Pré-Publicações do Departamento de Matemática
Universidade de Coimbra
Preprint Number 20–37

ACCURACY AND FAIRNESS TRADE-OFFS IN MACHINE LEARNING: A STOCHASTIC MULTI-OBJECTIVE APPROACH

S. LIU AND L. N. VICENTE

ABSTRACT: In the application of machine learning to real-life decision-making systems, e.g., credit scoring and criminal justice, the prediction outcomes might discriminate against people with sensitive attributes, leading to unfairness. The commonly used strategy in fair machine learning is to include fairness as a constraint or a penalization term in the minimization of the prediction loss, which ultimately limits the information given to decision-makers. In this paper, we introduce a new approach to handle fairness by formulating a stochastic multi-objective optimization problem for which the corresponding Pareto fronts uniquely and comprehensively define the accuracy-fairness trade-offs. We have then applied a stochastic approximation-type method to efficiently obtain well-spread and accurate Pareto fronts, and by doing so we can handle training data arriving in a streaming way.

KEYWORDS: Stochastic approximation, multi-objective optimization, supervised machine learning, fairness, Pareto fronts.
MATH. SUBJECT CLASSIFICATION (2010): 68T07, 68T09, 68T20, 90C15, 90C29.

1. Introduction

Machine learning (ML) plays an increasingly significant role in data-driven decision making, e.g., credit scoring, college admission, hiring decisions, and criminal justice. As the learning models became more and more sophisticated, concern regarding fairness started receiving more and more attention. In 2014, the Obama Administration's Big Data Report [24] claimed that discrimination against individuals and groups might be the "inadvertent outcome of the way big data technologies are structured and used". Two years later, a White House report [1] on the challenges of big data emphasized the necessity of promoting fairness and called for equal opportunity in insurance, education, employment, and other sectors.

In supervised machine learning, training samples consist of pairs of feature vectors (containing a number of features that are descriptive of each instance) and target values/labels. One tries to determine an accurate predictor, seen

Received December 28, 2020.
Support for LNV was partially provided by the Centre for Mathematics of the University of Coimbra under grant FCT/MCTES UIDB/MAT/00324/2020.

as a function mapping feature vectors into target labels. Such a predictor is typically characterized by a number of parameters, and the process of identifying the optimal parameters is called training or learning. The trained predictor can then be used to predict labels for unlabeled instances.

If a ML predictor does inequitably treat people from different groups defined by *sensitive* or *protected* attributes, such as gender, race, country, or disability, we say that such a predictor is *unfair*. The sources of unfairness in supervised ML are twofold. Firstly, the ML predictors are trained on data collected by humans (or automated agents developed by humans), which may contain inherent biases. Hence, by learning from biased or prejudiced targets, the prediction results obtained from standard learning processes can hardly be unbiased. Secondly, even if the targets are unbiased, the learning process may sacrifice fairness, as the main goal of ML is to make predictions as accurate as possible. In fact, previous research work [22, 30] has showed that simply excluding sensitive attributes from features data (also called *fairness through unawareness*) does not help due to the fact that the sensitive attributes can be inferred from the nonsensitive ones.

Hence, a proper framework for evaluating and promoting fairness in ML becomes indispensable and relevant. Depending on when the fairness criteria are imposed, there are three categories of approaches proposed to handle fairness, namely pre-processing, in-training, and post-processing. Pre-processing approaches [5, 30] modify the input data representation so that the prediction outcomes from any standard learning process become fair, while post-processing [13, 23] tries to adjust the results of a pre-trained predictor to increase fairness while maintaining the prediction accuracy as much as possible. Assuming that the sensitive attributes information are accessible in the training samples, most of in-training methods [2, 3, 15, 29, 27, 28] enforce fairness during the training process either by directly imposing fairness constraints and solving constrained optimization problems or by adding penalization terms to the learning objective.

The approach proposed in our paper falls into the in-training category. We will however explicitly recognize the presence of at least two conflicting objectives in fair machine learning: (1) maximizing prediction accuracy; (2) maximizing fairness (w.r.t. certain sensitive attributes).

1.1. Existing Fairness Criteria in Machine Learning. Fairness in machine learning basically requires that prediction outcomes do not disproportionally benefit people from majority and minority or historically advantageous and disadvantageous groups. In the literature of fair machine learning, several prevailing criteria for fairness include *disparate impact* [2] (also called *demographic parity* [3]), *equalized odds* [13], and its special case of *equal opportunity* [13], corresponding to different aspects of fairness requirements.

In this paper, we will focus on *binary classification* to present the formula for fairness criteria and the proposed accuracy and fairness trade-off framework, although they can all be easily generalized to other ML problems (such as regression or clustering). We point out that many real decision-making problems such as college admission, bank loan application, hiring decisions, etc. can be formulated into binary classification models.

Let $Z \in \mathbb{R}^n, A \in \{0, 1\}, Y \in \{-1, +1\}$ denote feature vector, binary-valued sensitive attribute (for simplicity we focus on the case of a single binary sensitive attribute), and target label respectively. Consider a general predictor $\hat{Y} \in \{-1, +1\}$ which could be a function of both Z and A or only Z. The predictor is free of disparate impact [2] if the prediction outcome is statistically independent of the sensitive attribute, i.e., for $\hat{y} \in \{-1, +1\}$,

$$\mathbb{P}\{\hat{Y} = \hat{y} | A = 0\} \;=\; \mathbb{P}\{\hat{Y} = \hat{y} | A = 1\}. \tag{1}$$

However, disparate impact could be unrealistic when one group is more likely to be classified as a positive class than others, an example being that women are more dominating in education and healthcare services than men [16]. As a result, disparate impact may never be aligned with a perfect predictor $\hat{Y} = Y$.

In terms of equalized odds [13], the predictor is defined to be fair if it is independent of the sensitive attribute but conditioning on the true outcome Y, namely for $y, \hat{y} \in \{-1, +1\}$,

$$\mathbb{P}\{\hat{Y} = \hat{y} | A = 0, Y = y\} \;=\; \mathbb{P}\{\hat{Y} = \hat{y} | A = 1, Y = y\}. \tag{2}$$

Under this definition, a perfectly accurate predictor can be possibly defined as a fair one, as the probabilities in (2) will always coincide when $\hat{Y} = Y$. Equal opportunity [13], a relaxed version of equalized odds, requires that condition (2) holds for only positive outcome instances ($Y = +1$), for example, students admitted to a college and candidates hired by a company.

1.2. Our Contribution. From the perspective of multi-objective optimization (MOO), most of the in-training methods in the literature [2, 3, 15, 27, 28, 29] are based on the so-called *a priori* methodology, where the decision-making preference regarding an objective (the level of fairness) must be specified before optimizing the other (the accuracy). For instance, the constrained optimization problems proposed in [28, 29] are to some extent nothing else than the ϵ–constraint method [12] in MOO. Such procedures highly rely on the decision-maker's advanced knowledge of the magnitude of fairness, which may vary from criterion to criterion and from dataset to dataset.

In order to better frame our discussion of accuracy vs fairness, let us introduce the general form of a multi-objective optimization problem

$$\min \quad F(x) \;=\; (f_1(x), \ldots, f_m(x)), \tag{3}$$

with m objectives, and where $F : \mathbb{R}^n \to \mathbb{R}^m$. Usually, there is no single point optimizing all the objectives simultaneously. The notion of *dominance* is used to define optimality in MOO. A point x is said to be nondominated if $F(y) \nleq F(x)$ holds element-wise for any other point y. An unambiguous way of considering the trade-offs among multiple objectives is given by the so-called *Pareto front*, which lies in the criteria space $\mathbb{R}^m$ and is defined as the set of points of the form $F(x)$ for all nondominated points x.

