Poisoning Attacks on Fair Machine Learning

Minh-Hao Van[1][0000−0001−7342−6801], Wei Du[1][0000−0002−3371−8305], Xintao Wu[1][0000−0002−2823−3063], and Aidong Lu[2][0000−0002−7684−4512]

[1] University of Arkansas, Fayetteville, AR 72701, USA
{haovan,wd005,xintaowu}@uark.edu
[2] University of North Carolina at Charlotte, Charlotte, NC 28223, USA
Aidong.Lu@uncc.edu

Abstract. Both fair machine learning and adversarial learning have been extensively studied. However, attacking fair machine learning models has received less attention. In this paper, we present a framework that seeks to effectively generate poisoning samples to attack both model accuracy and algorithmic fairness. Our attacking framework can target fair machine learning models trained with a variety of group based fairness notions such as demographic parity and equalized odds. We develop three online attacks, adversarial sampling , adversarial labeling, and adversarial feature modification. All three attacks effectively and efficiently produce poisoning samples via sampling, labeling, or modifying a fraction of training data in order to reduce the test accuracy. Our framework enables attackers to flexibly adjust the attack's focus on prediction accuracy or fairness and accurately quantify the impact of each candidate point to both accuracy loss and fairness violation, thus producing effective poisoning samples. Experiments on two real datasets demonstrate the effectiveness and efficiency of our framework.

Keywords: Poisoning attacks · Algorithmic fairness · Adversarial machine learning.

1 Introduction

Both fair machine learning and adversarial machine learning have received increasing attention in past years. Fair machine learning (FML) aims to learn a function for a target variable using input features, while ensuring the predicted value be fair with respect to some sensitive attributes based on given fairness criterion. FML models can be categorized into pre-processing, in-processing, and post-processing (see a survey [13]). Adversarial machine learning focuses on vulnerabilities in machine learning models and has been extensively studied from perspectives of attack settings and defense strategies (see surveys [21,4]).

There have been a few works on attacking FML models very recently. Solans et al. [18] developed a gradient-based poisoning attack to increase demographic disparities among different groups. Mehrabi et al. [14] also focused on demographic disparity and presented anchoring attack and influence attack. Chang et al. [5] focused on attacking FML models with equalized odds. To tackle the

challenge of intractable constrained optimization, they developed approximate algorithms for generating poisoning samples. However, how to effectively generate poisoning samples to attack algorithmic fairness still remains challenging due to its difficulty of quantifying impact of each poisoning sample to accuracy loss or fairness violation in the trained FML model.

In this paper, we present a poisoning sample based framework (PFML) for attacking fair machine learning models. The framework enables attackers to adjust their attack's focus on either decreasing prediction accuracy or increasing fairness violation in the trained FML model. Our framework supports a variety of group based fairness notions such as demographic parity and equalized odds. We present three training-time attacks, adversarial sampling, adversarial labeling, and adversarial feature modification. All of these attacks leave the test data unchanged and instead perturb the training data to affect the learned FML model. In adversarial sampling, the attacker is restricted to select a subset of samples from a candidate attack dataset that has the same underlying distribution of the clean data. Adversarial labeling and adversarial feature modification can further flip the labels or modify features of selected samples. All three developed attacking methods are online attacks, which are more efficient than those offline poisoning attacks. Our framework enables attackers to flexibly adjust the attack's focus on prediction accuracy or fairness and accurately quantify the impact of each candidate point to both accuracy loss and fairness violation, thus producing effective poisoning samples. Experiments on two real datasets demonstrate the effectiveness and efficiency of our framework.

2 Background

2.1 Fair Machine Learning

Consider a binary classification task $f_\theta : \mathcal{X} \to \mathcal{Y}$ from an input $x \in \mathcal{X}$ to an output $y \in \mathcal{Y}$. Let $l : \Theta \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ be a loss function, $\mathcal{D}$ be the training set and each $(x, y) \in \mathcal{D}$ be a data point. The classification model minimizes, $\mathcal{L}(\theta, \mathcal{D}) = \sum_{(x,y)\in\mathcal{D}} l(\theta; x, y)$, the cumulative loss of the model over the training data set $\mathcal{D}$, to obtain the optimal parameters. Without loss of generality, we assume $\mathcal{X}$ contains one binary sensitive feature $S \in \{0, 1\}$. FML aims to train a model such that its predictions are fair with respect to S based on a given fairness notion, e.g., disparate impact, equal opportunity and equalized odds.

Definition 1. *A binary classifier f_θ is δ-fair under a fairness notion Δ if $\Delta(\theta, \mathcal{D}) \leq \delta$, where $\Delta(\theta, \mathcal{D})$ is referred as the empirical fairness gap of the model and δ is a user-specified threshold. The model satisfies exact fairness when $\delta = 0$.*

Definition 2. *We denote demographic parity and equalized odds as Δ_{DP} and Δ_{EO}, respectively. They are defined as:*

$$\Delta_{DP}(\theta, \mathcal{D}) := |\Pr(f_\theta(X) = 1|S = 1) - \Pr(f_\theta(X) = 1|S = 0)| \tag{1}$$

$$\Delta_{EO}(\theta, \mathcal{D}) := \max_{y\in\{0,1\}} |Pr[f_\theta(X) \neq y|S = 0, Y = y] - Pr[f_\theta(X) \neq y|S = 1, Y = y]| \tag{2}$$

Algorithm 1 Online Learning for Generating Poisoning Data

Require: $\mathcal{D}_c$, $n = |\mathcal{D}_c|$, feasible poisoning set $\mathcal{F}(\mathcal{D}_k)$, number of poisoning data ϵn, learning rate η.
Ensure: Poisoning dataset $\mathcal{D}_p$.
1: Initialize $\theta^0 \in \Theta$, $\mathcal{D}_p \leftarrow Null$
2: **for** $t = 1 : \epsilon n$
3: $(x^t, y^t) \leftarrow argmax_{(x,y)\in\mathcal{F}(\mathcal{D}_k)}[l(\theta^{t-1}; x, y)$
4: $\mathcal{D}_p \leftarrow \mathcal{D}_p \cup \{(x^t, y^t)\}$, $\mathcal{F}(\mathcal{D}_k) \leftarrow \mathcal{F}(\mathcal{D}_k) - \{(x, y)\}$
5: $\theta^t \leftarrow \theta^{t-1} - \eta \frac{\nabla \mathcal{L}(\theta^{t-1}; \mathcal{D}_c \cup \mathcal{D}_p)}{n+t}$
6: **end for**

