2022 IEEE International Symposium on Information Theory (ISIT) | 978-1-6654-2159-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISIT50566.2022.9834475

Breaking Fair Binary Classification with Optimal Flipping Attacks

Changhun Jo
Dept. of Mathematics
University of Wisconsin-Madison
Madison, Wisconsin, USA
cjo4@wisc.edu

Jy-yong Sohn
Dept. of ECE
University of Wisconsin-Madison
Madison, Wisconsin, USA
sohn9@wisc.edu

Kangwook Lee
Dept. of ECE
University of Wisconsin-Madison
Madison, Wisconsin, USA
kangwook.lee@wisc.edu

***Abstract*—Minimizing risk with fairness constraints is one of the popular approaches to learning a fair classifier. Recent works showed that this approach yields an unfair classifier if the training set is corrupted. In this work, we study the minimum amount of data corruption required for a successful flipping attack. First, we find lower/upper bounds on this quantity and show that these bounds are tight when the target model is the unique unconstrained risk minimizer. Second, we propose a computationally efficient data poisoning attack algorithm that can compromise the performance of fair learning algorithms.**

A full version is at https://arxiv.org/pdf/2204.05472.pdf

I. Introduction

Fairness and robustness are two main requirements for trustworthy artificial intelligence (AI). According to the fairness principle in [1], AI systems should ensure that individuals and groups are free from unfair bias and discrimination. In recent years, researchers have proposed various definitions for fair classification [2], [3] and algorithms for learning fair models [3]–[13]. One popular approach is to solve risk minimization with constraints that capture the desired fairness definition.

While several works theoretically analyzed the risk minimization with fairness constraints [14], [15], our understanding of its performance on noisy or corrupted data is scarce. Given that the use of web-scale training data, crawled from the Internet and/or crowdsourced, has become an essential part of machine learning pipeline [16]–[18], it is of utmost importance to understand how one can learn fair models on data that is potentially corrupted by random or adversarial noise. To understand the robustness of risk minimization with fairness constraints, [19] studied the worst-case scenario – called *data poisoning attacks* – where adversaries can modify training data to make the model learned on it becomes unusable (either due to low accuracy or bias). They designed an online gradient descent algorithm, specifically tailored for attacking constrained risk minimization. Their experimental results showed that constrained risk minimization is so unstable under their attack that the models learned by this approach might be even more unfair than the models learned by unconstrained risk minimization. However, the optimality of the proposed attack algorithm was unknown.

In this work, we study the problem of developing the optimal flipping attack algorithm against risk minimization with fairness constraints. In particular, we consider a general problem setup where an attacker manipulates the data distribution such that the model learned on the poisoned data becomes a given target model. By formulating this attack problem as a bilevel optimization problem, we provide lower and upper bounds on the minimum amount of data perturbation required for a successful flipping attack. Furthermore, if the target model is the unique unconstrained risk minimizer (which generally is *unfair*), then our bounds are tight, and our upper bound provides an explicit construction of the optimal flipping attack algorithm. Fig. 1 illustrates how our attack algorithm makes a fair learning algorithm output the unconstrained risk minimizer, thereby compromising the fairness of it. In other words, when the attacker's goal is to counteract the fairness constraints, our attack algorithm can achieve the goal by perturbing the minimum amount of data. As a byproduct of our analysis, we also show that, under mild assumptions, there exist infinitely many non-trivial fair models that do not suffer from disparate treatment [20], which can be of independent theoretical interest.

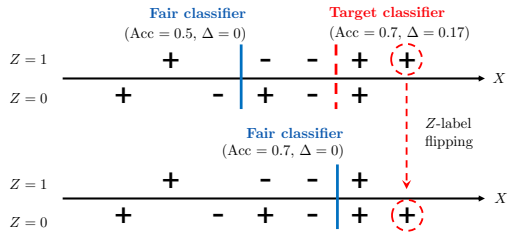


Fig. 1: A toy example that shows how our attack algorithm makes the fair learning algorithm output the unconstrained risk minimizer. We consider a dataset with 10 samples, where X denotes the feature that takes its value in $\mathbb{R}$, $+$ and $-$ denote Y labels to be predicted by the learning algorithm, and Z denotes a sensitive attribute (*e.g.,* gender). We consider linear classifiers that predict samples greater than their thresholds as positive, where the thresholds are shown as vertical lines in the figure. Acc denotes the accuracy, and Δ denotes the fairness gap that measures the unfairness of the classifier (see Def. 1 for details). When $\Delta = 0$, the classifier is perfectly fair and satisfies *equal opportunity* [3], one of the popular fairness metrics. On the clean dataset, the fair learning algorithm outputs the fair classifier, the blue solid line, with Acc $= 0.5$ and $\Delta = 0$. In this example, the attacker's goal is to make the fair learning algorithm to output the unconstrained risk minimizer, the red dashed line, which is unfair because $\Delta = 0.17$. By flipping the Z value of the rightmost sample, the attacker can achieve the goal with the minimum number of label flipping, thereby degrading the fairness of the fair learning algorithm.