In this paper, instead of looking for a single predictor that satisfies certain fairness constraints, our goal is to directly construct a complete Pareto front between prediction accuracy and fairness, and thus to identify a set of predictors associated with different levels of fairness. We propose a stochastic multi-objective optimization framework, and aim at obtaining good approximations of true Pareto fronts. We summarize below the three main advantages of the proposed framework.

- By applying an algorithm for stochastic multi-objective optimization (such as the Pareto front stochastic multi-gradient (PF-SMG) algorithm developed in [21]), we are able to obtain well-spread and accurate Pareto fronts in a flexible and efficient way. The approach works for a variety of scenarios, including binary and categorical multi-valued sensitive attributes. It also handles multiple objectives simultaneously, such as multiple sensitive attributes and multiple fairness measures. Compared to the constrained optimization approaches, e.g., [28, 29], our framework is proved to be computational efficient in constructing the whole Pareto fronts.

- The proposed framework is quite general in the sense that it has no restriction on the type of predictors and works for any convex or non-convex smooth objective functions. In fact, it can not only handle the fairness criteria mentioned in Section 1.1 based on covariance approximation, but also tackle other formula proposed in the literature, e.g., mutual information [14] and fairness as a risk measure [26].
- The PF-SMG algorithm falls into a Stochastic Approximation (SA) algorithmic approach, and thus it enables us to deal with the case where the training data is arriving on a streaming mode. By using such an SA framework, there is no need to reconstruct the Pareto front from scratch each time new data arrives. Instead, a Pareto front constructed based on consecutive arriving samples will eventually converge to the one corresponding to the overall true population.

The remainder of this paper is organized as follows. Our stochastic bi-objective formulation using disparate impact is suggested in Section 2. The PF-SMG algorithm, used to solve the multi-objective problems, is briefly introduced in Section 3 (more details in Appendix B). A number of numerical results for both synthetic (Subsection 4.1) and real data (Subsection 4.2) are presented in Section 4 to support our claims. Further exploring our line of thought, we introduce another stochastic bi-objective formulation, this time for trading-off accuracy vs equal opportunity (see Section 5), also reporting numerical results. In Section 6, we show how to handle multiple sensitive attributes and multiple fairness measures. For the purpose of getting more insight on the various trade-offs, two tri-objective problems are formulated and solved. Finally, a preliminary numerical experiment described in Section 7 will illustrate the applicability of our approach to streaming data. The paper is ended with some conclusions and prospects of future work in Section 8.

2. The Stochastic Bi-Objective Formulation Using Disparate Impact

Given that disparate impact is the most commonly used fairness criterion in the literature, we will first consider disparate impact in this section to present a stochastic bi-objective fairness and accuracy trade-off framework.

In our setting, the training samples consist of nonsensitive feature vectors Z, a binary sensitive attribute A, and binary labels Y. Assume that we have access to N samples $\{z_j, a_j, y_j\}_{j=1}^N$ from a given database. Let the binary predictor $\hat{Y} = \hat{Y}(Z; x) \in \{-1, +1\}$ be a function of the parameters x, and only learned from the nonsensitive feature Z.

Recall that the predictor $\hat{Y}$ is free of disparate impact if it satisfies equation (1). A general measurement of disparate impact, the so-called CV score [4], is defined by the maximum gap between the probabilities of getting positive outcomes in different sensitive groups, i.e.,

$$\mathrm{CV}(\hat{Y}) \;=\; |\mathbb{P}\{\hat{Y} = 1 | A = 0\} - \mathbb{P}\{\hat{Y} = 1 | A = 1\}|. \tag{4}$$

The trade-offs between prediction accuracy and fairness can then be formulated as a general stochastic bi-objective optimization problem as follows

$$\min \; f_1(x) \; = \mathbb{E}[\ell(\hat{Y}(Z; x), Y)], \tag{5}$$

$$\min \; f_2(x) \; = \mathrm{CV}(\hat{Y}(Z; x)), \tag{6}$$

where the first objective (5) is a composition function of a loss function $\ell(\cdot, \cdot)$ and the prediction function $\hat{Y}(Z; x)$, and the expectation is taken over the joint distribution of Z and Y.

The logistic regression model is one of the classical prediction models for binary classification problems. For a given feature vector z_i and corresponding true label y_i, one searches for a separating hyperplane $\phi(z_j; x) = \phi(z_j; c, b) = c^\top z_j + b$ such that (noting $x = (c, b)^\top$)

$$\begin{cases} c^\top z_j + b \;\geq\; 0 & \text{when } y_j \;=\; +1, \\ c^\top z_j + b \;<\; 0 & \text{when } y_j \;=\; -1. \end{cases}$$

The predictor defined by the separating hyperplane is known as the threshold classifier, i.e., $\hat{Y}(z_j; c, b) \;=\; 2 \times \mathbf{1}(c^\top z_j + b \geq 0) - 1$. The logistic loss function of the form $\ell(z, y; c, b) = \log(1 + \exp(-y(c^\top z + b)))$ is a smooth and convex version of the classical 0–1 loss. The first objective can then be approximated by the empirical logistic regression loss, i.e.,

$$f_1(c, b) \;=\; \frac{1}{N} \sum_{j=1}^{N} \log(1 + \exp(-y_j(c^\top z_j + b))), \tag{7}$$

based on N training samples. A regularization term $\frac{\lambda}{2}\|c\|^2$ can be added to avoid over-fitting.

Dealing with the second objective (6) is challenging since it is nonsmooth and nonconvex. Hence, we make use of the *decision boundary covariance* proposed by [29] as a convex approximate measurement of disparate impact. Specifically, the CV score (4) can be approximated by the empirical covariance between the sensitive attributes A and the hyperplane $\phi(Z; c, b)$, i.e.,

$$\begin{aligned} &\mathrm{Cov}(A, \phi(Z; c, b)) \\ &= \mathbb{E}[(A - \bar{A})(\phi(Z; c, b) - \overline{\phi(Z; c, b)})] \\ &= \mathbb{E}[(A - \bar{A})\phi(Z; c, b)] - \mathbb{E}[A - \bar{A}]\overline{\phi(Z; c, b)} \\ &\simeq \tfrac{1}{N} \textstyle\sum_{j=1}^{N} (a_j - \bar{a})\phi(z_j; c, b), \end{aligned}$$

where $\bar{A}$ is the expected value of the sensitive attribute, and $\bar{a}$ is an approximated value of $\bar{A}$ using N samples. The intuition behind this approximation is that the disparate impact (1) basically requires the predictor completely independent from the sensitive attribute.

Given that zero covariance is a necessary condition for independence, the second objective can be approximated as:

$$f_2^{\mathrm{DI}}(c, b) \; = \; \left(\frac{1}{N} \sum_{j=1}^{N} (a_j - \bar{a})(c^\top z_j + b) \right)^2, \tag{8}$$

which, as we will see later in the paper, is monotonically increasing with disparate impact. We were thus able to construct a finite-sum bi-objective problem

$$\min \; \left(f_1(c, b), f_2^{\mathrm{DI}}(c, b) \right), \tag{9}$$

where both functions are now convex and smooth.

3. The Stochastic Multi-Gradient Method and Its Pareto Front Version

Consider again a stochastic MOO of the same form as in (3), where some or all of the objectives involve uncertainty. Denote by $g_i(x, w)$ a stochastic gradient of the i-th objective function, where w indicates the batch of samples used in the estimation. The stochastic multi-gradient (SMG) algorithm is described in Algorithm 1 (see Appendix A). It essentially takes a step along the stochastic multi-gradient $g(x_k, w_k)$ which is a convex linear combination of $g_i(x, w)$, $i = 1, \ldots, m$. The SMG method is a generalization of stochastic gradient (SG) to multiple objectives. It was first proposed by [25] and further analyzed by [21]. In the latter paper it was proved that the SMG algorithm has the same convergence rates as SG (although now to a nondominated point), for both convex and strongly convex objectives. As we said before, when $m = 1$ SMG reduces to SG. When $m > 1$ and the f's are deterministic, $-g(x_k) = -g(x_k, w_k)$ is the direction that is the most descent among all the m functions [10, 11].

