Loss-Aversively Fair Classification

Junaid Ali, Muhammad Bilal Zafar, Adish Singla, Krishna P. Gummadi
Max Planck Institute for Software Systems (MPI-SWS)
{junaid, mzafar, adishs, gummadi}@mpi-sws.org

ABSTRACT

The use of algorithmic (learning-based) decision making in scenarios that affect human lives has motivated a number of recent studies to investigate such decision making systems for potential unfairness, such as discrimination against subjects based on their sensitive features like gender or race. However, when judging the fairness of a newly designed decision making system, these studies have overlooked an important influence on people's perceptions of fairness, which is how the new algorithm changes the status quo, i.e., decisions of the existing decision making system. Motivated by extensive literature in behavioral economics and behavioral psychology (prospect theory), we propose a notion of fair updates that we refer to as *loss-averse updates*. Loss-averse updates constrain the updates to yield improved (more beneficial) outcomes to subjects compared to the status quo. We propose tractable proxy measures that would allow this notion to be incorporated in the training of a variety of linear and non-linear classifiers. We show how our proxy measures can be combined with existing measures for training nondiscriminatory classifiers. Our evaluation using synthetic and real-world datasets demonstrates that the proposed proxy measures are effective for their desired tasks.

CCS CONCEPTS

• **Social and professional topics**; • **Computing methodologies** → **Machine learning algorithms**;

KEYWORDS

Algorithmic Fairness, Fair Updates, Fairness in Machine Learning, Loss-averse Fairness

ACM Reference Format:
Junaid Ali, Muhammad Bilal Zafar, Adish Singla, Krishna P. Gummadi . 2019. Loss-Aversively Fair Classification. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES '19), January 27–28, 2019, Honolulu, HI, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3306618.3314266

1 INTRODUCTION

The use of algorithmic (data-driven and learning-based) decision making systems in domains ranging from judiciary (recidivism risk estimation) and banking (credit ratings and loan approval risk) to welfare (benefits eligibility) and insurance (accident risks) has raised numerous concerns about their fairness. Consequently, in recent years, a number of notions of algorithmic (un)fairness have been proposed [15, 18, 24] and numerous learning mechanisms have been devised to train algorithmic decision making systems that satisfy these notions [5, 12, 13, 15, 18, 24–26]. These fairness notions have focussed on both the *decision outcomes* as well as the *decision making process*, i.e., the inputs used to make the decisions and the objectives of the learning algorithms.

In this paper, we focus on a crucial aspect of algorithmic decision making systems ignored by existing studies on fair learning namely, fairness of *updates to decision making systems*. In many decision making scenarios such as banking or judiciary or insurance, a newly deployed system replaces an already existing decision making system, be it run by a human decision maker or an older learning model (e.g., learning models without discrimination-awareness) or a learning model trained over outdated training data (e.g., when features of users in a society evolve). Existing literature in behavioral economics and psychology shows that peoples' perceptions of fairness of the new decision making system are influenced by *how the decision outcomes change from the status quo* i.e., how the new outcomes differ from the old outcomes [4, 16, 17, 22]. However, current works on fair learning do not account for the status quo when reasoning about fairness of a decision making system.

In this work, inspired by existing literature in behavioral economics, we formally define a notion of update fairness namely, **loss-aversively fair updates**. Intuitively, our notion of loss-averse updates accounts for the "endowment effect" in human behavior [16, 17], where an individual or a group of users perceives the fairness of the new system based on whether their new outcomes were more or less beneficial than their status quo outcomes from the existing system.

We design intuitive measures for this notion that can be incorporated into a variety of linear and non-linear classifiers as convex constraints and be efficiently learned. A classifier trained using our constraints would account for the existing outcomes from the status quo classifier.

We also show that our new notion of fair update can be easily integrated with existing mechanisms for training non-discriminatory classifiers. For instance, when attempting to equalize rates of beneficial outcomes such as positive class acceptance rate or true positive rate across different groups, adding our loss-averse update constraint ensures that "no group of users is worse-off" than before. Such a constraint may be necessary in practice when training non-discriminatory classifiers as Bazerman et al. [4] point out that same "don't make anyone worse off' principle likely underlines Supreme Courts decision [21] that firing personnel from historically advantaged groups to achieve parity (in order to overcome past discrimination) is prohibited.

In the rest of the paper, we first formally define our notion of fair update in the context of training classifiers. We also propose tractable and efficient mechanisms to train fair classifiers while satisfying this practical consideration. Experiments with synthetic

AIES '19, January 27–28, 2019, Honolulu, HI, USA

ACM ISBN 978-1-4503-6324-2/19/01.
https://doi.org/10.1145/3306618.3314266

and real-world datasets show the effectiveness of our mechanism in enforcing this consideration.

2 RELATED WORK

Fairness in ML. A plethora of recent studies have focused on proposing notions [15, 18, 23, 24] and mechanisms for fairness-aware classification [5, 10, 12, 13, 15, 18, 23–26]. For more discussion into these notions, we point the interested readers to [3, 7, 19, 25]. While classification has received most attention in the area of fairness-aware machine learning, some recent work has also focused on prediction tasks beyond classification, such as regression [6], ranking [8, 20] and clustering [9]. In this paper, we primarily focus on updates to classification tasks, leaving fairness of updates to regression, ranking, and clustering tasks to future studies.

Indvidual-level vs. Group-level Fairness Notions. Fairness in classification has been divided into two broad areas: individual- and group-level fairness [12]. Loss-averse updates can be applied at both individual and group-levels. However, in this work, we only show results at the group-level.