Demographic parity requires that the predicted labels are independent of the protected attribute. Equalized odds [9] requires the protected feature S and predicted outcome $\hat{Y}$ are conditionally independent given the true label Y. Equalized opportunity is a weaker notion of equalized odds and requires non-discrimination only within the advantaged outcome group. Our framework naturally covers equalized opportunity. The FML model achieves δ-fairness empirically by minimizing the model's empirical accuracy loss under the fairness constraint:

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \frac{1}{|\mathcal{D}|}\mathcal{L}(\theta; \mathcal{D}) \text{ s.t. } C(\theta, \mathcal{D}) = \Delta(\theta, \mathcal{D}) - \delta \leq 0 \tag{3}$$

2.2 Data Poisoning Attack

Data poisoning attacks [2,3,15] seek to increase the misclassification rate for test data by perturbing the training data to affect the learned model. The perturbation can generally include inserting, modifying or deleting points from the training data so that the trained classification model can change its decision boundaries and thus yields an adversarial output. The modification can be done by either directly modifying the labels of the training data or manipulating the input features depending on the adversary's capabilities. In this study, we assume that an attacker can access to the training data during the data preparation process and have the knowledge of the structure and fairness constraint of the classification model. We focus on three data poisoning attacks, adversarial sampling, adversarial labeling, and adversarial feature modification, against group-based FML models. In all three attacks, the adversary can select the feature vector of the poisoning data from an attack dataset $\mathcal{D}_k$, which is sampled from the same underlying distribution of the clean dataset $\mathcal{D}_c$, and can control sampling, labeling, or modifying for a fraction of training data in order to reduce the test accuracy.

Algorithm 1 shows the general online gradient descent algorithm for generating poisoning samples. The input parameter n denotes the size of the clean set $\mathcal{D}_c$, ϵ is the fraction of the size of generated poisoning data over the clean data in the training data set, $\mathcal{F}(\mathcal{D}_k)$ is feasible poisoning set. Specifically, $\mathcal{F}(\mathcal{D}_k)$ is

the same as $\mathcal{D}_k$ for adversarial sampling. A fraction of data points $(x, y) \in \mathcal{D}_k$ are changed to $(x, 1-y)$ for adversarial labeling, and to $(\tilde{x}, y)$ for adversarial feature modification where $\tilde{x}$ is a modified version of feature vector x. In line 1, it first initializes the model with $\theta^0 \in \Theta$. Using the feasible set of poisoning points, the algorithm generates ϵn poisoning data points iteratively. In line 3, it selects a data point with the highest impact on the loss function with respect to θ^{t-1}. In line 4, it adds the generated data point to $\mathcal{D}_p$. In line 5, the model parameters θ are updated to minimize the loss function based on the selected data point (x^t, y^t).

3 Data Poisoning Attack on FML

3.1 Problem Formulation

The attacker's goal is to find a poisoning dataset that maximizes the linear combination of the accuracy loss and the model's violation from the fairness constraint. The fairness constraint is defined as $C(\theta, \mathcal{D}) = \Delta(\theta, \mathcal{D}) - \delta \leq 0$. We formulate the data poisoning attack on algorithmic fairness as a bi-level optimization problem:

$$\begin{aligned} &\max_{\mathcal{D}_p} \mathbb{E}_{(x,y)}[\alpha \cdot l(\hat{\theta}; x, y) + (1-\alpha) \cdot \gamma \cdot l_f(\hat{\theta}; x, y)] \\ &\text{where } \hat{\theta} = \arg\min_{\theta \in \Theta} \frac{\mathcal{L}(\theta; \mathcal{D}_c \cup \mathcal{D}_p)}{|\mathcal{D}_c \cup \mathcal{D}_p|} \\ &\qquad \text{s.t. } C(\theta, \mathcal{D}_c \cup \mathcal{D}_p) = \Delta(\theta, \mathcal{D}_c \cup \mathcal{D}_p) - \delta \leq 0 \end{aligned} \tag{4}$$

where $\alpha \in [0, 1]$ is a hyperparameter that controls the balance of the attack's focus on accuracy and fairness, $l(\hat{\theta}; x, y)$ is the prediction accuracy loss of the sample (x, y), $l_f(\hat{\theta}; x, y)$ is the fairness loss, and γ is a hyperparameter to have l_c and l_f at the same scale.

We can solve Eq. 4 by optimizing user and attacker's objectives separately. Intuitively, the user (inner optimization) minimizes the classification loss subject to fairness constraint. The attacker (outer optimization) tries to maximize the joint loss $\mathbb{E}_{(x,y)}[\alpha \cdot l(\hat{\theta}; x, y) + (1-\alpha) \cdot \gamma \cdot l_f(\hat{\theta}; x, y)]$ by creating a poisoning set $\mathcal{D}_p$ based on $\hat{\theta}$ obtained by the user to degrade the performance of classifier either from accuracy or fairness aspect. For example, if the value of α approaches to 1, then the attacker tends to degrade more on the accuracy of the model. Note that the loss expectation is taken over the underlying distribution of the clean data. Inspired by [5], we also approximate the loss function in the outer optimization via the loss on the clean training data and the poisoning data and have

$$\max_{\mathcal{D}_p} [\alpha \cdot \mathcal{L}(\hat{\theta}; \mathcal{D}_c \cup \mathcal{D}_p) + (1-\alpha) \cdot \gamma \cdot \Delta(\hat{\theta}; \mathcal{D}_c \cup \mathcal{D}_p)] \tag{5}$$

The accuracy loss $\mathcal{L}$ and fairness loss Δ may be at different scales due to the use of different loss functions and data distribution. Figure 1 shows the curves of

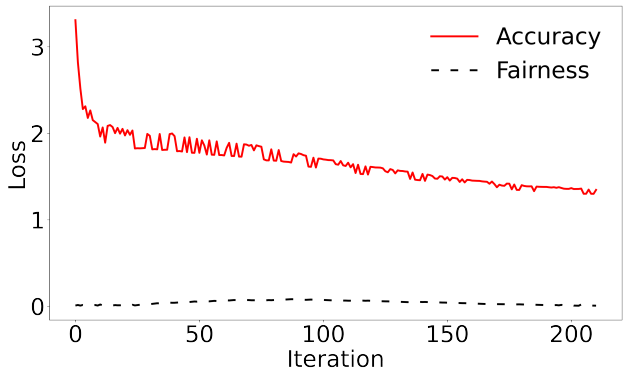


Fig. 1: Accuracy and fairness loss (in terms of equalized odds) with different iterations of PFML-AS ($\alpha = 0.8$) on COMPAS.

accuracy loss and fairness loss of equalized odds when we increase the generated poisoning samples from 1 to 211 on COMPAS dataset (see experiment section for the detailed experimental setting). This shows the importance of introducing hyperparameter γ to have accuracy loss and fairness loss at the same scale.