Note that the two smooth objective functions (7) and (8) are both given in a finite-sum form, for which one can efficiently compute stochastic gradients using batches of samples.

To compute good approximations of the entire Pareto front in a single run, we use the Pareto Front SMG algorithm (PF-SMG) developed by [21]. PF-SMG essentially maintains a list of nondominated points using SMG updates. It solves stochastic multi-objective problems in an *a posteriori* way, by determining Pareto fronts without predefining weights or adjusting levels of preferences. One starts with an initial list of randomly generated points (5 in our experiments).

At each iteration of PF-SMG, we apply SMG multiple times at each point in the current list, and by doing so one obtains different final points due to stochasticity. At the end of each iteration, all the dominated points are removed to get a new list for the next iteration (see Appendix B for an illustration). The process can be stopped when either the number of nondominated points is greater than a certain budget (1,500 in our experiments) or when the total number of SMG iterates applied in any trajectory exceeds a certain budget (1,000 in our experiments). We refer to the paper [21] for more details.

4. Numerical Results for Disparate Impact

To numerically illustrate our approach based on the bi-objective formulation (9), we have used synthetic data and the *Adult Income* dataset [17], which is available in the UCI Machine Learning Repository [9].

There are several parameters to be tuned in PF-SMG for a better performance: (1) p_1: number of times SMG is applied at each point in the current list; (2) p_2: number of SMG iterations each time SMG is called; (3) $\{\alpha_k\}_1^T$: step size sequence; (4) $\{b_{1,k}\}_1^T, \{b_{2,k}\}_1^T$: batch size sequences used in computing stochastic gradients for the two objectives. To control the rate of generated nondominated points, we remove nondominated points from regions where such points tend to grow too densely.

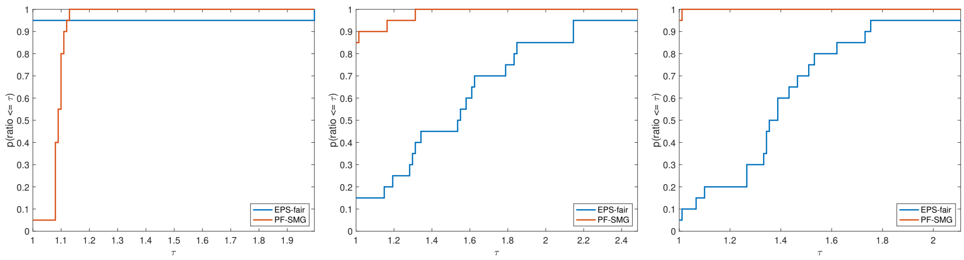
4.1. Synthetic Data. Using synthetic data, our approach is first compared to the ϵ-constrained optimization model proposed in [29, Equation (4)]. From now on, we note their ϵ-constrained method as EPS-fair. It basically minimizes prediction loss subject to disparate impact being bounded above by a constant ϵ, i.e.,

$$\min \;(7) \quad \text{s.t.} \quad |\tfrac{1}{N}\sum_{j=1}^{N}(a_j - \bar{a})\phi(z_j; c, b)| \leq \epsilon.$$

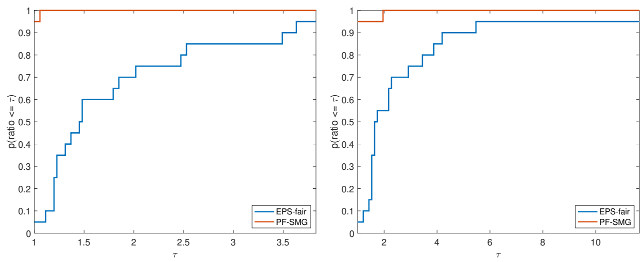
Since the bi-objective problem (9) under investigation is convex, EPS-fair is able to compute a set of nondominated points by varying the value of ϵ. The implementation details of EPS-fair method can be found in [29]. First, by solely minimizing prediction loss, a reasonable upper bound is obtained for disparate impact. Then, to obtain the Pareto front, a sequence of thresholds ϵ is evenly chosen from 0 to such an upper bound, leading to a set of convex constrained optimization problems. The Sequential Least SQuares Programming (SLSQP) solver [18] based on Quasi-Newton methods is then used for solving those problems. We found that 70-80% of the final points produced by this process were actually dominated ones, and we removed them for the purpose of analyzing results.

The synthetic data is formed by 20 sets of 2,000 binary classification data instances randomly generated from the same distributions setting specified in [29, Section 4], specifically using an uniform distribution for generating binary labels Y, two different Gaussian distributions for generating 2-dimensional nonsensitive features Z, and a Bernoulli distribution for generating the binary sensitive attribute A. We evaluated the performance of the two approaches by comparing CPU time, number of gradient evaluations, and the quality of Pareto fronts. Such a quality is measured by a formula called *purity* (which tries to evaluate how the fronts under analysis dominate each other) and two formulas for the spread of the fronts (Γ and Δ, measuring how well the nondominated points on a Pareto front are distributed). Higher purity corresponds to higher accuracy, while smaller Γ and Δ indicate better spread. The detailed formulas of the three measures are given in Appendix C.

The five performance profiles (see [8]) are shown in Figure 1. The purity (see (a)) of the Pareto fronts produced by the EPS-fair method is only slightly better than the one of those determined by PF-SMG. However, notice that PF-SMG produced better spread fronts than EPS-fair without compromising accuracy too much (see (b)–(c)). In addition, PF-SMG outperforms EPS-fair in terms of computational cost quantified by CPU time and gradient evaluations (see (d)–(e)).



(a) Purity. (b) Spread: Γ. (c) Spread: Δ.



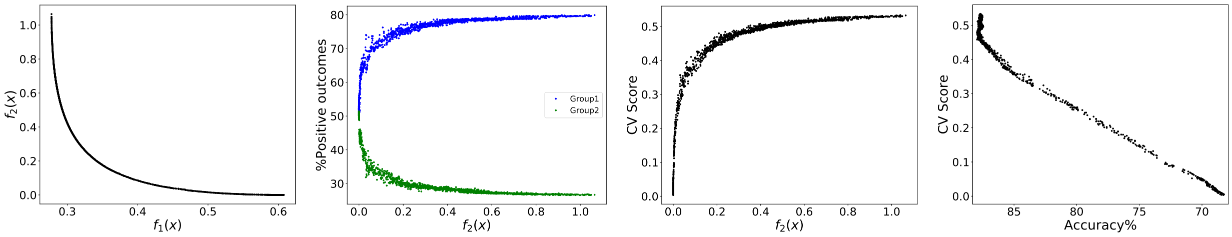
(d) CPU time. (e) #Gradient evaluations.

FIGURE 1. Performance profiles for 20 synthetic datasets: PF-SMG versus EPS-fair. Parameters used in PF-SMG: $p_1 = 1$, $p_2 = 1$, $\alpha_k = 0.3$, $b_{1,k} = 5 \times 1.01^k$, and $b_{2,k} = 200 \times 1.01^k$.