Normative vs. Descriptive Notions of Fairness. Our fairness consideration for updating decision making systems has roots in normative vs. descriptive approaches in behavioral economics [16, 17]. For example, Kahneman et al. [16] show how certain changes to an economic model that are accepted on the normative standards might be deemed unacceptable on the descriptive standards. Our work here is motivated by such observations: while anti-discrimination laws (normatively) prescribe how nondiscriminatory decisions ought to be done, if people (descriptively) preceived the changes in outcomes with the new nondiscriminatory decision system to be too disruptively disadvantageous to them, they would resist adopting the new system. Our notions of update fairness can be thought of as addressing such practical considerations.

3 FORMALIZING NOTION OF LOSS-AVERSE UPDATES

In this section, we formally define a notion of fairness that can be useful when updating algorithmic decision making systems. Specifically, we focus on decision making tasks centered around binary classification.

Preliminaries. In a binary classification task, given a training dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, the goal is to learn a function $\boldsymbol{\theta} : \mathbb{R}^d \rightarrow \{-1, 1\}$ between the feature vectors $\boldsymbol{x} \in \mathbb{R}^d$ and class labels $y \in \{-1, 1\}$. For convex decision boundary-based classifiers like logistic regression and (non)linear SVM, this task boils down to finding a decision boundary $\boldsymbol{\theta}^*$ in the feature space that minimizes a given loss $L(\boldsymbol{\theta})$ over $\mathcal{D}$, *i.e.*, $\boldsymbol{\theta}^* = \text{argmin}_{\boldsymbol{\theta}}\, L(\boldsymbol{\theta})$. The convexity of the loss function ensures that the optimal decision boundary parameters can be found in an efficient manner. Then, for a given (potentially unseen) feature vector $\boldsymbol{x}$, one predicts the class label $\hat{y} = 1$ if $d_{\boldsymbol{\theta}^*}(\boldsymbol{x}) \geq 0$ and $\hat{y} = -1$ otherwise, where $d_{\boldsymbol{\theta}^*}(\boldsymbol{x})$ denotes the signed distance from $\boldsymbol{x}$ to the decision boundary. Without loss of generality, we consider $\hat{y} = 1$ to be the beneficial (desired) label, *e.g.*, being granted the loan or being released on bail.

Setup. We consider scenarios where we need to update an existing, *status quo*, binary classifier, whose decision boundary is denoted by $\boldsymbol{\theta}_{sqo}$. We assume that the boundary of the new classifier, $\boldsymbol{\theta}_{new}$ is learnt from the training dataset $\mathcal{D}$. The outcomes of the updated (new) classifier may differ from the status quo for many reasons such as the status quo classifier being a human or an older (simpler) learning model, or the status quo classifier being trained on outdated training data, or the status quo classifier being trained using models without awareness of potential for discrimination. Our notion of fair update defines the conditions in which the *changes in decision outcomes caused by an update* would be deemed as fair.

Existing Notions: Discrimination in Classification.
Anti-discrimination laws require classification outcomes are also required to be nondiscriminatory with respect to a sensitive feature $z \in \{0, 1\}$, *e.g.*, gender, race. Most of the existing studies differentiate between the following two notions of discrimination: *statistical parity* [12, 13]—also referred to as disparate impact, and *equality of opportunity* [15, 24]—also referred to as disparate mistreatment. Both notions require that certain group-conditional beneficial outcome rates be the same for each group, *i.e.*,:

$$\mathcal{B}_{z=0}(\boldsymbol{\theta}) = \mathcal{B}_{z=1}(\boldsymbol{\theta}), \tag{1}$$

where the definition of the benefit function $\mathcal{B}_z$ depends on the notion of discrimination under consideration.

Under the notion of *statistical parity (SP)* [12, 13], the benefits function is defined as the positive class acceptance rate (AR), *i.e.*, the positive class acceptance rate should be the same for both the groups. More formally,

$$-\ \textbf{\textit{SP:}} \qquad P(\hat{y} = 1|z = 0) = P(\hat{y} = 1|z = 1), \tag{2}$$

Under *equality of opportunity (EOP)* notion [15, 24], the benefit function is defined as the true positive rate, i.e., the true positive rate (TPR) should be the same for both the groups. More formally,

$$-\ \textbf{\textit{EOP:}}\ P(\hat{y} = 1|y = 1, z = 0) = P(\hat{y} = 1|y = 1, z = 1), \tag{3}$$

Note that, current notions of nondiscrimination do not take into account status quo classifier. In the following section we introduce a notion of updating status quo classifier.

New Notion: Loss-Averse Updates. We now formally describe a new consideration of fair updates, introduced in Section 1. We draw inspiration from human behavior and behavioral economics and we consider how people might perceive fairness of an updated classifier in comparison to status quo. Specifically, any disadvantageous effect of an updated classifier would be considered unfair. Prospect theory, proposed by Kahneman and Tversky [17], states that equal amounts of loses result in a bigger loss in utility than the increase in utility by the same amount of gains. In other words people percieve losses much worse than gains, *i.e.*, they are loss-averse. Given the *status quo* classifier $\boldsymbol{\theta}_{sqo}$, a new classifier $\boldsymbol{\theta}_{new}$ constitutes a *loss-averse* update only when the new classifier increases the beneficial outcome rates for all groups. More formally,

$$\mathcal{B}_{z=k}(\boldsymbol{\theta}_{new}) \ \geq \ \mathcal{B}_{z=k}(\boldsymbol{\theta}_{sqo}), \quad \text{for all } k \in \{0, 1\} \tag{4}$$

where $\mathcal{B}_z$ can be any one of the benefit functions proposed in the existing literature on nondiscriminatory classification.

4 UPDATING CLASSIFIERS LOSS-AVERSIVELY

In this section, we devise mechanisms to update status quo classifier, $\boldsymbol{\theta}_{sqo}$ to $\boldsymbol{\theta}_{new}$ that follow the practical considerations of "loss-averse

updates". We specifically focus on training convex decision boundary based classifiers (*e.g.*, logistic regression, linear and non-linear SVMs), *i.e.*, the classifiers that learn the decision boundary parameters by optimizing a convex loss function $L(\boldsymbol{\theta})$.