The user tries to achieve optimal and fair $\hat{\theta}$ under the poisoning set $\mathcal{D}_p$. As the constrained optimization is intractable, we further transform the inner optimization to its dual form as the following:

$$\hat{\theta} = \min_{\theta \in \Theta} \left(\frac{1}{n+t} \mathcal{L}(\theta; \mathcal{D}_c \cup \mathcal{D}_p) + \lambda \Delta(\theta, \mathcal{D}_c \cup \mathcal{D}_p) \right) \tag{6}$$

where λ is the Lagrange multiplier and t is the current size of poisoning samples $\mathcal{D}_p$. By Eq. 5 and Eq. 6, we effectively capture the contribution of each poisoning point (x, y) to both accuracy loss and fairness gap.

3.2 Convex Relaxation of Fairness Constraint

The dual optimization problem in Eq. 6 involves the calculation of $\Delta(\theta, \mathcal{D}_c \cup \mathcal{D}_p)$ over the current $\mathcal{D}_c \cup \mathcal{D}_p$. However, fairness notions such as demographic parity and equalized odds are non-convex. We adopt simplifications proposed by [22] for demographic parity and [7] for equalized odds to reach convex relaxations of fairness constraints. Demographic parity can be approximated by the decision boundary fairness. The decision boundary fairness over $\mathcal{D}_c \cup \mathcal{D}_p$ is defined as the covariance between the sensitive attribute and the signed distance from the non-sensitive attribute vector to the decision boundary. It can be written as:

$$C(\theta, \mathcal{D}_c \cup \mathcal{D}_p) = \frac{1}{n+t} \sum_{i=1}^{n+t} (s_i - \bar{s}) d_{\theta(x_i)} \tag{7}$$

where t is the size of the current poisoning samples $\mathcal{D}_p$, s_i is the value of the sensitive attribute of the sample x_i, $d_{\theta(x_i)} = \theta^T x_i$ is the distance to the decision boundary of the classifier f_θ, $\bar{s}$ is the mean value of the sensitive attribute over $\mathcal{D}_c \cup \mathcal{D}_p$. We require that $|C(\theta, \mathcal{D}_c \cup \mathcal{D}_p)| \leq \tau$ to achieve fairness.

We adopt the fairness definition for equalized odds by balancing the risk among two sensitive groups. Let the linear loss be $\mathcal{L}_l$ (e.g., $\mathcal{L}_l = 0.5(1 - f_\theta(x))$ for SVM model) and denote $\mathcal{D} = \mathcal{D}_c \cup \mathcal{D}_p$. We can write down the convex relaxation for the fairness gap of equalized odds as the following:

$$C(\theta, \mathcal{D}) = \frac{1}{2} \sum_{y=0,1} |R^{y,s=0}(\theta, \mathcal{D}) - R^{y,s=1}(\theta, \mathcal{D})| \tag{8}$$

where $R^{y,s}(\theta, \mathcal{D}) = \frac{1}{n^{y,s}} \sum_{(x,y)\in\mathcal{D}_{y,s}} \mathcal{L}_l(x, y; \theta)$. $\mathcal{D}_{y,s}$ is the dataset of points with group s and label y and $n^{y,s}$ is the size of $\mathcal{D}_{y,s}$. Similar to the approximation of equalized odds, we can use $C(\theta, \mathcal{D}) = |R^{y=1,s=0}(\theta, \mathcal{D}) - R^{y=1,s=1}(\theta, \mathcal{D})|$ for the convex relaxation of equalized opportunity.

Algorithm 2 Poisoning Attack on Fair Machine Learning (PFML)

Require: $\mathcal{D}_c$, $n = |\mathcal{D}_c|$, feasible poisoning set $\mathcal{F}(\mathcal{D}_k)$, number of poisoning data ϵn, penalty parameter (Lagranger multiplier) λ, learning rate η, scaling factor γ, balance ratio α, fairness notion Δ.

Ensure: Poisoning dataset $\mathcal{D}_p$.

Initialize $\theta^0 \in \Theta$
for $i = 1 : I$
  $\theta^i \leftarrow \theta^{i-1} - \eta \left(\frac{\nabla \mathcal{L}(\theta^{i-1}; \mathcal{D}_c)}{n} + \nabla \left[\lambda \Delta \left(\theta^{i-1}, \mathcal{D}_c \right) \right] \right)$
end for
$\theta^0 \leftarrow \theta^I$, $\mathcal{D}_p \leftarrow Null$
for $t = 1 : \epsilon n$
  $(x^t, y^t) \leftarrow argmax_{(x,y)\in\mathcal{F}(\mathcal{D}_k)} [\alpha \cdot l(\theta^{t-1}; x, y) +$
    $(1 - \alpha) \cdot \gamma \cdot \Delta \left(\theta^{t-1}, \mathcal{D}_c \cup \mathcal{D}_p \cup \{(x, y)\} \right)]$
  $\mathcal{D}_p \leftarrow \mathcal{D}_p \cup \left\{ \left(x^t, y^t \right) \right\}$, $\mathcal{F}(\mathcal{D}_k) \leftarrow \mathcal{F}(\mathcal{D}_k) - \{(x, y)\}$
  $\theta^t \leftarrow \theta^{t-1} - \eta \left(\frac{\nabla \mathcal{L}(\theta^{t-1}; \mathcal{D}_c \cup \mathcal{D}_p)}{n+t} + \nabla \left[\lambda \Delta \left(\theta^{t-1}, \mathcal{D}_c \cup \mathcal{D}_p \right) \right] \right)$
end for

3.3 Attack Algorithm

Algorithm 2 shows pseudo code of our poisoning attack framework on fair machine learning (PFML). Our three algorithms are denoted as PFML-AS for adversarial sampling, PFML-AF for adversarial flipping, and PFML-AM for adversarial feature modification. In each algorithm, we can adjust the attack's focus on prediction accuracy or fairness by choosing different α values. For example, when 1 (0), the attack's focus is purely on accuracy (fairness) and when 0.5, the

Table 1: Test accuracy and fairness gap of fair reduction [1] and post-processing [9] with **equalized odds** under PFML and baselines (COMPAS).