Figure 2 gives the detailed trade-off results for one of the synthetic data sets. The Pareto front in (a) confirms the conflict between two objectives. Given a nondominated solution $x = (c, b)$ from (a), the probability of getting positive prediction for each sensitive group is approximated by the percentage of positive outcomes for the data samples, i.e.,

$$\mathbb{P}\{\hat{Y}(Z;x) = 1 | A = a\} \simeq \frac{N(\hat{Y} = 1, A = a)}{N(A = a)},$$

where $N(\hat{Y} = 1, A = a)$ denotes the number of instances predicted as positive in group a and $N(A = a)$ is the number of instances in group a. For conciseness, we will only compute the proportion of positive outcomes for analysis. Figure 2 (b) presents how the proportions of positive outcomes for the two groups change over f_2^{DI}. As the covariance goes to zero, one can observe a smaller gap between the percentages of positive outcomes. Furthermore, Figure 2 (c) confirms that the value of f_2^{DI} is monotonically increasing with CV score and hence a good approximation of disparate impact. The last plot in Figure 2 indicates that requiring lower CV scores results in lower prediction accuracy.



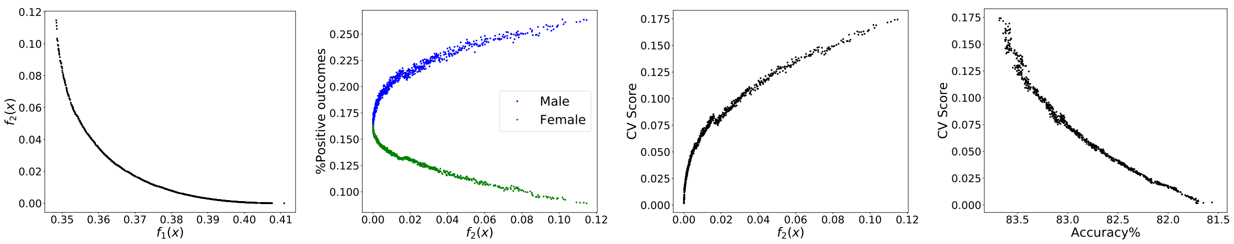
(a) Pareto front. (b) $f_2(x)$ vs %pos. outcomes. (c) $f_2(x)$ vs CV score. (d) Accuracy vs CV score.

FIGURE 2. Trade-off results for synthetic data. Parameters used in PF-SMG: $p_1 = 1$, $p_2 = 1$, $\alpha_k = 0.3$, $b_{1,k} = 5 \times 1.01^k$, and $b_{2,k} = 200 \times 1.01^k$.

4.2. Real Datasets. The cleaned up version of *Adult Income* dataset contains 45,222 samples. Each instance is characterized by 12 nonsensitive attributes (including age, education, marital status, and occupation), a binary sensitive attribute (gender), and a multi-valued sensitive attribute (race). The prediction target is to determine whether a person makes over 50K per year. Tables 1 and 2 in Appendix D show the detailed demographic composition of the dataset with respect to gender and race.

In the following experiment, we have randomly chosen 5,000 training instances, using the remaining instances as the testing dataset. The PF-SMG algorithm is applied using the training dataset, but all the Pareto fronts and the corresponding trade-off information will be presented using the testing dataset.

Considering gender as the sensitive attribute, the obtained Pareto front is plotted in Figure 3 (a), reconfirming the conflicting nature of the two objectives. It is observed from (b) that as f_2^{DI} increases, the proportion of high income adults in females decreases, which means the predictors of high accuracy are actually unfair for females. Similar to the results for synthetic data, from (c) we can conclude that the value of f_2^{DI} has positive correlation with CV score for this dataset. Figure 3 (d) implies that zero disparate impact can be achieved by reducing 2% of accuracy (the range of the x-axis is nearly 2%). To eliminate the impact of the fact that female is a minority in the dataset, we ran the algorithms for several sets of training samples with 50% females and 50% males. It turns out that the conflict is not alleviated at all.



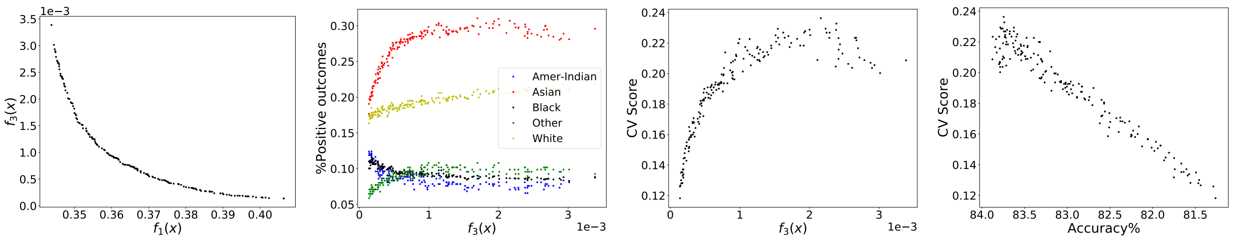
(a) Pareto front. (b) $f_2^{DI}(x)$ vs %pos. outcomes. (c) $f_2^{DI}(x)$ vs CV score. (d) Accuracy vs CV score.

FIGURE 3. Trade-off results for Adult Income dataset w.r.t. gender. Parameters used in PF-SMG: $p_1 = 2, p_2 = 3, \alpha_0 = 2.1$ and then multiplied by 1/3 every 500 iterates of SMG, and $b_{1,k} = b_{2,k} = 80 \times 1.01^k$.

Dealing with multi-valued sensitive attribute race is more complicated. In general, if a multi-valued sensitive attribute has K categorical values, we convert it to K binary attributes denoted by $A^1, \ldots, A^K \in \{0, 1\}$. Note that the binary attribute A^i indicates whether the original sensitive attribute has i-th categorical value or not. The second objective is then modified as follows

$$f_3^{\mathrm{DI}}(c, b) = \max_{i=1,\ldots,K} \left(\frac{1}{N} \sum_{j=1}^{N} (a_j^i - \bar{a}^i)(c^\top z_j + b) \right)^2, \tag{10}$$

which is still a convex function. We have observed that the non-smoothness introduced by the max operator in (10) led to more discontinuity in the true trade-off curves, and besides stochastic gradient type methods are designed for smooth objective functions. We have thus approximated the max operator in (10) using $S_\beta(\max(x^1, \ldots, x^\ell)) = \sum_{i=1}^{\ell} x^i e^{\beta x^i} / \sum_{i=1}^{\ell} e^{\beta x^i}$. In our experiments, we set $\beta = 8$. Figure 4 (a) plots the obtained Pareto front of the bi-objective problem of $\min(f_1(c, b), f_3^{\mathrm{DI}}(c, b))$. Figure 4 (b) implies that solely optimizing over prediction accuracy might result in unfair predictors for American-Indian, Black, and Other. Regardless of the noise, it is observed that the value of f_3^{DI} is increasing with CV score (Figure 4 (c)) and that the prediction accuracy and CV score have positive correlation (Figure 4 (d)). Note that CV score in this case was computed as the absolute difference between maximum and minimum proportions of positive outcomes among K groups.



(a) Pareto front. (b) $f_3^{DI}(x)$ vs %pos. outcomes. (c) $f_3^{DI}(x)$ vs CV score. (d) Accuracy vs CV score.

FIGURE 4. Trade-off results for Adult dataset w.r.t. race. Parameters used in PF-SMG: $p_1 = 3, p_2 = 2, \alpha_0 = 2.6$ and multiplied by 1/3 every 100 iterates of SMG, $b_{1,k} = 50 \times 1.005^k$, and $b_{2,k} = 80 \times 1.005^k$.

5. Equal Opportunity

Recall that equal opportunity focuses on positive outcomes $Y = +1$ and requires the following for $\hat{y} \in \{-1, +1\}$

$$\mathbb{P}\{\hat{Y} = \hat{y} | A = 0, Y = +1\} \; = \; \mathbb{P}\{\hat{Y} = \hat{y} | A = 1, Y = +1\}.$$

When $\hat{y} = -1$ in the above equation, this condition essentially suggests equalized *false negative rate* (FNR) across different groups. Similarly, the case of $\hat{y} = +1$ corresponds to equalized *true positive rate* (TPR). Given that $\mathrm{FNR} + \mathrm{TPR} = 1$ always holds, we will focus on the $\hat{y} = -1$ case where qualified candidates are falsely classified in a negative class by the predictor $\hat{Y}$.