Existing Mechanisms: Nondiscriminatory Classification. Existing mechanisms to train nondiscriminatory classifiers involve solving an optimization problem maximizing accuracy while equalizing benefits, *i.e.*, enforcing Eq. (1), for different sensitive feature groups. More formally,

$$\begin{aligned} &\text{minimize} \quad L(\boldsymbol{\theta}) \\ &\text{subject to} \quad \mathcal{B}_{z=0}(\boldsymbol{\theta}) = \mathcal{B}_{z=1}(\boldsymbol{\theta}), \end{aligned} \tag{P1}$$

Constraints in Problem (P1), as operationalized in Eqs. (2) and (3) are non-convex. However, prior studies [5, 24, 25] propose tractable convex or convex-concave proxies for enforcing the equality of benefits constraint in Eqs. (2) and (3). Borrowing these proxies from [5, 24, 25], one can replace the equal benefits condition with proxies as follows:

$$-\textbf{\textit{SP:}} \qquad \frac{1}{|\mathcal{D}|}\left|\sum_{(\boldsymbol{x},z)\in\mathcal{D}}(z-\bar{z})d_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\right| \leq c, \tag{5}$$

$$-\textbf{\textit{EOP:}} \qquad \frac{1}{|\mathcal{D}_+|}\left|\sum_{(\boldsymbol{x},z)\in\mathcal{D}_+}(z-\bar{z})d_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\right| \leq c, \tag{6}$$

where $\mathcal{D}_+$ are data points with $y = 1$. Here equality of opportunity limits discrimination in true positive rates of different groups. The covariance threshold $c \in \mathbb{R}^+$ determines the level of discrimination, with $c = 0$ aiming for a perfectly fair classifier.

New Mechanism: Loss-Averse Updates. For updating the status quo classifier, $\boldsymbol{\theta}_{sqo}$, in a nondiscriminatory and loss-aversive manner, one can add the respective conditions to the classifier formulation as a constraint, *i.e.*,

$$\begin{aligned} &\text{minimize} \quad L(\boldsymbol{\theta}) \\ &\text{subject to} \quad \mathcal{B}_{z=0}(\boldsymbol{\theta}) = \mathcal{B}_{z=1}(\boldsymbol{\theta}) \\ &\qquad\qquad \boldsymbol{\mathcal{B}_{z=k}(\theta) \geq \mathcal{B}_{z=k}(\theta_{sqo})}, \quad \text{for all } k \in \{0,1\}. \end{aligned} \tag{P2}$$

The constraints in the above problem are nonconvex functions of the classifier parameters $\boldsymbol{\theta}$, if $\mathcal{B}$ is defined in terms of probabilities as given in Eqs. (2) and (3), for example, this would make it very challenging to solve the resulting problem in an efficient manner.

We used the convex proxies from prior studies [5, 24, 25] for the first constraint as given by Eqs. (5) and (6). We propose the following convex proxies to approximate the new loss-averse constraints in Problem (P2):

Under SP, when the benefit function is AR we suggest:

$$\frac{1}{|\mathcal{D}_{z=k}|}\sum_{\boldsymbol{x}\in\mathcal{D}_{z=k}} d_{\boldsymbol{\theta}}(\boldsymbol{x}) \geq \frac{1}{|\mathcal{D}_{z=k}|}\sum_{\boldsymbol{x}\in\mathcal{D}_{z=k}} d_{\boldsymbol{\theta}_{sqo}}(\boldsymbol{x}) + \gamma, \tag{7}$$
$$\text{for all } k \in \{0,1\}, \gamma \in \mathbb{R}^+.$$

Under EOP, when the benefit function is TPR we suggest:

$$\frac{1}{|\mathcal{D}^+_{z=k}|}\sum_{\boldsymbol{x}\in\mathcal{D}^+_{z=k}} d_{\boldsymbol{\theta}}(\boldsymbol{x}) \geq \frac{1}{|\mathcal{D}^+_{z=k}|}\sum_{\boldsymbol{x}\in\mathcal{D}^+_{z=k}} d_{\boldsymbol{\theta}_{sqo}}(\boldsymbol{x}) + \gamma, \tag{8}$$
$$\text{for all } k \in \{0,1\}, \gamma \in \mathbb{R}^+,$$

where $\mathcal{D}_{z=k}$ are the data points whose sensitive attribute value $z = k$, and $\mathcal{D}^+_{z=k}$ are data points in the dataset with label $y = 1$ and sensitive attribute value $z = k$. Here, γ controls the strength of the constraint. We pick an appropriate γ using a validation set. Note that the right hand side in Eqs. (7) and (8) represents constant terms since $\boldsymbol{\theta}_{sqo}$ is already known.

Both of the proposed proxies are convex with respect to the optimization variables. The convexity of the proxies (7 and 8) means that for any convex function $L(\boldsymbol{\theta})$ the optimization problem stays convex and can be solved in an efficient manner.