	Accuracy					Fairness				
Method	Fair Reduction (δ)				Post (δ)	Fair Reduction (δ)				Post (δ)
	0.12	0.1	0.07	0.05	0	0.12	0.1	0.07	0.05	0
Benign	0.950	0.949	0.949	0.948	0.877	0.108	0.103	0.086	0.082	0.095
RS	0.936	0.930	0.919	0.912	0.839	0.101	0.105	0.104	0.103	0.081
LF	0.935	0.931	0.919	0.911	0.839	0.062	0.066	0.072	0.080	0.109
HE	0.915	0.908	0.899	0.891	0.829	0.076	0.082	0.100	0.109	0.131
INFL	0.850	0.848	0.845	0.841	0.653	0.089	0.081	0.078	0.081	0.054
KKT	0.890	0.891	0.891	0.886	0.701	0.136	0.137	0.137	0.142	0.096
min-max	0.891	0.887	0.878	0.874	0.678	0.096	0.125	0.089	0.075	0.082
AS	0.830	0.824	0.816	0.810	0.740	0.051	0.069	0.111	0.143	0.156
AF	0.823	0.817	0.808	0.803	0.740	0.046	0.059	0.100	0.130	0.136
PFML-, α										
AS, 0	0.853	0.847	0.833	0.802	0.753	0.126	0.148	0.164	0.185	0.190
AS, 0.2	0.843	0.837	0.820	0.792	0.728	0.112	0.124	0.138	0.127	0.188
AS, 0.5	0.824	0.820	0.814	0.809	0.705	0.110	0.118	0.130	0.142	0.143
AS, 0.8	0.820	0.816	0.809	0.800	0.715	0.101	0.105	0.116	0.120	0.099
AS, 1.0	0.811	0.807	0.800	0.796	0.724	0.083	0.071	0.061	0.061	0.074
AF, 0	0.847	0.841	0.832	0.805	0.752	0.120	0.144	0.172	0.184	0.193
AF, 0.2	0.843	0.838	0.817	0.792	0.728	0.107	0.117	0.125	0.126	0.186
AF, 0.5	0.818	0.814	0.808	0.804	0.711	0.101	0.110	0.126	0.139	0.136
AF, 0.8	0.804	0.797	0.791	0.786	0.714	0.093	0.090	0.097	0.107	0.090
AF, 1.0	0.803	0.797	0.794	0.788	0.722	0.088	0.068	0.043	0.039	0.097
AM, 0	0.908	0.906	0.904	0.897	0.764	0.195	0.200	0.215	0.207	0.198
AM, 0.2	0.811	0.805	0.798	0.794	0.731	0.102	0.086	0.076	0.076	0.153
AM, 0.5	0.793	0.788	0.780	0.775	0.688	0.079	0.059	0.071	0.080	0.124
AM, 0.8	0.791	0.789	0.782	0.773	0.696	0.082	0.055	0.053	0.077	0.096
AM, 1.0	0.828	0.823	0.817	0.813	0.696	0.063	0.045	0.055	0.073	0.076

focus is on the combination of fairness and accuracy. In line 2 - 4, we first train FML model on the clean data $\mathcal{D}_c$ and use the fitted parameter θ^I to start generating poisoning samples. We then execute the loop of line 6 - 9 to iteratively generate ϵn poisoning samples. In line 7, when generating the data point (x^t, y^t) with highest impact on a weighted sum of the accuracy loss and the fairness violation with respect to θ^{t-1}, we add both the previously generated data points in $\mathcal{D}_p$ and the data point (x^t, y^t) to $\mathcal{D}_c$. As a result, we can measure the incremental contribution of that data point to the fairness gap $\Delta\left(\theta^{t-1}, \mathcal{D}_c \cup \mathcal{D}_p \cup \{(x, y)\}\right)$. Note that in this step, the accuracy loss can be simply calculated over each point $(x, y) \in \mathcal{F}(\mathcal{D}_k)$ as the accuracy loss of existing data points from $\mathcal{D}_c \cup \mathcal{D}_p$ is unchanged. In line 8, we add the chosen poisoning point (x^t, y^t) to $\mathcal{D}_p$ and also remove it from the feasible poisoning set. In line 9, when updating the model parameters θ, we minimize the penalized loss function over $\mathcal{D}_c$ and $\mathcal{D}_p$. We see

Table 2: Test accuracy and fairness gap of fair reduction and post-processing with **demographic parity** (COMPAS).