For simplicity, let $\mathrm{FNR}_a(\hat{Y}) = \mathbb{P}\{\hat{Y} = -1 | A = a, Y = +1\}$, $a \in \{0, 1\}$. The CV score associated with equal opportunity is now defined as follows

$$\mathrm{CV}_{\mathrm{FNR}}(\hat{Y}) = |\mathrm{FNR}_0(\hat{Y}) - \mathrm{FNR}_1(\hat{Y})|. \tag{11}$$

Since equalized FNR indicates statistical independence between sensitive attributes and instances that have positive targets but falsely predicted as negative, $\mathrm{CV}_{\mathrm{FNR}}(\hat{Y})$ could thus be approximated [28] by

$$\mathrm{Cov}(A, \psi(Z, Y; c, b)) \simeq \tfrac{1}{N} \sum_{j=1}^{N} (a_j - \bar{a}) \psi(z_j, y_j; c, b),$$

where $\psi(z, y; c, b) = \min\{0, \frac{(1+y)}{2} y \phi(z; c, b)\}$. Here, $(1 + y)/2$ excludes truly negative instances $y = -1$ and $y\phi(z, y; c, b) < 0$ implies wrong prediction. Similar to (8), the objective function for equalized FNR is given by

$$f_4^{\mathrm{FNR}}(c, b) = \left(\tfrac{1}{N} \sum_{j=1}^{N} (a_j - \bar{a}) \psi(z_j, y_j; c, b) \right)^2,$$

which is a nonconvex finite-sum function. (Note that as in (10) we have also smoothed here the min operator in $\psi(z, y; c, b)$.) Now, the finite-sum bi-objective problem becomes

$$\min \ \left(f_1(c, b), f_4^{\mathrm{FNR}}(c, b) \right). \tag{12}$$

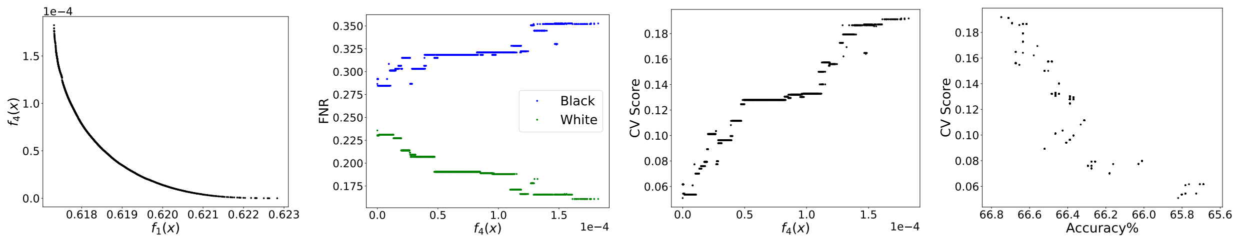
The ProPublica COMPAS dataset [20] contains features that are used by COMPAS algorithms [19] for scoring defendants together with binary labels indicating whether or not a defendant recidivated within 2 years after the screening. For analysis, we take blacks and whites from the *two-years-violent* dataset (see the link in the reference [20]) and consider features including gender, age, number of prior offenses, and charge for which the person was arrested. For consistency with the word "opportunity", we marked the case where a defendant is non-recidivist as the positive outcome. The demographic composition of the dataset is given in Table 3 in Appendix D. Due to shortage of data, we use the whole dataset for both training and testing.

By applying PF-SMG to the bi-objective problem (12), we obtained the trade-off results in Figure 5. The conflicting nature of prediction loss and equalized FNR is confirmed by the Pareto front in Figure 5 (a). For each nondominated solution x, we approximated FNR using samples by

$$\mathrm{FNR}_a(\hat{Y}(Z; x)) \simeq \frac{N(\hat{Y}(Z; x) = -1, A = a, Y = +1)}{N(A = a, Y = +1)},$$

where $N(\cdot)$ is the number of instances satisfying all the conditions.

From the rightmost part of (b), we can draw a similar conclusion as in [19] that black defendants (blue curve) who did not reoffend are accidentally predicted as recidivists twice as often as white defendants (green curve) when using the most accurate predictor obtained (i.e., 0.35 versus 0.175). However, the predictor associated with zero covariance (see the leftmost part) mitigates the situation to 0.28 versus 0.23, although by definition the two rates should converge to the same point. This is potentially due to the fact that the covariance is not well approximated using a limited number of samples. In fact, the leftmost part of Figure 5 (c) shows that zero covariance does not correspond to zero $\mathrm{CV_{FNR}}$. Finally, Figure 5 (d) provides a rough confirmation of positive correlation between CV score and prediction accuracy.



(a) Pareto front. (b) $f_4^{FNR}(x)$ vs FNR. (c) $f_4^{FNR}(x)$ vs CV score. (d) Accuracy vs CV score.

FIGURE 5. Trade-off results for COMPAS dataset w.r.t. race. Parameters used in PF-SMG: $p_1 = 3, p_2 = 3, \alpha_0 = 4$ and multiplied by 1/3 every 100 iterates of SMG, and $b_{1,k} = b_{2,k} = 80 \times 1.005^k$.

The results for equal opportunity presented in this section show the applicability of our multi-objective optimization framework when dealing with nonconvex fairness measures.

6. Handling Multiple Sensitive Attributes and Multiple Fairness Measures

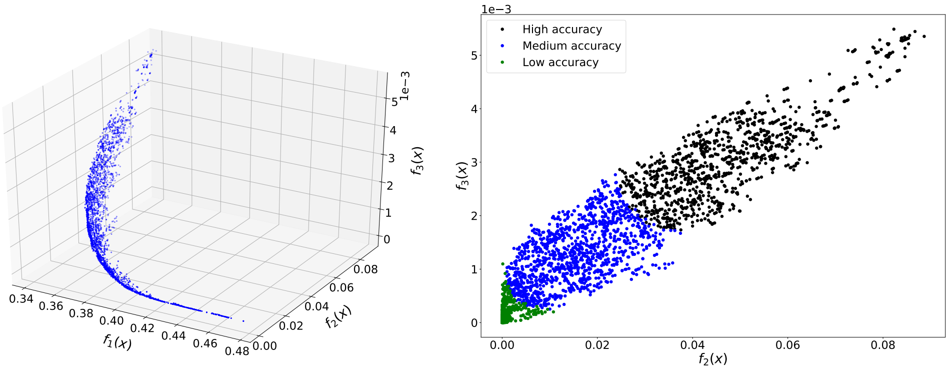
A main advantage of handling fairness in machine learning through multi-objective optimization is the possibility of considering any number of criteria. In this section, we explore two possibilities, multiple sensitive attributes and multiple fairness measures.

6.1. Multiple Sensitive Attributes. Let us see first how we can handle more than one sensitive attribute. One can consider a binary sensitive attribute (e.g. gender) and a multi-valued sensitive attribute (e.g. race), and formulate the following tri-objective problem

$$\min \; (f_1(c,b), f_2^{\mathrm{DI}}(c,b), f_3^{\mathrm{DI}}(c,b)). \tag{13}$$

In our experiments, we use the *Adult Income* dataset and the splitting of training and testing samples of Subsection 4.2. A 3D Pareto front is plotted in Figure 6 (a) resulting from the application of PF-SMG to (13), with gender (f_2^{DI}) and race (f_3^{DI}) as the two sensitive attributes.