Logistic Regression: SP. We can specialize Problem (P2), using logistic regression classifier with L-2 norm regularizer, SP as a notion of discrimination, given by Eq. (5), and loss-averse constraint, given by Eq. (8), as follows:

$$\begin{aligned} &\text{minimize} \quad -\frac{1}{|\mathcal{D}|}\sum_{(\boldsymbol{x},y)\in\mathcal{D}} \log p(y|\boldsymbol{x},\boldsymbol{\theta}) + \lambda||\boldsymbol{\theta}||^2 \\ &\text{subject to} \quad \frac{1}{|\mathcal{D}|}\left|\sum_{(\boldsymbol{x},z)\in\mathcal{D}}(z-\bar{z})d_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\right| < c \\ &\frac{1}{|\mathcal{D}_{z=k}|}\sum_{\boldsymbol{x}\in\mathcal{D}_{z=k}} d_{\boldsymbol{\theta}}(\boldsymbol{x}) \geq \frac{1}{|\mathcal{D}_{z=k}|}\sum_{\boldsymbol{x}\in\mathcal{D}_{z=k}} d_{\boldsymbol{\theta}_{sqo}}(\boldsymbol{x}) + \gamma, \\ &\quad \text{for all } k \in \{0,1\}, \gamma \in \mathbb{R}^+. \end{aligned} \tag{P3}$$

Logistic Regression: EOP. Similarly, considering equality of opportunity as a notion of nondiscrimination we can approximate Problem (P2), by adding Eqs. (6 and 8) as constraints to logistic loss, as follows:

$$\begin{aligned} &\text{minimize} \quad -\frac{1}{|\mathcal{D}|}\sum_{(\boldsymbol{x},y)\in\mathcal{D}} \log p(y|\boldsymbol{x},\boldsymbol{\theta}) + \lambda||\boldsymbol{\theta}||^2 \\ &\text{subject to} \quad \frac{1}{|\mathcal{D}_+|}\left|\sum_{(\boldsymbol{x},z)\in\mathcal{D}_+}(z-\bar{z})d_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\right| < c \\ &\frac{1}{|\mathcal{D}^+_{z=k}|}\sum_{\boldsymbol{x}\in\mathcal{D}^+_{z=k}} d_{\boldsymbol{\theta}}(\boldsymbol{x}) \geq \frac{1}{|\mathcal{D}^+_{z=k}|}\sum_{\boldsymbol{x}\in\mathcal{D}^+_{z=k}} d_{\boldsymbol{\theta}_{sqo}}(\boldsymbol{x}) + \gamma, \\ &\quad \text{for all } k \in \{0,1\}, \gamma \in \mathbb{R}^+. \end{aligned} \tag{P4}$$

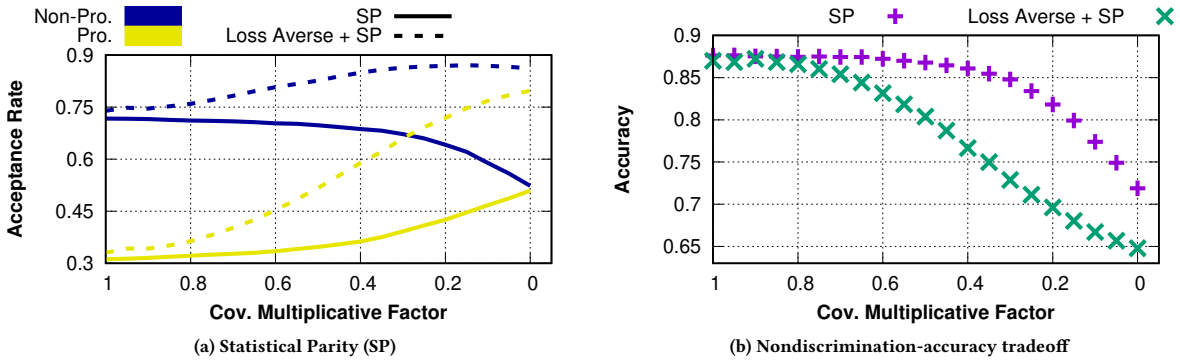
5 EVALUATION ON SYNTHETIC DATASET

In this section we evaluate the effectiveness "Loss-averse" constraint (7), using a synthetic dataset on a binary classification task. We consider a well known notion of nondiscrimination, namely statistical parity. Due to space considerations, we show the results of loss-averse formulation, given by Eq. (8), combined with equality of opportunity, using synthetic data in Appendix A.

5.1 Dataset and Experimental Set up

We used synthetic dataset with binary ground truth class labels $y \in \{+1, -1\}$. Each data point comprises of 2 features besides a binary sensitive feature, *i.e.*, $z \in \{0, 1\}$, where $z = 0$ is the protected group. We do not use the sensitive attribute during training.

Synthetic Dataset. For demonstrating the results of loss-averse updates with statistical parity, given by Eq. (2), as a notion of nondiscrimination, we used the dataset proposed by Zafar et al. [25]. This dataset comprises of 6000 data points, the class labels were drawn



(a) Statistical Parity (SP)

(b) Nondiscrimination-accuracy tradeoff

Figure 1: [Synthetic dataset. Enforcing statistical parity] These figures show a comparison between the solutions of Problem (P1)**, using SP proxies, and Problem** (P3)**. Left panel shows the beneficial outcome rates, *i.e.*, positive class acceptance rates, for a classifier only enforcing SP constraint (solid lines), and a classifier additionally enforcing the "loss-averse" constraint (dotted lines). Right panel shows the nondiscrimination-accuracy tradeoff for both the classifiers. Enforcing "loss-averse" constraint, defined in Eq.** (7)**, leads to significant additional loss in accuracy for the same level of discrimination.**

uniformly at random. Conditioned on the class membership, each data point was sampled from the following distributions:

$$p(\boldsymbol{x}|y=1) = N([2;2][5,1;1,5]),$$
$$p(\boldsymbol{x}|y=-1) = N([-2;-2][10,1;1,3]).$$

Value of the sensitive attribute was sampled from the following Bernoulli probability distributions:

$$p(z=1) = \frac{p(\boldsymbol{x}'|y=1)}{p(\boldsymbol{x}'|y=1) + p(\boldsymbol{x}'|y=-1)},$$

where, $\boldsymbol{x}' = [\cos(\phi), -\sin(\phi); \sin(\phi), \cos(\phi)]\boldsymbol{x}$, *i.e.*, the rotated feature vector, $\boldsymbol{x}$. On average there were 3280 points in the protected group and 2720 were in non-protected group.