	Accuracy					Fairness				
Method	Fair Reduction (δ)				Post (δ)	Fair Reduction (δ)				Post (δ)
	0.12	0.1	0.07	0.05	0	0.12	0.1	0.07	0.05	0
Benign	0.887	0.867	0.803	0.768	0.859	0.175	0.169	0.107	0.095	0.046
RS	0.882	0.839	0.813	0.767	0.867	0.187	0.155	0.130	0.076	0.023
LF	0.890	0.852	0.814	0.775	0.868	0.194	0.166	0.138	0.099	0.021
HE	0.901	0.859	0.808	0.766	0.840	0.205	0.181	0.135	0.098	0.036
INFL	0.879	0.855	0.774	0.748	0.784	0.200	0.186	0.097	0.108	0.015
KKT	0.884	0.875	0.788	0.768	0.817	0.221	0.214	0.127	0.136	0.016
min-max	0.870	0.870	0.843	0.818	0.810	0.201	0.204	0.182	0.167	0.036
PFML-, α										
AS, 0	0.853	0.829	0.771	0.750	0.824	0.195	0.168	0.109	0.099	0.041
AS, 0.2	0.847	0.819	0.766	0.736	0.798	0.189	0.171	0.106	0.100	0.039
AS, 0.5	0.844	0.812	0.763	0.731	0.795	0.182	0.167	0.101	0.092	0.038
AS, 0.8	0.845	0.811	0.758	0.731	0.791	0.175	0.166	0.096	0.094	0.036
AS, 1.0	0.829	0.816	0.757	0.722	0.790	0.171	0.151	0.083	0.075	0.032
AF, 0	0.848	0.822	0.786	0.761	0.822	0.192	0.185	0.098	0.080	0.057
AF, 0.2	0.841	0.805	0.766	0.742	0.806	0.188	0.163	0.095	0.086	0.056
AF, 0.5	0.842	0.809	0.762	0.733	0.801	0.174	0.136	0.087	0.086	0.036
AF, 0.8	0.838	0.803	0.755	0.729	0.798	0.167	0.134	0.086	0.079	0.027
AF, 1.0	0.831	0.808	0.752	0.721	0.793	0.160	0.132	0.082	0.069	0.032
AM, 0	0.883	0.853	0.816	0.791	0.833	0.246	0.219	0.183	0.159	0.031
AM, 0.2	0.840	0.820	0.762	0.730	0.814	0.218	0.208	0.138	0.128	0.038
AM, 0.5	0.838	0.802	0.757	0.733	0.793	0.212	0.170	0.120	0.114	0.030
AM, 0.8	0.826	0.800	0.758	0.720	0.787	0.193	0.147	0.115	0.065	0.031
AM, 1.0	0.853	0.805	0.767	0.726	0.815	0.184	0.138	0.105	0.060	0.029

the execution time is mostly spent on line 6 - 11. In fact, line 9 and line 10 only involve one time operation. The time complexity of line 8 is $\mathcal{O}(m)$, where m is the size of feasible poisoning set $\mathcal{F}(\mathcal{D}_k)$. Therefore, the time complexity of the loop from line 6 - 11 is $\mathcal{O}(\epsilon nm)$. In practice, the size of $\mathcal{F}(\mathcal{D}_k)$ is fixed, and we can simplify time complexity as $\mathcal{O}(\epsilon n)$.

Remarks. Chang et al. [5] presented an online gradient descent algorithm that generates poisoning data points for fair machine learning model with equalized odds. As the fairness gap is not an additive function of the training data points, they used $\mathcal{D}_c \cup \{(x^t, y^t)^{\epsilon n}\}$ (denoted as $\mathcal{D}_t$) to measure the contribution of that data point to the fairness gap where $\mathcal{D}_t$ is equivalent to adding ϵn copies of (x^t, y^t) to $\mathcal{D}_c$. The weighted loss function used for selecting poisoning samples is shown as $\left[\epsilon \cdot l\left(\theta^{t-1}; x, y\right) + \lambda \cdot \Delta\left(\theta^{t-1}, \mathcal{D}_t\right)\right]$. The algorithm then updates the model parameters θ via the gradient descent, i.e., $\theta^t \leftarrow \theta^{t-1} - \eta(\frac{\nabla \mathcal{L}\left(\theta^{t-1}; \mathcal{D}_c\right)}{n} + \nabla\left[\epsilon \cdot l\left(\theta^{t-1}; x^t, y^t\right) + \lambda \cdot \Delta\left(\theta^{t-1}, \mathcal{D}_t\right)\right])$. However, both the use of $\mathcal{D}_c \cup \{(x, y)^{\epsilon n}\}$ to quantify the (x^t, y^t)'s contribution to the fairness gap and the use of param-

eters (ϵ and λ) to define the weighted loss are heuristic, thus hard to produce effective poisoning samples on algorithmic fairness. Moreover, different from [5] that covers only a single fairness notion (i.e., equalized odds) and two attacks (adversarial sampling and adversarial label flipping), our paper presents a general framework with algorithms for three group based fairness notions and a new important adversarial feature modification attack. In our evaluation, we compare our methods with [5] and three new state-of-the-art baselines (influence attack, KKT, and min-max attack) from [10].

4 Experiments

Datasets. We conduct our experiments on COMPAS [11] and Adult [8] which are two benchmark datasets for FML community. COMPAS is a collection of personal information such as criminal history, demographics, jail and prison time. Adult is also a collection of individual's information including gender, race, martial status, and so forth. The task for COMPAS is a binary classification to predict whether the individual will be re-offended based on personal information, while the task for Adult dataset is to predict if an individual's annual income will be over \$50k based on his personal information. We use race (only black/white) as the sensitive attribute for COMPAS and gender as sensitive attribute for Adult. After preprocessing, COMPAS has 5278 data points and 11 features, while Adult has 48842 data points and 14 features. For each dataset, we first train a SVM model on the entire dataset. For the 60% data with the smallest loss, we randomly split them into clean dataset $\mathcal{D}_c$, attack candidate dataset $\mathcal{D}_k$, and test dataset $\mathcal{D}_{test}$ with ratio 4:1:1. The rest 40% data is treated as hard examples and added into $\mathcal{D}_k$. For COMPAS, $\mathcal{D}_c$ contains 2111 samples and $\mathcal{D}_{test}$ has 528 samples. $\mathcal{D}_k$ has 2639 samples including 2112 hard examples. For adversarial label flipping, we randomly flip the label of 15% data from $\mathcal{D}_k$ to build the feasible poisoning candidate set $\mathcal{F}(\mathcal{D}_k)$. For adversarial feature modification, we randomly flip one binary feature of each data point from $\mathcal{D}_k$ and include them to $\mathcal{F}(\mathcal{D}_k)$. Following the similar pre-processing strategy, $\mathcal{D}_c$, $\mathcal{D}_{test}$, and $\mathcal{D}_k$ of Adult contain 15385 samples, 6594 samples, and 26863 samples, respectively. Due to space limit, we report detailed results of COMPAS in the majority of this experiment section and only show the summarized results of Adult in Figure 2 at the end of this experiment section.

Baselines. We consider the following baselines: (a) Random Sampling (RS): attacker randomly selects data samples from $\mathcal{D}_k$; (b) Label Flipping (LF): attacker randomly selects data samples from $\mathcal{D}_k$ and flips their labels; (c) Hard Examples (HE): attacker randomly selects data samples from hard examples set; (d) influence attack (INFL), (e) KKT attack, (f) min-max attack, (g) adversarial sampling (AS), and (h) adversarial flipping (AF). Attacks (d)-(f) are stronger data poisoning attacks breaking data sanitization defenses and all control both the label y and input features x of the poisoned points [10]. Attacks (g) and (h) are designed for attacking FML from [5]. As attacks (g) and (h) are only designed

for equalized odds, we exclude them from baselines when reporting comparisons based on demographic parity.