Figure 6 (b) depicts all the nondominated points projected onto the f_2–f_3 objective space, where the green, blue, and black points correspond to low, medium, and high prediction accuracy, respectively. It is observed that there is no conflict between f_2^{DI} and f_3^{DI}. Although it could happen for other datasets, eliminating disparate impact with respect to gender does not hinder that with respect to race for this dataset. Intuitively, one could come up with a predictor where the proportions of positive predictions for female and male are equalized and the proportions of positive predictions for different races are equalized within the female and male groups separately, which would lead to zero disparate impact in terms of gender and race simultaneously.



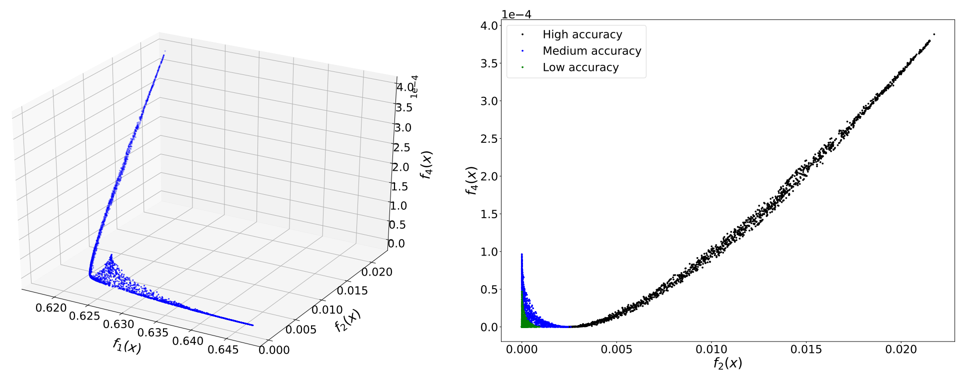
(a) Pareto front. (b) Projection to f_2-f_3 objective space.

FIGURE 6. Trade-off results for problem (13) using Adult Income dataset. Parameters used in PF-SMG: same as in Fig. 4 except for $b_{1,k} = b_{2,k} = b_{3,k} = 80 \times 1.005^k$.

6.2. Multiple Fairness Measures. Now we see how to handle more than one fairness measure. As an example, we consider handling two fairness measures (disparate impact and equal opportunity) in the case of a binary sensitive attribute, and formulate the following tri-objective problem

$$\min \, (f_1(c,b), f_2^{\mathrm{DI}}(c,b), f_4^{\mathrm{FNR}}(c,b)). \tag{14}$$

In our experiments, we use the whole ProPublica COMPAS *two-years-violent* dataset (see Section 5) for both training and testing. Figure 7 (a) shows an approximated 3D Pareto front (resulting from the application of PF-SMG to (14)). By projecting all the obtained nondominated points onto the 2D f_2–f_4 objective space, we have subplot (b), where the three colors indicate the three levels of prediction accuracy. From Figure 7 (b), one can easily find that an unique minimizer (in the green area with lower prediction accuracy) exists for both f_2^{DI} and f_4^{FNR}, and thus conclude that there is indeed no conflict between disparate impact and equal opportunity. In fact, by definition, the CV score (11) generalized to equal opportunity is a component of the CV score (4) measuring disparate impact. Therefore, in the black area where the accuracy is high enough, the values of the two fairness measures are aligned and increasing as the prediction accuracy increases. Interestingly, we have discovered a little Pareto front between f_2 and f_4 when the accuracy is fixed in a certain medium level, marked in blue.



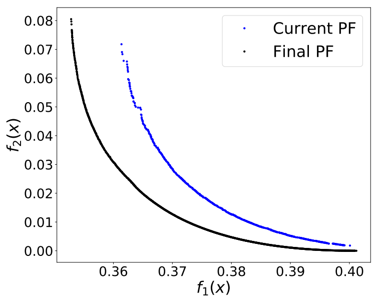
(a) Pareto front. (b) Projection to f_2-f_4 objective space.

FIGURE 7. Trade-off results for problem (14) using COMPAS dataset. Parameters used in PF-SMG: same as in Fig. 5 except for $b_{1,k} = b_{2,k} = b_{3,k} = 80 \times 1.005^k$.

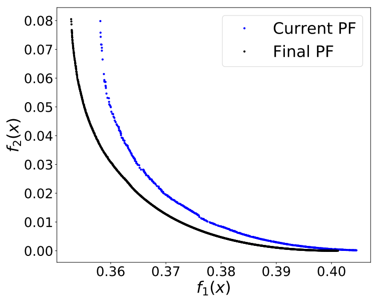
The proposed multi-objective approach works well in handling more than one sensitive attribute or multiple fairness measures. We point out that looking at Pareto fronts for three objectives helps us identifying the existence of conflicts among any subset of two objectives (compared to looking at Pareto fronts obtained just by solving the corresponding bi-objective problems). In the above experiments, by including f_1, we were able to obtain additional helpful information in terms of decision-making reasoning.

7. Streaming Data

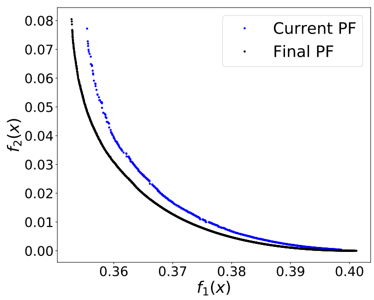
As we claimed in the Abstract and Introduction, another advantage of an SA-based approach like ours is its ability to handle streaming training data. We conducted a preliminary test using the Adult Income dataset and gender as the binary sensitive attribute. To simulate the streaming scenario, the whole dataset is split into batches of 2,000. The initial Pareto front is constructed by applying PF-SMG to one batch of 2,000 samples. Each time a new batch of samples is given, the Pareto front is then updated by selecting a number of nondominated points from the current Pareto front as the starting list for PF-SMG. Figures 8 given below shows how the successive Pareto fronts approach the final one computed for the whole dataset.



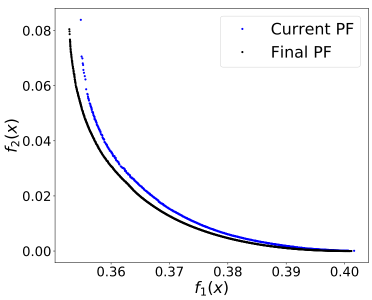
(a) 2,000 samples.



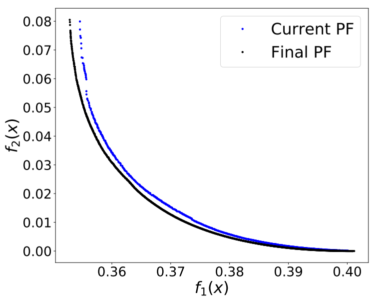
(b) 4,000 samples.



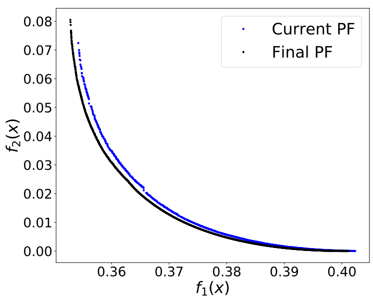
(c) 6,000 samples.



(d) 8,000 samples.



(e) 10,000 samples.



(f) 12,000 samples.

FIGURE 8. Updating Pareto fronts using streaming data.

8. Concluding Remarks

We have proposed a stochastic multi-objective optimization framework to evaluate trade-offs between prediction accuracy and fairness for binary classification. The fairness criterion used was the covariance approximation of disparate impact and equal opportunity, but we could have handled equalized odds in the same vein. A Stochastic Approximation (SA) algorithm like PF-SMG was proved to be computationally efficient to produce well-spread and sufficiently accurate Pareto fronts. We have confirmed the conflicting nature of prediction accuracy and fairness, and presented complete accuracy vs fairness trade-off results. The proposed multi-objective framework can handle both binary and categorical multi-valued sensitive attributes as well as handle more than one sensitive attribute or different fairness measures simultaneously. Using an SA-type approach has allowed us to handle streaming data.