Experimental Setup. The dataset is split into 70%-30%, train-test folds. Additionally, hyperparameters are validated using a 30% hold out set from the training data. All the results have been averaged over 5 shuffles of the data initialized by different random seed. In order to pick the penalization parameter, λ in Problem (P3), multiplied with the regularizer, we trained the unconstrained classifier for $\lambda \in [1e-5, 1e-2]$. Then, we picked a value which yielded the highest accuracy on the validation set, for a particular shuffle of the data . We used this value of the parameter for *all* the experiments on that shuffle of the data. We use CVXPY [11] library to solve all the optimization problems.

5.2 Loss-aversively Fair Updates

In this section we experiment with Problems (P1 and P3). First we consider statistical parity, where beneficial outcome rates are defined as positive class acceptance rate, as a notion of discrimination, *i.e.*, solving Problems (P1) using SP proxies. Then, we show results combining SP and loss-averse constraints and we update $\boldsymbol{\theta}_{sqo}$ with loss-averse nondiscriminatory classifiers.

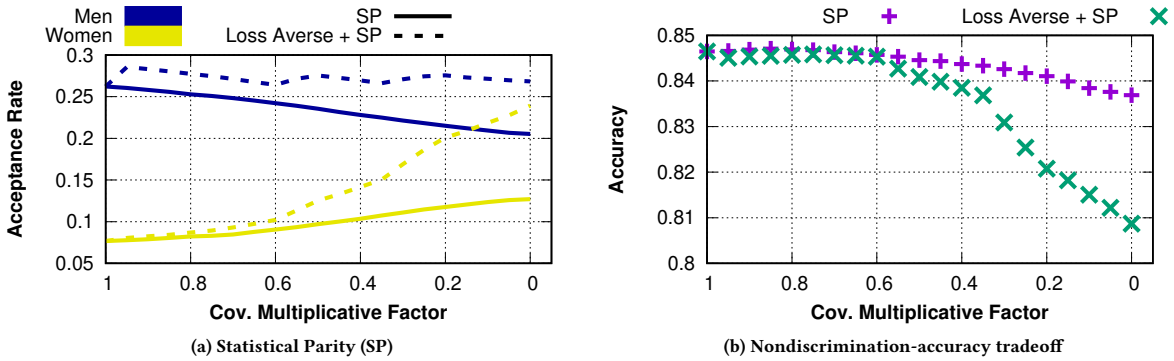
Training Loss-aversively Fair Classifier. We initialize θ_{sqo} with the solution of unconstrained problem. Then, given a value of covariance threshold c, as used in Eqs.(5 and 6), and a range of γ, as used in Eqs.(7 and 8), we solve Problem (P3). We, then, pick the gamma values whose solutions yield a higher benefits compared to $\boldsymbol{\theta}_{sqo}$, for all the groups, on the validation set. In case there are multiple such values, we pick the one whose solution yields maximum accuracy. We then report the results on the test set.

SP. Accuracy of an unconstrained classifier, on *Synthetic* dataset, is 88%, and the acceptance rates for the protected and non-protected groups are 31% and 72%, respectively. There is a clear disparity in acceptance rates of both the groups. In order to remove this disparity we solve Problem (P1), replacing the first constraint with SP proxy, given by Eq. (5). For a covariance threshold $c = 0$, this leads to a classifier with an acceptance rate of 51% and 52%, for protected and non-protected groups respectively, and an accuracy of 72%.

The results for this formulation, Problem (P1) specialized with SP, are shown in Figure (1). The x-axis is covariance multiplicative factor $m : c = m \times c^*$, where c^* is the covariance values of the unconstrained classifier and c is covariance threshold as given in Eq.(5). *Solid lines* in Figure (1a) represent the statistics of the classifiers resulting from the solutions of this formulation. Figure (1b) shows the accuracies of classifiers resulting from solving this formulation in *purple* colored points.

Note that: i) Figure (1b) demonstrates that as the covariance is decreased the accuracy of the resulting, less discriminatory, classifiers also decreases. ii) Figure (1a) shows that as the covariance decreases, the discrimination also reduces. iii) However it should be noted that discrimination is decreased by reducing the acceptance rate of the non-protected group.

Loss-Aversiveness + SP. In order to train a classifier enforcing loss-averse update of $\boldsymbol{\theta}_{sqo}$, Eq. (4), combined with statistical parity, Eq (2), on the *Synthetic* dataset, we solve Problem (P3). Loss-averse updates yield a classifier with an accuracy of 65% and acceptance rates of 80% and 86% for protected and non-protected groups, respectively, for the covariance value $c = 0$.



(a) Statistical Parity (SP)

(b) Nondiscrimination-accuracy tradeoff

Figure 2: [Adult dataset. Enforcing statistical parity] Left panel shows the beneficial outcome rates, *i.e.*, positive class acceptance rates, for a classifier only enforcing SP constraint, *i.e.*, solution of Problem (P1) using SP proxies (solid lines), and a classifier additionally enforcing the "loss-averse" constraint, *i.e.*, solution of Problem (P3) (dotted lines). Right panel shows the nondiscrimination-accuracy tradeoff for both the classifiers. Enforcing "loss-averse" constraint, defined in Eq. (7), leads to a significant additional loss in accuracy for the same level of discrimination.

The results are shown in Figure (1a) in *dotted lines* and in *green* colored points in Figure (3b). i) The figures demonstrate that loss-aversively fair updates yield a less discriminatory classifier while increasing the benefits for both the groups, ii) however this comes at a higher cost of accuracy.

Summary. In this section we demonstrated the effectiveness of our proposed formulation on synthetic datasets. We illustrated the effectiveness of loss-aversively making the status quo classifiers nondiscriminatory, albeit at a higher cost of accuracy.