Fair Classification Models. We use SVM as the classification model and choose fair reduction [1] and post-processing [9] as FML under attack. Post-processing adjusts an unconstrained trained model to remove discrimination based on fairness notions such as demographic parity and equalized odds. After adjustment, the unconstrained model behaves like a randomized classifier that assigns each data point a probability conditional on its protected attribute and predicted label. These probabilities are calculated by a linear program to minimize the expected loss. Fair reduction is an advanced in-processing approach that reduces fair classification to a sequence of cost-sensitive classification and achieves better accuracy-fairness tradeoff than previous FML models. Hence, we do not report results from other in-processing FML models.

Hyperparameters. In our default setting, we choose the number of pretrain steps with $\mathcal{D}_c$ as 2000, learning rate lr as 0.001, penalty parameter λ as 5, and ϵ as 0.1. The scaling factor γ is calculated as the ratio of accuracy loss and fairness loss over $\mathcal{D}_c$. **Metrics.** We run our attacks, PFML-AS, PFML-AF and PFML-AM, each with five α values, and baseline attacks to generate the poisoning data $\mathcal{D}_p$ and then train fair reduction (with four δ values as fairness threshold) and post-processing models with $\mathcal{D}_c \cup \mathcal{D}_p$. Finally we run the trained FML models on the test data $\mathcal{D}_{test}$ and report the test accuracy and fairness gap. For each experiment, we report the average value of five runs. Due to space limit, we skip reporting their standard deviation and instead we summarize comparisons based on t-test.

Reproducibility. All datasets, source code and setting details are released in GitHub with `https://github.com/minhhao97vn/PFML` for reproducibility.

4.1 Evaluation of PFML with Equalized Odds

Table 1 shows the comparison of our PFML attacks under different α with other baseline models in terms of both accuracy and fairness on two FML models (fair reduction and post-processing) trained with equalized odds under different fairness threshold values of δ. In each cell of Table 1, we report the average value of five runs. Due to space limit, we skip reporting their standard deviation and instead we summarize comparisons based on t-test at the end of this subsection.

First, the accuracy of FML model under all three PFML attacks (PFML-AS, PFML-AF and PFML-AM) is significantly lower than the benign case. For each fixed δ, both the accuracy value and fairness gap of FML under PFML attacks decrease when we increase α. Recall that larger α indicates that PFML attacks more on accuracy and smaller α indicates more attack's focus on fairness. Note that larger fairness gap caused by smaller α indicates higher model unfairness. Taking PFML-AS as an example, the accuracy of fair reduction with $\delta = 0.12$ is 0.853 and the fairness gap is 0.126 when $\alpha = 0$; the accuracy is 0.811 and the fairness gap is 0.083 when $\alpha = 1$. This result demonstrates that controlling α is flexible and effective for attackers to tune attack target on either prediction accuracy or fairness. Second, PFML-AF and PFML-AM outperform PFML-AS

in terms of attacking performance from both accuracy and fairness perspectives, which shows modifying input features or flipping labels is more powerful than adversarial sampling. Third, compared to RS, LF and HE, our PFML attacks can reduce more accuracy or incur more unfairness with the same δ for both fair reduction and post-processing. Taking PFML-AS with $\delta = 0.12$ and $\alpha = 0$ as an example, the accuracy is 0.811, which is 0.125, 0.124, and 0.104 lower than that of RS, LF and HE, respectively. Compared to previous FML attacks (AS, AF) [5] and sanitization attacks (INFL, KKT, min-max) [10], our PFML attacks achieve better attack performance in terms of accuracy (fairness) with large (small) α values, which is consistent with our expectation. In particular, the accuracy of PFML-AS with $\alpha = 1$ and $\delta = 0.12$ is 0.811, which is 0.020 lower than AS. If we choose smaller α, the attack performance of PFML on accuracy under performs [5], which is consistent with our expectation.

We also notice, for each fixed α, the accuracy of fair reduction under our PFML attacks decreases when we decrease δ. The fairness gap of fair reduction under PFML-AS (PFML-AF) attack increases when we decrease δ, which indicates the fair reduction model is less robust or more vulnerable when stricter fairness constraint is enforced. However, the fairness gap of fair reduction under PFML-AM attack instead decreases along the decrease of δ. Theoretical analysis is needed to understand the robustness of fair reduction approach with equalized odds under different poisoning attacks.

4.2 Evaluation of PFML with Demographic Parity

Table 2 shows the comparison results of adversarial fair machine learning with demographic parity. Note that we do not compare with online FML attacks (AS, AF) as they do not support demographic parity. Generally we see similar patterns as equalized odds shown in Table 2. For each fixed δ, both the accuracy and fairness gap of FML models under all three PFML attacks decrease when we increase α. This is because smaller α means more attack's focus on fairness. Compared to RS, LF and HE, our PFML attacks can reduce more model accuracy of FML with the same δ for both fair reduction and post-processing. Compared to INFL, KKT and min-max attacks, PFML attacks achieve better attack performance in terms of accuracy drop (fairness gap) of FML models when we set large (small) α values.

For each fixed α, the accuracy of fair reduction under PFML attacks decreases when we decrease δ. This pattern is similar as equalized odds. However, the fairness gap of fair reduction has a clear decreasing trend when δ decreases, which is different from equalized odds. This result actually indicates the fair reduction model with stricter fairness requirement (small δ) is less vulnerable under poisoning attacks.

4.3 Sensitivity Analysis of Hyperparameters

In this section, we evaluate the sensitivity of PFML attacks under different hyperparameters. Table 3 shows the accuracy, fairness gap and execution time

for COMPAS with equalized odds when we change the size of poisoning samples ϵ against fair reduction. In all experiments, we fix $\delta = 0.07$ and $\alpha = 0.8$. In general, with increasing ϵ, the accuracy of fair reduction drops while its fairness gap increases when fair reduction is under each of our PMFL attacks. Note that larger ϵ corresponds to injecting more poisoning data points into the training data, thus causing more accuracy drop and unfairness of the trained FML model.

