al., 2020). While some applications may directly call for statistical parity as a criterion, Corollary 6.1 emphasizes the inevitability of fairness constraints as trade-offs between multiple objectives, and frames these trade-offs explicitly in terms of welfare measures.



(a) Distribution of profit and welfare scores. Marker size indicates population sizes.

(b) The fairness parameter ϵ determines the profit trade-off and corresponds to the welfare weight α.

Figure 7. Trade-offs between profit and fairness in lending can be equivalently encoded by a multi-objective framework.

7. Conclusions

We present a methodology for developing welfare-aware policies that jointly optimize a private return (such as profit) with a public objective (such as social welfare). Taking care to consider data-limited regimes, we develop theory around the optimality of using learned predictors to make decisions. Experiments corroborate our theoretical results, showing that thresholding on predicted scores can approach a Pareto-optimal policy.

This score-based approach to balancing competing objectives with noisy data is attractive for several reasons:

- Score-based policies can trade off multiple objectives with scalar predictions, with error bounded by a weighted sum of the errors in the learned scores.
- The plug-in policy is a learned decision rule that is easily explained and diagnosed — in line with the desire for transparent classification rules in practice.
- It provides a crisp and interpretable connection to fair-constrained profit maximization, but reframes the problem as one of multi-objective optimization (see Sec. 6).

While separating the problem of instantiating learned policies from the problem of learning scores has desirable benefits, we note the limitations of this approach as well. First, the plug-in policy is not guaranteed to be the optimal policy learned from data. Thus, when further assumptions on the problem structure are appropriate, it may be worthwhile to consider more general policy classes learned from data. Second, the score-based approach shifts much of the difficulty of welfare-aware machine learning toward defining and predicting welfare, which is an area of active academic and policy debate (Griffin, 1986; Kahneman & Krueger, 2006).

When welfare utilities are estimable, the ability to trade off context-sensitive measures with general policies can improve upon the status quo of applying machine learning policies in welfare-sensitive domains. Further, a multi-objective framework could allow communities to understand the trade-offs between competing definitions of welfare or fairness in data constrained situations.

Taken together, these results help illustrate how machine learning can be used to design policies that prioritize the social impact of an algorithmic decision from the outset, rather than as an afterthought. By elucidating the possible trade-offs between competing objectives, and by illustrating the importance of measurement and prediction error in multi-objective optimization, we hope this work encourages new ways of thinking about welfare-aware machine learning.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant Nos. 1942702 and DGE1752814 , Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. This work was supported by the Bill and Melinda Gates Foundation, the Center for Effective Global Action, and DARPA and NIWC under contract N66001-15-C-4066. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes not withstanding any copyright notation thereon. The views, opinions, and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government. MS is supported by the Open Philanthropy AI Fellowship. LTL is supported by the Open Philanthropy AI Fellowship and the Microsoft Ada Lovelace Fellowship. DB was partly supported by Microsoft Research. MH is a paid consultant for Twitter.

References

Arrow, K. J. *Social Choice and Individual Values*. Number 12. Yale University Press, 1963.

Awasthi, P., Kleindessner, M., and Morgenstern, J. Equalized odds postprocessing under imperfect group information, 2019.

Balkanski, E. and Singer, Y. The sample complexity of optimizing a convex function. In *Conference on Learning Theory*, pp. 275–301, 2017.

Barocas, S. and Selbst, A. D. Big data's disparate impact. *UCLA Law Review*, 2016.

Barocas, S., Hardt, M., and Narayanan, A. *Fairness and Machine Learning*. fairmlbook.org, 2019. http://www.fairmlbook.org.

Burk, A. A reformulation of certain aspects of welfare economics. *The Quarterly Journal of Economics*, 52(2): 310–334, 1938. Publisher: MIT Press.

Chen, J., Kallus, N., Mao, X., Svacha, G., and Udell, M. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 339–348, 2019.

Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5, 2017.