6 EVALUATION ON REAL-WORLD DATASET

In this section, we evaluate the effectiveness of our proposed schemes in updating the status quo classifier, $\boldsymbol{\theta}_{sqo}$, compliant with the "loss-aversively fair updates" consideration, on real-world dataset using statistical parity as a notion of nondiscrimination. We also consider another widely used notion of discrimination, *i.e.*, equality of opportunity, and show loss-averse constraints combined with EOP on a real-world dataset in Appendix B, due to space limitations.

6.1 Dataset and Experimental Setup

In this section we explain the real-world dataset used to evaluate our proposed considerations.

Adult Dataset. We show result for loss-aversively fair update mechanism, introduced in section 4, using *Adult dataset* [2]. Specifically, we illustrate the effectiveness of Problem (P3) to train loss-aversively fair classifiers, using Adult dataset. For experiments in this section, we consider statistical parity as a notion of nondiscrimination.

The *Adult Dataset* consists of 45, 222 subjects and 14 features like gender, race, educational level, *etc.* The classification task is to predict whether a person earns more than 50K USD per annum (positive class) or not (negative class). We consider gender to be a sensitive feature for this dataset.

Experimental Setup. For the experiments conducted on the *Adult dataset* we use the same data split as used for *Synthetic dataset*. We also randomize the data, as well as validate the hyperparameters in a similar manner.

6.2 Loss-Aversively Fair Updates

In this section we compare the results of Problem (P1), using SP proxies, and Loss-aversively fair updates given by Problem (P3) using *Adult dataset*.

SP. On the Adult dataset, logistic regression classifier leads to an accuracy of 84.6%. However, the classifier leads to the beneficial outcome rates of 8% and 26% for women and men respectively, showing a clear disparity in the beneficial outcome rates for the two groups. Next, using the method of Zafar et al. [25], we train a nondiscriminatory classifier while reducing the value of the covariance threshold c, (Eq. (5)), towards 0. The results are shown in *solid lines* in Figure (2a) and in *purple* colored points in Figure (2b). The least discriminatory classifier in this case achieves the beneficial outcome rates of 13% and 20% for women and men respectively, with an accuracy of 83.7%. We notice that the discrimination is reduced by lowering the beneficial outcome rates for men, which leads to a violation of "loss-averse" consideration.

Loss-Aversiveness + SP. We next train classifier with the loss-averse constraints (Eq. (7)) combined with SP, *i.e.*, solve Problem. (P3). The least discriminatory classifier in this case achieves the beneficial outcome rates of 24% and 27% for women and men, respectively, while achieving an accuracy of 80.8%. However, the reduction in discrimination is achieved by only increasing the beneficial outcome rate for both groups. Results are shown in Figures (2a and 2b), in *dotted lines* and *green* colored points, respectively.

The figure shows the beneficial outcome rates for (i) a classifier with statistical parity constraint and (ii) a classifier with loss-averse and statistical parity constraints. The figure shows that at successively decreasing values of the covariance threshold c, while classifier (i) achieves lower discrimination by increasing benefits for one group and decreasing them for the other, classifier (ii) does so by *only increasing benefits for both the groups*. Figure (2b) shows the

nondiscrimination-accuracy tradeoff achieved by both the classifiers. The figure demonstrates that, as expected, classifier (ii) incurs a much higher cost in terms of accuracy for the same level of discrimination due to the additional loss-averse constraint.

Summary. Our proposed methodology, in Section 4, successfully enforces the loss averse constraint while updating the status quo classifier, $\boldsymbol{\theta}_{sqo}$, to a nondiscriminatory classifier. However, enforcing these constraints could be at a significant additional cost in terms of accuracy.

7 CONCLUDING DISCUSSION

A number of recent works have explored various aspects of fairness related to algorithmic decision making. In this paper, we focus on an aspect of decision making that crucially affects people's fairness perceptions, yet has been overlooked: it is the *fairness of updating decision making*, i.e., how the decision outcomes change when updating a decision making system.

Based on observations in behavioral economics and psychology, we note that any "disadvantageous" changes in outcomes to individual subjects or groups of subjects would be perceived as unfair. Accordingly, we propose a complementary notion of update fairness that we call *loss-averse updates*. Loss-averse updates try to constrain updates to only yield more advantageous (more beneficial) outcomes compared to status quo.

In this work, we formalize this notion in the context of classification tasks. We proposed measures that would allow these notions to be incorporated in the training of any convex decision-boundary based classifiers (like logistic regression or linear/non-linear SVM) as convex constraints. We also show how this notion can be combined with prior notions and measures of non-discrimination in classification. Our evaluation using synthetic and real-world datasets demonstrates the benefits of loss-averse updates in practice.

Our work here also opens up a number of new and interesting research directions. The motivation behind our notions of fair updates generalize to any algorithmic decision making scenario that affects people's lives including search and recommender algorithms such as Google's search, Facebook's NewsFeed, Amazon's product recommendations or market-matching algorithms like Uber's rider-driver matching algorithms. Exploring how our notion loss-averse updates can be applied to these more complex algorithmic decision making scenarios (beyond binary classification) remains an open challenge.