Table 3: Effects of ratio ϵ for fairness reduction with equalized odds (COMPAS).

Dataset		$\epsilon = 0.025$	$\epsilon = 0.05$	$\epsilon = 0.1$	$\epsilon = 0.15$
Accuracy	INFL	0.891	0.857	0.845	0.820
	KKT	0.912	0.899	0.891	0.884
	min-max	0.918	0.902	0.878	0.850
	PFML-AS	0.882	0.821	0.809	0.799
	PFML-AF	0.867	0.824	0.791	0.794
	PFML-AM	0.891	0.839	0.782	0.777
Fairness Gap	INFL	0.068	0.054	0.078	0.063
	KKT	0.086	0.102	0.137	0.158
	min-max	0.083	0.108	0.089	0.215
	PFML-AS	0.086	0.082	0.116	0.134
	PFML-AF	0.089	0.081	0.097	0.142
	PFML-AM	0.089	0.092	0.077	0.107
Exec. Time (s)	INFL	497.1	915.7	1569.1	2009.5
	KKT	1633.6	2903.3	5503.3	8400.0
	min-max	337.9	597.1	1137.6	1714.6
	PFML-AS	4.8	6.1	8.8	12.1
	PFML-AF	5.1	6.7	9.9	13.9
	PFML-AM	6.5	10.9	16.5	22.5

We also compare with baseline attack models (INFL, KKT, and min-max). We can see with the same ϵ our PFML attacks can degrade the model accuracy more than the baselines, and cause similar or higher level of model unfairness than the baselines in most scenarios. We also report the execution time in Table 3 and we can see the execution time of our PFML attacks increases linearly with increasing ϵ, which is consistent with our time complexity analysis in Algorithm 2. Compared to the baseline models, our PFML attacks takes two or three orders of magnitude less time to generate poisoning samples than baselines.

Table 4: Effects of penalty parameter λ on PFML-AS for post-processing with equalized odds (COMPAS).

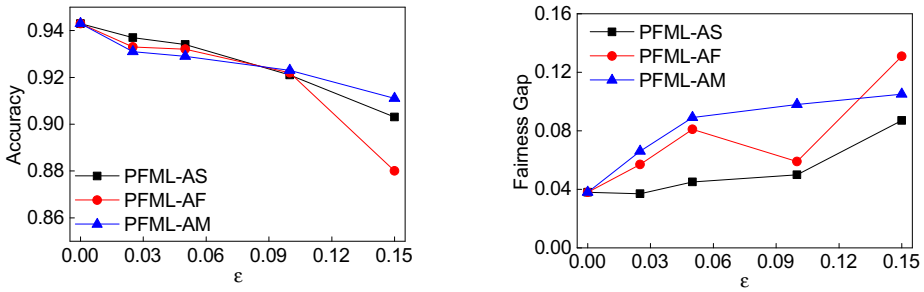
	$\lambda = 1$	$\lambda = 5$	$\lambda = 10$	$\lambda = 15$	$\lambda = 50$	$\lambda = 150$
Accuracy	0.709	0.715	0.718	0.720	0.726	0.723
Fairness Gap	0.094	0.099	0.122	0.118	0.125	0.156

Table 4 shows the accuracy and fairness gap with equalized odds when we use PFML-AS ($\alpha = 0.8$) to attack post-processing FML [9] under different λ values. The post-processing approach has strict fairness constraint $\delta = 0$. We can see with larger λ, the PFML attack focuses more on attacking fairness, which leads to larger fairness gap and smaller accuracy drop of the FML model.

4.4 Significance Testing

For each experiment, we have run our methods and other baseline models five times as shown in all our tables. We apply independent two-sample t-test to compare each of three PFML models (with a given fairness notion and α) with each of baseline models in terms of accuracy reduction and fairness respectively. The t-test results show our PFML attacks significantly outperform baselines from both accuracy and fairness perspectives. Due to space limit, we only report our summarized results here. All p-values except three are less than 0.01 and the left three are still less than 0.1, which demonstrates statistical significance of our PFML methods over baselines.

4.5 Summarized Results of Adult Dataset



(a) Accuracy

(b) Fairness gap

Fig. 2: Effects of ratio ϵ for fairness reduction with equalized odds (Adult).

We also report our summarized results on Adult with the setting of $\delta = 0.1$, $\alpha = 0.8$, and varied ϵ values [3]. Figures 2a and 2b plot the curves of accuracy and fairness gap for each of three PFML attacks with the increasing ϵ. The accuracy of fair reduction under PFML-AS, PFML-AF and PFML-AM attacks decreases when we increase ϵ. This pattern is consistent with our observation on COMPAS. As we analyzed previously, larger ϵ indicates stronger attack as more poisoning data are injected during the model training thus cause more

[3] We conduct experiments on Adult in other settings as COMPAS and observe similar patterns. We skip them due to space limit.

performance degradation. Similarly, the fairness gap under PFML-AS, PFML-AF and PFML-AM generally increases with increasing ϵ. In terms of execution time, our PFML methods take from 187.4 s to 1399.2 s with increasing ϵ, which is significantly less than the baseline models (e.g., two orders of magnitude faster than min-max attack).

5 Related Work

The bulk of recent research on adversarial machine learning has focused on test-time attacks where the attacker perturbs the test data to obtain a desired classification. Train-time attacks leave the test data unchanged, and instead perturb the training data to affect the learned model. Data poisoning attacks are among the most common train-time attack methods in adversarial learning.

Barreno et al. [2] first proposed poisoning attacks which modify the training dataset to potentially change the decision boundaries of the targeted model. The modification can be done by either direct modifying the labels of the training data or manipulating the input features depending on the adversary's capabilities. Biggio et al. [3] developed an approach of crafting poisoning samples using gradient ascent. Shortly speaking, the method identifies the inputs corresponding to local maxima in the test error of the classification model. Mei et al. [15] developed a method that finds an optimal change to the training data when the targeted learning model is trained using a convex optimization loss and its input domain is continuous. Recent approaches include optimization-based methods [10] (e.g., influence, KKT, and min-max), poisoning Generative Adversarial Net (pGAN) model [16], and class-oriented poisoning attacks against neural network models [23]. The influence attack is a gradient-based attack that iteratively modifies each attack sample to increase the test loss, the KKT attack selects poisoned samples to achieve pre-defined decoy parameters, and the min-max attack efficiently solves for the poisoned samples that maximize train loss as a proxy for test loss. All three attacks control both the label and input features of the poisoned points. The structure of pGAN includes a generator, a discriminator, and an additional target classifier. The pGAN model generates poisoning data points to fool the model and degrade the prediction accuracy. Defense methods [10,19] typically require additional information, e.g., a labeled set of outliers or a clean set, and apply supervised classification to separate outliers from normal samples.