The proposed framework can be generalized to accommodate different types of predictors and loss functions. Hence, one could frame other prediction models, e.g., SVM and neural networks, to multi-objective optimization problems and report accuracy and fairness trade-offs for various machine learning tasks, including multi-class classification and regression. Moreover, our approach allows us to handle nonconvex approximations of disparate impact, equalized odds, or equal opportunity, two potential ones being mutual information [14] and fairness risk measures [26].

Appendix A. The Stochastic Multi-Gradient (SMG) Algorithm

Algorithm 1 Stochastic Multi-Gradient (SMG) Algorithm

Input: an initial point $x_1 \in \mathbb{R}^n$, a step size sequence $\{\alpha_k\}_{k\in\mathbb{N}} > 0$, and maximum iterates T.

for $k = 1, \ldots, T$ **do**

Compute the stochastic gradients $g_i(x_k, w_k)$ for the individual functions, $i = 1, \ldots, m$.

Solve the quadratic subproblem

$$\begin{aligned} \lambda^k \;\in\; & \operatorname{argmin}_{\lambda\in\mathbb{R}^m} \left\|\sum_{i=1}^m \lambda_i g_i(x_k, w_k)\right\|^2 \\ \text{s.t.} \quad & \sum_{i=1}^m \lambda_i = 1, \lambda_i \geq 0, \forall i = 1, ..., m. \end{aligned}$$

Calculate the *stochastic multi-gradient* $g(x_k, w_k) = \sum_{i=1}^m \lambda_i^k g_i(x_k, w_k)$.

Update the iterate $x_{k+1} = x_k - \alpha_k g(x_k, w_k)$.

end for

Appendix B. Illustration of the Pareto-Front Stochastic Multi-Gradient algorithm

In Figure 9, the blue curve represents the true Pareto front. The PF-SMG algorithm first randomly generates a list of starting feasible points (see blue points in (a)). For each point in the current list, a certain number of perturbed points (see green circles in (a)) are added to the list, after which multiple runs of the SMG algorithm are applied to each point in the current list. These newly generated points are marked by red circles in (b). At the end of the current iteration, a new list for the next iteration is obtained by removing all the dominated points. As the algorithm proceeds, the front will move towards the true Pareto front.

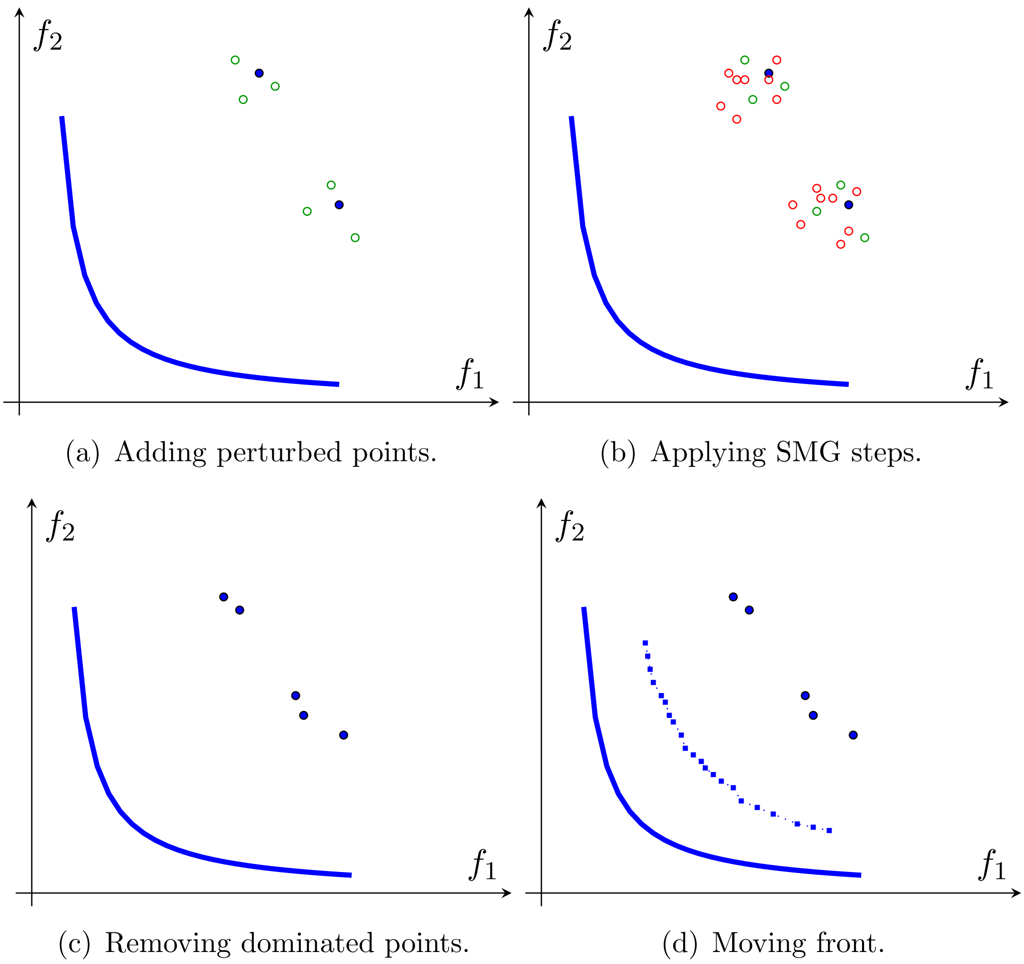


FIGURE 9. Illustration of Pareto-Front stochastic multi-gradient algorithm.

The complexity rates to determine a point in the Pareto front using stochastic multi-gradient are reported in [21]. However, in multiobjective optimization, as far as we know, there are no convergence or complexity results to determine the whole Pareto front (under reasonable assumptions that do not reduce to evaluating the objective functions in a set that is dense in the decision space).

Appendix C. Metrics for Pareto front comparison

Let $\mathcal{A}$ denote the set of algorithms/solvers and $\mathcal{T}$ denote the set of test problems. The Purity metric measures the accuracy of an approximated Pareto front. Let us denote $F(\mathcal{P}_{a,t})$ as an approximated Pareto front of problem t computed by algorithm a. We approximate the "true" Pareto front $F(\mathcal{P}_t)$ for problem t by all the nondominated points in $\cup_{a\in\mathcal{A}}F(\mathcal{P}_{a,t})$. Then, the Purity of a Pareto front computed by algorithm a for problem t is the ratio $r_{a,t} = |F(\mathcal{P}_{a,t}) \cap F(\mathcal{P}_t)|/|F(\mathcal{P}_{a,t})| \in [0,1]$, which calculates the percentage of "true" nondominated solutions among all the nondominated points generated by algorithm a. A higher ratio value corresponds to a more accurate Pareto front.

The Spread metric is designed to measure the extent of the point spread in a computed Pareto front, which requires the computation of extreme points in the objective function space $\mathbb{R}^m$. Among the m objective functions, we select a pair of nondominated points in $\mathcal{P}_t$ with the highest pairwise distance (measured using f_i) as the pair of extreme points. More specifically, for a particular algorithm a, let $(x^i_{\min}, x^i_{\max}) \in \mathcal{P}_{a,t}$ denote the pair of nondominated points where $x^i_{\min} = \operatorname{argmin}_{x\in\mathcal{P}_{a,t}} f_i(x)$ and $x^i_{\max} = \operatorname{argmax}_{x\in\mathcal{P}_{a,t}} f_i(x)$. Then, the pair of extreme points is $(x^k_{\min}, x^k_{\max})$ with $k = \operatorname{argmax}_{i=1,\ldots,m} f_i(x^i_{\max}) - f_i(x^i_{\min})$.