REFERENCES

[1] 2017. http://www.nyc.gov/html/nypd/html/analysis_and_planning/stop_question_and_frisk_report.shtml.
[2] Adult. 1996. http://tinyurl.com/UCI-Adult.
[3] Solon Barocas and Andrew D. Selbst. 2016. Big Data's Disparate Impact. *California Law Review* (2016).
[4] Max H Bazerman, Sally Blount White, and George F Loewenstein. 1995. Perceptions of Fairness in Interpersonal and Individual Choice Situations. *Current Directions in Psychological Science* (1995).
[5] Yahav Bechavod and Katrina Ligett. 2017. Learning Fair Classifiers: A Regularization Approach. FATML.
[6] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2017. A Convex Framework for Fair Regression. *arXiv preprint arXiv:1706.02409* (2017).
[7] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2017. Fairness in criminal justice risk assessments: the state of the art. *arXiv preprint arXiv:1703.09207* (2017).
[8] Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. 2018. Equity of Attention: Amortizing Individual Fairness in Rankings. In *SIGIR*.
[9] Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii. 2017. Fair Clustering Through Fairlets. In *NIPS*.
[10] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. In *KDD*.
[11] Steven Diamond and Stephen Boyd. 2016. CVXPY: A Python-Embedded Modeling Language for Convex Optimization. *Journal of Machine Learning Research* 17, 83 (2016), 1–5.
[12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, and Omer Reingold. 2012. Fairness Through Awareness. In *ITCSC*.
[13] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *KDD*.
[14] Sharad Goel, Justin M. Rao, and Ravi Shroff. 2015. Precinct or Prejudice? Understanding Racial Disparities in New York City's Stop-and-Frisk Policy. *Annals of Applied Statistics* (2015).
[15] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. In *NIPS*.
[16] Daniel Kahneman, Jack L Knetsch, and Richard Thaler. 1986. Fairness as a Constraint on Profit Seeking: Entitlements in the Market. *The American economic review* (1986).
[17] Daniel Kahneman and Amos Tversky. 1979. Prospect Theory: An Analysis of Decisions Under Risk. *Econometrica* (1979).
[18] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware Data Mining. In *KDD*.
[19] Andrea Romei and Salvatore Ruggieri. 2014. A Multidisciplinary Survey on Discrimination Analysis. *KER* (2014).
[20] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of Exposure in Rankings. *arXiv preprint arXiv:1802.07281* (2018).
[21] Supreme Court of the United States. 1989. Martin vs. Wilks.
[22] Joel E Urbany, Thomas J Madden, and Peter R Dickson. 1989. All's not fair in pricing: an initial look at the dual entitlement principle. *Marketing Letters* (1989).
[23] Muhammad Bilal Zafar, Isabel Valera, Manuel Rodriguez, Krishna P. Gummadi, and Adrian Weller. 2017. From Parity to Preference-based Notions of Fairness in Classification. In *NIPS*.
[24] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *WWW*.
[25] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness Constraints: Mechanisms for Fair Classification. In *AISTATS*.
[26] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. 2013. Learning Fair Representations. In *ICML*.

ACKNOWLEDGEMENTS

This research was supported in part by a European Research Council (ERC) Advanced Grant for the project "Foundations for Fair Social Computing", funded under the European Union's Horizon 2020 Framework Programme (grant agreement no. 789373)

A EVALUATION ON SYNTHETIC DATASET: EOP

In this section we will present the "loss-averse" fairness results combined with equality of opportunity, using synthetic dataset. We show the results of the optimization Problem (P4).

A.1 Dataset and Experimental Setup

In this section we explain the synthetic dataset used for demonstrating the loss-averse consideration and the experimental setup used to solve the optimization Problem (P4).

Synthetic Dataset. Each data point comprises of 2 features apart from the sensitive attribute. Each data point also has a binary ground truth label. For equality of opportunity, as given by Eq. (3), we are considering true positive rates as a notion of benefit. To demonstrate the results of fair updates combined with EOP, we use a synthetic dataset proposed by Zafar et al. [24], except that we flip the ground truth labels in order to have a disparity in the false negative rates instead of the false positive rates. We generated 16000 data points with the probability distributions of the features given as follows:

$$p(\boldsymbol{x}|z=0, y=1) = N([2;2][3,1;1,3])$$
$$p(\boldsymbol{x}|z=1, y=1) = N([2;2][3,1;1,3])$$
$$p(\boldsymbol{x}|z=0, y=-1) = N([1;1][3,3;1,3])$$
$$p(\boldsymbol{x}|z=1, y=-1) = N([-2;-2][3,1;1,3])$$

Both, class labels, y, and value of the sensitive attribute, z, were sampled uniformly at random.

Experimental Setup. We use the same data split and method of validating the hyperparameters as explained in section 5.

A.2 Loss-Aversively Fair Updates

In this section we show the results of Problem (P1), using EOP as a notion of nondiscrimination. We also show results for the loss-averse formulation combined with EOP, given by Problem (P4).

EOP. An unconstrained classifier trained on *Synthetic* dataset yields an accuracy of 86% and true positive rates (TPRs) of 94% and 77% for non-protected and protected groups, respectively. To equalize the TPRs we solve Problem (P1) using proxies for EOP given in Eq. (6).

These results are show in Figure (3a) in *solid lines* and Figure (3b) in *purple* colored points. i) In order to reduce discrimination, this formulation yields a classifier which lowers the TPR of the non-protected class to 72% and raises the TPR of the protected group to 79%, for covariance threshold $c = 0$, while achieving an accuracy of 74%. ii) Figure (3a) shows the limitation of equality of opportunity proxy proposed by Zafar et al. [24], as it achieves a lower discrimination for higher value of the covariance.

Loss-Aversiveness + EOP. To avoid lowering the benefits for any group while reducing discrimination, we solve the Problem (P4). We encountered some issues in convergence for some values of covariance factor, specifically smaller ones. Out of 7 random seeds that we tried we find the results for *all* covariance factors for only 5 seeds, we report the average of these results. One reason for the lack of convergence could be that a very high base TPR might make it difficult to find a nondiscriminatory classifier.For covariance threshold $c = 0$, this formulation leads to a classifier whose true positive rates are 95% and 99% for non-protected and protected groups, respectively, with an accuracy of 64%.

We show these results in Figure (3a) in *dotted lines* and Figure (3b) in *green* colored crosses. i) These figures illustrate the effectiveness of the loss-averse formulation, as the resulting classifiers achieve nondiscrimination by increasing TPR for both groups, ii) however this results in a significant drop in the accuracy.

B EVALUATION ON REAL-WORLD DATASET: EOP

In this section we will present the "loss-averse" fairness results combined with equality of opportunity, using a real-world dataset.