There have been a few works on attacking fair machine learning models very recently [5,18,17,14]. Solans et al. [18] introduced an optimization framework for poisoning attacks against algorithmic fairness and developed a gradient-based poisoning attack to increase classification disparities among different groups. Mehrabi et al. [14] also focused on attacking FML models trained with fairness constraint of demographic disparity. They developed anchoring attack and influence attack and focused on demographic disparity. Chang et al. [5] formulated the adversarial FML as a bi-level optimization and focused on attacking FML models trained with equalized odds. To tackle the challenges of the non-convex loss functions and the non-additive function of equalized odds, they further de-

veloped two approximate algorithms. Roh et al. [17] developed a GAN-based model that tries to achieve accuracy, fairness and robustness against adversary attacks.

6 Conclusions and Future Work

In this paper, we present a poisoning sample based framework that can attack model accuracy and algorithmic fairness. Our attacking framework can target fair machine learning models trained with a variety of group based fairness notions such as demographic parity and equalized odds. Our framework enables attackers to flexibly adjust the attack's focus on prediction accuracy or fairness and accurately quantify the impact of each candidate point to both accuracy loss and fairness violation, thus producing effective poisoning samples. We developed three online attacks, adversarial sampling, adversarial labeling, and adversarial feature modification. All three attacks effectively and efficiently produce poisoning samples via sampling, labeling, or modifying a fraction of training data in order to reduce the test accuracy. The three attacks studied in this paper are special cases of gradient-based attacks and belong to indiscriminate attacks. In our future work, we will extend our approach to other attacks, e.g., the targeted attacks that seek to cause errors on specific test examples. We will also investigate robust defense approaches against attacks on fair machine learning models, e.g., by applying multi-gradient algorithms for multi-objective optimization [12,6] and robust learning [20].

Acknowledgement

This work was supported in part by NSF grants 1564250, 1937010 and 1946391.

References

1. Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., Wallach, H.: A reductions approach to fair classification. In: International Conference on Machine Learning. pp. 60–69. PMLR (2018)
2. Barreno, M., Nelson, B., Sears, R., Joseph, A.D., Tygar, J.D.: Can machine learning be secure? In: Proceedings of the 2006 ACM Symposium on Information, computer and communications security. pp. 16–25 (2006)
3. Biggio, B., Nelson, B., Laskov, P.: Poisoning attacks against support vector machines. In: Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012. icml.cc / Omnipress (2012)
4. Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., Mukhopadhyay, D.: Adversarial attacks and defences: A survey. arXiv preprint arXiv:1810.00069 (2018)
5. Chang, H., Nguyen, T.D., Murakonda, S.K., Kazemi, E., Shokri, R.: On adversarial bias and the robustness of fair machine learning. arXiv preprint arXiv:2006.08669 (2020)

6. Désidéri, J.A.: Multiple-gradient descent algorithm (mgda) for multiobjective optimization. Comptes Rendus Mathematique **350**(5-6), 313–318 (2012)
7. Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J., Pontil, M.: Empirical risk minimization under fairness constraints. In: Bengio, S., Wallach, H.M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada. pp. 2796–2806 (2018), https://proceedings.neurips.cc/paper/2018/hash/83cdcec08fbf90370fcf53bdd56604ff-Abstract.html
8. Dua, D., Graf, C.: Adult dataset. https://archive.ics.uci.edu/ml/datasets/adult (1994)
9. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: Advances in neural information processing systems. pp. 3315–3323 (2016)
10. Koh, P.W., Steinhardt, J., Liang, P.: Stronger data poisoning attacks break data sanitization defenses. arXiv preprint arXiv:1811.00741 (2018)
11. Larson, J., Mattu, S., Kirchner, L., Angwin, J.: Compas dataset. https://github.com/propublica/compas-analysis (2017)
12. Liu, S., Vicente, L.N.: The stochastic multi-gradient algorithm for multi-objective optimization and its application to supervised machine learning. arXiv preprint arXiv:1907.04472 (2019)
13. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. arXiv preprint arXiv:1908.09635 (2019)
14. Mehrabi, N., Naveed, M., Morstatter, F., Galstyan, A.: Exacerbating algorithmic bias through fairness attacks. CoRR **abs/2012.08723** (2020), https://arxiv.org/abs/2012.08723
15. Mei, S., Zhu, X.: Using machine teaching to identify optimal training-set attacks on machine learners. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 29 (2015)
16. Muñoz-González, L., Pfitzner, B., Russo, M., Carnerero-Cano, J., Lupu, E.C.: Poisoning attacks with generative adversarial nets. arXiv preprint arXiv:1906.07773 (2019)
17. Roh, Y., Lee, K., Whang, S., Suh, C.: Fr-train: A mutual information-based approach to fair and robust training. In: International Conference on Machine Learning. pp. 8147–8157. PMLR (2020)
18. Solans, D., Biggio, B., Castillo, C.: Poisoning attacks on algorithmic fairness. arXiv preprint arXiv:2004.07401 (2020)
19. Steinhardt, J., Koh, P.W., Liang, P.: Certified defenses for data poisoning attacks. arXiv preprint arXiv:1706.03691 (2017)
20. Taskesen, B., Nguyen, V.A., Kuhn, D., Blanchet, J.: A distributionally robust approach to fair classification. arXiv preprint arXiv:2007.09530 (2020)
21. Yuan, X., He, P., Zhu, Q., Li, X.: Adversarial examples: Attacks and defenses for deep learning. IEEE transactions on neural networks and learning systems **30**(9), 2805–2824 (2019)
22. Zafar, M.B., Valera, I., Rogriguez, M.G., Gummadi, K.P.: Fairness constraints: Mechanisms for fair classification. In: AISTATS (2017)
23. Zhao, B., Lao, Y.: Class-oriented poisoning attack. arXiv preprint arXiv:2008.00047 (2020)