The first Spread formula calculates the maximum size of the holes for a Pareto front. Assume algorithm a generates an approximated Pareto front with M points, indexed by $1,\ldots,M$, to which the extreme points $F(x^k_{\min})$,$F(x^k_{\max})$ indexed by 0 and $M+1$ are added. Denote the maximum size of the holes by Γ. We have

$$\Gamma \;=\; \Gamma_{a,t} \;=\; \max_{i\in\{1,\ldots,m\}} \left(\max_{j\in\{1,\ldots,M\}} \{\delta_{i,j}\} \right),$$

where $\delta_{i,j} = f_{i,j+1} - f_{i,j}$, and we assume each of the objective function values f_i is sorted in an increasing order.

Table 1. Adult Income dataset: Gender

Gender	$\leq 50K$	$> 50K$	Total
Males	$20,988$	$9,539$	$30,527$
Females	$13,026$	$1,669$	$14,695$
Total	$34,014$	$11,208$	$45,222$

The second formula was proposed by [7] for the case $m = 2$ (and further extended to the case $m \geq 2$ in [6]) and indicates how well the points are distributed in a Pareto front. Denote the point spread by Δ. It is computed by the following formula:

$$\Delta \;=\; \Delta_{a,t} \;=\; \max_{i \in \{1,\ldots,m\}} \left(\frac{\delta_{i,0} + \delta_{i,M} + \sum_{j=1}^{M-1} |\delta_{i,j} - \bar{\delta}_i|}{\delta_{i,0} + \delta_{i,M} + (M-1)\bar{\delta}_i} \right),$$

where $\bar{\delta}_i, i = 1, \ldots, m$ is the average of $\delta_{i,j}$ over $j = 1, \ldots, M-1$. Note that the lower Γ and Δ are, the more well distributed the Pareto front is.

Appendix D. Demographic composition of the real datasets

The data pre-processing details for the Adult Income dataset are given below.

(1) First, we combine all instances in *adult.data* and *adult.test* and remove those that values are missing for some attributes.
(2) We consider the list of features: Age, Workclass, Education, Education number, Martial Status, Occupation, Relationship, Race, Sex, Capital gain, Capital loss, Hours per week, and Country. In the same way as the authors [28] did for attribute Country, we reduced its dimension by merging all non-United-Stated countries into one group. Similarly for feature Education, where "Preschool", "1st-4th", "5th-6th", and "7th-8th" are merged into one group, and "9th", "10th", "11th", and "12th" into another.
(3) Last, we did one-hot encoding for all the categorical attributes, and we normalized attributes of continuous value.

In terms of gender, the dataset contains 67.5% males (31.3% high income) and 32.5% females (11.4% high income). Similarly, the demographic compositions in terms of race are 2.88% Asian (28.3%), 0.96% American-Indian (12.2%), 86.03% White (26.2%), 9.35% Black (1.2%), and 0.78%

TABLE 2. Adult Income dataset: Race

Race	$\leq 50K$	$> 50K$	Total
Asian	934	369	1,303
American-Indian	382	53	435
White	28,696	10,207	38,903
Black	3,694	534	4,228
Other	308	45	353
Total	34,014	11,208	45,222

TABLE 3. COMPAS dataset: Race

Race	Reoffend	Not reoffend	Total
White	822	1,281	2,103
Black	1,661	1,514	3,175
Total	2,483	2,795	5,278

Other (12.7%), where the numbers in brackets are the percentages of high-income instances.

References

[1] *Big data: A report on algorithmic systems, opportunity, and civil rights.* Executive Office of the President, 2016.

[2] S. Barocas and A. D. Selbst. Big data's disparate impact. *California Law Review*, page 671, 2016.

[3] T. Calders, F. Kamiran, and M. Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18. IEEE, 2009.

[4] T. Calders and S. Verwer. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21:277–292, 2010.

[5] F. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pages 3992–4001, 2017.

[6] A. L. Custódio, J. A. Madeira, A. I. F. Vaz, and L. N. Vicente. Direct multisearch for multi-objective optimization. *SIAM J. Optim.*, 21:1109–1140, 2011.

[7] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6:182–197, 2002.

[8] E. D. Dolan and J. J. Moré. Benchmarking optimization software with performance profiles. *Mathematical programming*, 91:201–213, 2002.

[9] D. Dua and C. Graff. UCI Machine Learning Repository, 2017.

[10] J. Fliege and B. F. Svaiter. Steepest descent methods for multicriteria optimization. *Math. Methods Oper. Res.*, 51:479–494, 2000.

[11] J. Fliege, A. I. F. Vaz, and L. N. Vicente. Complexity of gradient descent for multiobjective optimization. *Optimization Methods and Software*, 34:949–959, 2019.

[12] Y. V. Haimes. On a bicriterion formulation of the problems of integrated system identification and system optimization. *IEEE Transactions on Systems, Man, and Cybernetics*, 1:296–297, 1971.

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

[14] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 35–50. Springer, 2012.

[15] T. Kamishima, S. Akaho, and J. Sakuma. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 643–650. IEEE, 2011.

[16] J. Kelly. Women now hold more jobs than men in the u.s. workforce. `https://www.forbes.com/sites/jackkelly/2020/01/13/women-now-hold-more-jobs-than-men/#63c4134b8f8a`, 2020.

[17] R. Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD96, pages 202–207. AAAI Press, 1996.

[18] D. Kraft. A software package for sequential quadratic programming. *Forschungsbericht-Deutsche Forschungs- und Versuchsanstalt fur Luft- und Raumfahrt*, 1988.

[19] J. Larson, S. Mattu, L. Kirchner, and J. Angwin. How we analyzed the COMPAS recidivism algorithm. *ProPublica*, 2016.

[20] J. Larson, S. Mattu, L. Kirchner, and J. Angwin. ProPublica COMPAS dataset. `https://github.com/propublica/compas-analysis`, 2016.

[21] S. Liu and L. N. Vicente. The stochastic multi-gradient algorithm for multi-objective optimization and its application to supervised machine learning. Technical report, 2019.

[22] D. Pedreshi, S. Ruggieri, and F. Turini. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 560–568. ACM, 2008.

[23] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems*, pages 5680–5689, 2017.

[24] J. Podesta, P. Pritzker, E. J. Moniz, J. Holdren, and J. Zients. Big data: Seizing opportunities, preserving values, 2014.

[25] M. Quentin, P. Fabrice, and J. A. Désidéri. A stochastic multiple gradient descent algorithm. *European J. Oper. Res.*, 271:808 – 817, 2018.

[26] R. C. Williamson and A. K. Menon. Fairness risk measures. In *International Conference on Machine Learning*, pages 6786–6797, 2019.

[27] B. Woodworth, S. Gunasekar, M. I. Ohannessian, and N. Srebro. Learning non-discriminatory predictors. In *Conference on Learning Theory*, pages 1920–1953, 2017.

[28] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1171–1180. International World Wide Web Conferences Steering Committee, 2017.

[29] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pages 962–970, 2017.

[30] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013.

S. Liu
Department of Industrial and Systems Engineering, Lehigh University, 200 West Packer Avenue, Bethlehem, PA 18015-1582, USA.

E-mail address: `sul217@lehigh.edu`

L. N. Vicente
Department of Industrial and Systems Engineering, Lehigh University, 200 West Packer Avenue, Bethlehem, PA 18015-1582, USA and Centre for Mathematics of the University of Coimbra (CMUC), Portugal.

E-mail address: `lnv@mat.uc.pt`