Clark, A. and Oswald, A. Satisfaction and comparison income. *Journal of Public Economics*, 61(3):359–381, 1996. ISSN 0047-2727. URL https://econpapers.repec.org/article/eeepubeco/v_3a61_3ay_3a1996_3ai_3a3_3ap_3a359-381.htm. Publisher: Elsevier.

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

Deaton, A. *The measurement of welfare: Theory and practical guidelines*. World Bank, Development Research Center, 1980.

Deaton, A. Measuring and understanding behavior, welfare, and poverty. *American Economic Review*, 106(6):1221–43, 2016.

Deb, K. and Kalyanmoy, D. *Multi-Objective Optimization Using Evolutionary Algorithms*. John Wiley & Sons, Inc., New York, NY, USA, 2001. ISBN 047187339X.

Désidéri, J.-A. Multiple-gradient descent algorithm (mgda) for multiobjective optimization. *Comptes Rendus Mathematique*, 350(5-6):313–318, 2012.

Donahue, K. and Kleinberg, J. Fairness and utilization in allocating resources with uncertain demand. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 658–668, 2020.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, pp. 214–226, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1115-1. doi: 10.1145/2090236.2090255. URL http://doi.acm.org/10.1145/2090236.2090255.

Elzayn, H., Jabbari, S., Jung, C., Kearns, M., Neel, S., Roth, A., and Schutzman, Z. Fair algorithms for learning in allocation problems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 170–179, 2019.

Ensign, D., Friedler, S. A., Nevlle, S., Scheidegger, C., and Venkatasubramanian, S. Runaway feedback loops in predictive policing. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81, 2018.

Faddoul, M., Chaslot, G., and Farid, H. A longitudinal analysis of youtube's promotion of conspiracy videos. *In Preparation*, 2020.

Fleurbaey, M. and Maniquet, F. Optimal income taxation theory and principles of fairness. *Journal of Economic Literature*, 56(3):1029–79, 2018.

Florio, M. *Applied welfare economics: Cost-benefit analysis of projects and policies*. Routledge, 2014.

Griffin, J. Well-being: Its meaning, measurement and moral importance. 1986.

Gupta, M. R., Cotter, A., Fard, M. M., and Wang, S. Proxy fairness. *CoRR*, abs/1806.11212, 2018. URL http://arxiv.org/abs/1806.11212.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pp. 3315–3323, 2016.

Hu, L. and Chen, Y. Welfare and Distributional Impacts of Fair Classification. *Fairness, Accountability, and Transparency in Machine Learning (FATML). Stockholm, Sweden.*, July 2018. URL http://arxiv.org/abs/1807.01134. arXiv: 1807.01134.

Hu, L. and Chen, Y. Fair classification and social welfare. *ACM FAT**, 2020.

Jin, Y. *Multi-objective machine learning*, volume 16. Springer Science & Business Media, 2006.

Jin, Y. and Sendhoff, B. Pareto-based multiobjective machine learning: An overview and case studies. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(3):397–415, 2008.

Kahneman, D. and Krueger, A. B. Developments in the measurement of subjective well-being. *Journal of Economic perspectives*, 20(1):3–24, 2006.

Kallus, N., Mao, X., and Zhou, A. Assessing algorithmic fairness with unobserved protected class using data combination. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 110–110, 2020.

Kearns, M., Neel, S., Roth, A., and Wu, Z. S. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pp. 2564–2572, 2018.

Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. Avoiding discrimination through causal reasoning. In *In Proc. 30th NIPS*, pp. 656–666, 2017.

Kim, I. Y. and de Weck, O. L. Adaptive weighted-sum method for bi-objective optimization: Pareto front generation. *Structural and multidisciplinary optimization*, 29 (2):149–158, 2005.

Kim, M. P., Ghorbani, A., and Zou, J. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 247–254. ACM, 2019.

Kleinberg, J. M., Mullainathan, S., and Raghavan, M. Inherent trade-offs in the fair determination of risk scores. *Proc. 8th ITCS*, 2017.