B.1 Dataset and Experimental Setup

In this section we explain the dataset and the experimental setup. We show result of Problem (P1), with EOP as a notion of nondiscrimination, as well as Problem (P4), which combines EOP and loss-averse constraints.

SQF Dataset. For experiments in this section we consider *NYPD SQF dataset* [1]. The NYPD *SQF dataset* consists of pedestrians who were stopped in the year 2012 on the suspicion of having a weapon. The task is a binary prediction task which indicates whether (negative class) or not (positive class) a weapon was discovered. For our analysis, we consider the race to be the sensitive feature with values African-American and white. After balancing the classes and considering same features as Goel et al. [14], with the exception that we exclude the highly sparse features 'precinct' and 'timestamp of the stop', we obtain 5,832 subjects and 19 features.

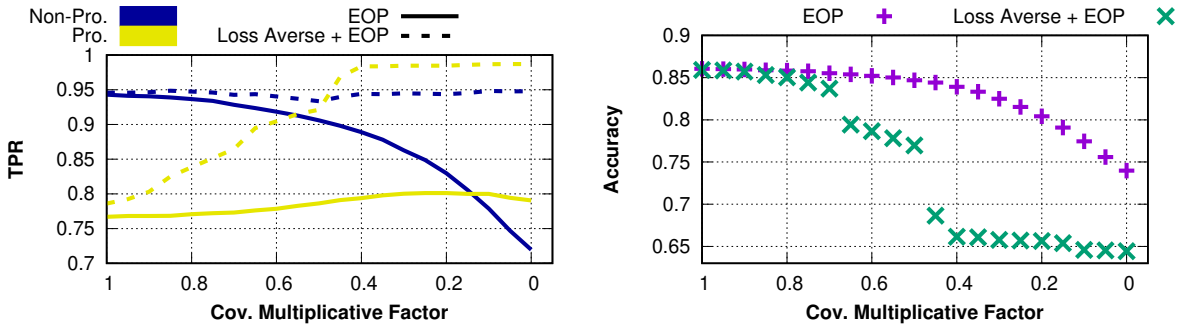
Experimental Setup. We used similar experimental setup as explained in section 5.

B.2 Loss-Aversively Fair Updates

In this section we show the results of Problem (P4), which enforces EOP and loss-averse constraints and compare them with the results of Problem (P1), which only enforces EOP using the proxy given by Eq. (6), on *NYPD SQF dataset.*

EOP. With equality of opportunity constraint, where beneficial outcome rates are defined in terms of true positive rate, we experiment with NYPD SQF dataset. Unconstrained logistic regression on SQF yields an accuracy of 74.4%, while the beneficial outcome rates are 69% and 82% for Whites and African-Americans, respectively. Least discriminatory classifier, trained with $c = 0$, given in constraint Eq. (6), yields benefits of 72% and 76% for Whites and African-Americans, respectively, while achieving an accuracy of 71.4%. Similar to the previous cases, this classifier also achieves lower discriminations by raising the benefits for one group while increasing them for the other group.

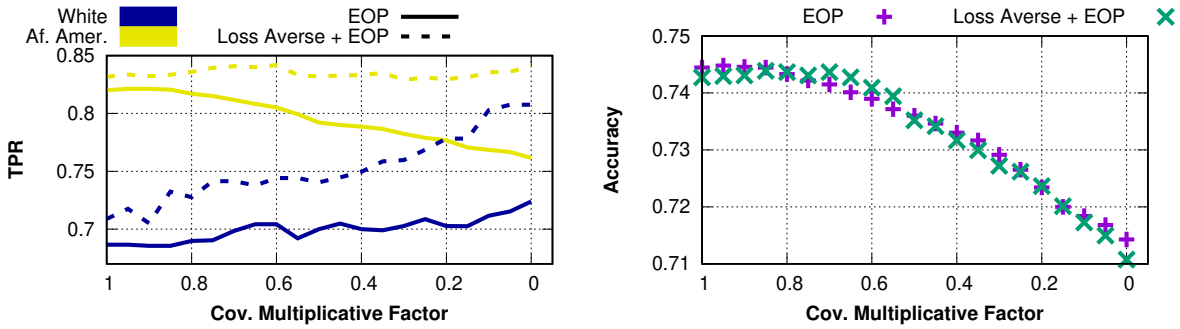
Loss-Aversiveness + EOP. Next, we combine the nondiscrimination constraint with the loss-averse constraint, given by Problem (P4), in order to update $\boldsymbol{\theta}_{sqo}$. A least discriminatory loss-averse classifier trained on NYPD SQF dataset yields an accuracy of 71% and benefits of 84% and 81% for African-Americans and White, respectively. Figure (4a) shows the beneficial outcome rates for (i) a classifier with only nondiscrimination constraints and (ii) a loss-averse classifier with nondiscrimination constraints. Again,



(a) Equality of Opportunity (EOP)

(b) Nondiscrimination-accuracy tradeoff

Figure 3: [Synthetic dataset. Enforcing equality of opportunity] Figure on the left shows the beneficial outcome rates, *i.e.*, true positive rates, for a classifier only enforcing EOP constraint (solid lines) and a classifier additionally enforcing the "loss-averse" constraint, given in Eq. (8), is shown in dotted lines. Figure on the right shows nondiscrimination-accuracy tradeoff for both the classifiers.



(a) Equality of Opportunity (EOP)

(b) Nondiscrimination-accuracy tradeoff

Figure 4: [SQF dataset. Enforcing equality of opportunity] These figures show similar results as Figure (3) using SQF dataset.

we notice that classifier (ii) removes discrimination by *only increasing the beneficial outcome rates* whereas classifier (i) does so by increasing benefits for one group and decreasing them for the other.

Finally, the comparison of nondiscrimination-accuracy tradeoff in Figure (4b) shows no significant difference between both the classifiers.