II. Related Work

A. Learning Fair Classifiers

Various metrics have been proposed to measure the fairness of a classification model such as demographic parity [2], equalized odds [3], and equal opportunity [3]. Many methods have been proposed to learn fair classifiers, and they can be grouped in four categories: (1) pre-processing methods [4]–[8] that preprocess or reweight training data, (2) in-processing methods [9]–[11], [14], [21]–[24] that enforce fairness constraints or regularizers during the training period, (3) post-processing methods [3], [25]–[27] that manipulate trained models, and (4) adaptive batch selection methods [12], [28]. Several works have also studied fair classification with missing/noisy sensitive attributes [29]–[34].

One prominent approach to learning fair classifiers is Fair Empirical Risk Minimization (FERM), an in-processing method, that solves empirical risk minimization with constraints that capture the desired fairness notion. As fairness constraints are generally non-convex, various relaxations and approximate algorithms have been proposed [10], [15]. While these algorithms are shown to successfully learn fair classifiers, the robustness to adversarial attack is not fully understood yet.

B. Data Poisoning Attacks and Defenses

Data poisoning attacks poison the training set to achieve the adversary's goal [35], [36], and there are two popular approaches; objective-driven attacks and model-targeted attacks. The goal of objective-driven attacks [19], [37]–[41] is to make the learner output a model satisfying a target property, *e.g.,* low accuracy. The goal of model-targeted attacks [39], [40], [42] is to make the learner output a predefined target model.

A few works suggested data poisoning attacks for degrading fairness of learned models. In [41], Solans et al. proposed a gradient-based poisoning attack against ERM to degrade model fairness without significantly degrading accuracy, but theoretical guarantees are missing. Recent works proposed online gradient descent algorithms for poisoning attacks against FERM, with respect to various fairness notions [19], [43], [44]. These existing attack methods are objective-driven attacks aiming at degrading fairness and/or accuracy, while we study a model-targeted attack. We consider the setting where attackers are able to flip the labels and sensitive attributes of data, inspired by recent works on label-flipping attacks [45]–[47].

Several works have theoretically analyzed the behavior of the fairness-aware learner under data poisoning attacks. The authors of [48] proposed a fair learning algorithm with guaranteed accuracy and fairness, under adversarial perturbation on labels and sensitive attributes. The authors of [49], [50] analyzed how the risk and unfairness of the fair learner change as a function of the fraction of the corrupted data, against the attacker who can perturb features, labels, and sensitive attributes of a random subset of the training set. Specifically, [49] provided order-optimal upper/lower bounds on the achievable risk and unfairness performances in a PAC learning sense. Compared with these existing works, the present paper has two key differences in attacker's goal and the attack model. First, while [48]–[50] focused on objective-driven attacks where the attacker's goal is to degrade the accuracy/fairness performance, we consider model-targeted attacks and analyze the minimum amount of perturbation required for a fair learner outputting a predefined target model. Second, given a fixed budget (number of samples) for data poisoning, the attacker considered in [49], [50] poisons a random subset of the samples, while the attacker in our work can choose which subset to poison.