Knowles, J. Parego: a hybrid algorithm with on-line landscape approximation for expensive multiobjective optimization problems. *IEEE Transactions on Evolutionary Computation*, 10(1):50–66, 2006.

Komiya, H. Elementary proof for sion's minimax theorem. *Kodai Mathematical Journal*, 11(1):5–7, 1988.

Lamy, A., Zhong, Z., Menon, A. K., and Verma, N. Noise-tolerant fair classification. In *Advances in Neural Information Processing Systems 32*, pp. 294–305. Curran Associates, Inc., 2019.

Liu, C., Xu, X., and Hu, D. Multiobjective reinforcement learning: A comprehensive overview. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(3):385–398, 2014.

Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., and Hardt, M. Delayed impact of fair machine learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3156–3164, Stockholm, Sweden, 2018.

Liu, L. T., Simchowitz, M., and Hardt, M. The implicit fairness criterion of unconstrained learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4051–4060, Long Beach, California, USA, 2019. PMLR.

Liu, L. T., Wilson, A., Haghtalab, N., Kalai, A. T., Borgs, C., and Chayes, J. The disparate equilibria of algorithmic decision making when individuals invest rationally. *ACM FAT**, 2020.

Loshchilov, I., Schoenauer, M., and Sebag, M. A mono surrogate for multiobjective optimization. In *Proceedings of the 12th annual conference on Genetic and evolutionary computation*, pp. 471–478. ACM, 2010.

Mouzannar, H., Ohannessian, M. I., and Srebro, N. From Fair Decision Making to Social Equality. *ACM FAT**, 2019.

Nash, W. J., Sellers, T. L., Talbot, S. R., Cawthorn, A. J., and Ford, W. B. The population biology of abalone (haliotis species) in tasmania. i. blacklip abalone (h. rubra) from the north coast and islands of bass strait. *Sea Fisheries Division, Technical Report*, 48:p411, 1994.

Noriega, A., Garcia-Bulle, B., Tejerina, L., and Pentland, A. Algorithmic fairness and efficiency in targeting social welfare programs at scale. *Bloomberg Data for Good Exchange Conference*, 2018.

Pareto, V. *Manuale di economia politica*, volume 13. Societa Editrice, 1906.

Paria, B., Kandasamy, K., and Póczos, B. A flexible multi-objective bayesian optimization approach using random scalarizations. *Uncertainty in Artificial Intelligence*, 2019.

Peitz, S. and Dellnitz, M. Gradient-based multiobjective optimization with uncertainties. In *NEO 2016*, pp. 159–182. Springer, 2018.

Persily, N. The 2016 US Election: Can democracy survive the internet? *Journal of democracy*, 28(2):63–76, 2017.

Pigou, A. C. *The economics of welfare*. Macmillan London, 1920.

Roijers, D. M. and Whiteson, S. Multi-objective decision making. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 11(1):1–129, 2017.

Samuelson, P. A. *Foundations of Economic Analysis*. Harvard University Press, 1947.

Sen, A. Behaviour and the concept of preference. *Economica*, 40(159):241–259, 1973.

Skiba, P. M. and Tobacman, J. Do Payday Loans Cause Bankruptcy? SSRN Scholarly Paper ID 1266215, Social Science Research Network, Rochester, NY, November 2009. URL https://papers.ssrn.com/abstract=1266215.

Stiglitz, J., Sen, A., and Fitoussi, J.-P. The measurement of economic performance and social progress revisited. *Reflections and overview. Commission on the measurement of economic performance and social progress, Paris*, 2009.

US Federal Reserve. Report to the congress on credit scoring and its effects on the availability and affordability of credit, 2007.

Van Moffaert, K. and Nowé, A. Multi-objective reinforcement learning using sets of pareto dominating policies. *The Journal of Machine Learning Research*, 15(1):3483–3512, 2014.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Zliobaite, I. On the relation between accuracy and fairness in binary classification. *arXiv preprint arXiv:1505.05723*, 2015.