III. Problem Formulation

Let $\mathcal{X}$ denote the set of feature vectors, $\mathcal{Y}$ denote the set of labels, and $\mathcal{Z}$ denote the set of sensitive attributes, *e.g.,* gender and race. We restrict our attention to the case where $\mathcal{X}$ is the n-dimensional real space for any natural number n, and $\mathcal{Y}$ and $\mathcal{Z}$ are binary, *i.e.,* $\mathcal{X} = \mathbb{R}^n$ and $\mathcal{Y} = \mathcal{Z} = \{0, 1\}$. Let X, Y, and Z be the jointly distributed random variables that take values in $\mathcal{X}, \mathcal{Y}$, and $\mathcal{Z}$, respectively. Let $\mathcal{D}$ be the joint distribution of X, Y, and Z. Then $\Pr_{\mathcal{D}}(\cdot)$ and $\mathbb{E}_{\mathcal{D}}[\cdot]$ denote the probability[1] and expectation over $\mathcal{D}$, respectively. In this work, we consider a model $h : \mathcal{X} \to \mathcal{Y}$ that does not suffer from disparate treatment, *i.e.,* h does not take the sensitive attribute $z \in \mathcal{Z}$ as input. Let $\mathcal{H}$ be the hypothesis class. Let $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathcal{Y}$ be the $0/1$ loss function, *i.e.,* $\ell(\hat{y}, y) = \mathbb{1}(y \neq \hat{y})$ where $\mathbb{1}(\cdot)$ is the indicator function. Let $R_\ell(h; \mathcal{D})$ be the true risk of h on $\mathcal{D}$, *i.e.,* $R_\ell(h; \mathcal{D}) = \mathbb{E}_{\mathcal{D}}[\ell(h(X), Y)]$.

We build our theory upon *equal opportunity* [3], but our analysis can be generalized to *demographic parity* [2] (see the full version for details). A model $h : \mathcal{X} \to \mathcal{Y}$ satisfies equal opportunity on the distribution $\mathcal{D}$ if $\Pr_{\mathcal{D}}(h(X) = 1 | Y = 1, Z = 0) = \Pr_{\mathcal{D}}(h(X) = 1 | Y = 1, Z = 1)$. We measure the unfairness of a model by capturing the dissimilarity between true positive rates across the sensitive attributes, which is similar to methods used in [12], [19], [28].

Definition 1. *The* fairness gap *of a model $h : \mathcal{X} \to \mathcal{Y}$ on the distribution $\mathcal{D}$, denoted $\Delta(h, \mathcal{D})$, is*

$$\max_{z \in \mathcal{Z}} \left| \Pr_{\mathcal{D}}(h(X) = 1 | Y = 1, Z = z) - \Pr_{\mathcal{D}}(h(X) = 1 | Y = 1) \right|.$$

For $\delta \in [0, 1]$, h is δ-fair *on $\mathcal{D}$ if $\Delta(h, \mathcal{D}) \leq \delta$. The model h is* perfectly fair *on $\mathcal{D}$ if it is 0-fair. We similarly define the fairness gap, δ-fairness, and perfect fairness of h on the training set D by using the empirical probability $\Pr_D(\cdot)$ over D.*

We use $\mathcal{D}_{X|Y=y,Z=z}$ to denote the distribution of X conditioned on $Y = y, Z = z$ for each $(y, z) \in \mathcal{Y} \times \mathcal{Z}$, and $\mathcal{D}_X$ to denote the marginal distribution of X. For analytical purposes, we assume that, for each $(y, z) \in \{0, 1\} \times \{0, 1\}$, $\mathcal{D}_{X|Y=y,Z=z}$ has the density function[2] $f_{X|Y=y,Z=z}(x|y, z)$

[1]For ease of presentation, we did not mention the σ-algebra over which $\Pr_{\mathcal{D}}(\cdot)$ is defined. When the ambient space is $\mathbb{R}^n$, we consider the Lebesgue σ-algebra, the collection of all Lebesgue measurable sets. When the ambient space is a finite set, we use its power set, the collection of all subsets of it.

[2]Having a density function is closely related to absolute continuity. In this work, we consider a probability distribution over $\mathbb{R}^n$ whose probability space is a triple $(\mathbb{R}^n, \mathcal{L}(\mathbb{R}^n), \nu)$, where $\mathcal{L}(\mathbb{R}^n)$ is the collection of all Lebesgue measurable sets, and the measure ν assigns the probability for $E \in \mathcal{L}(\mathbb{R}^n)$. Then, by the Radon–Nikodym theorem [51], the measure ν has the density function with respect to the Lebesgue measure μ if and only if ν is absolutely continuous with respect to μ.

with respect to the Lebesgue measure μ satisfying $\Pr(X \in E|Y=y,Z=z) = \int_E f_{X|Y=y,Z=z}\,\mathrm{d}\mu$ for any Lebesgue measurable set $E \in \mathbb{R}^n$. Then the joint density function $f(x,y,z)$ of $\mathcal{D}$ is $f_{X|Y=y,Z=z}(x|y,z)\Pr_{\mathcal{D}}(Y=y,Z=z)$, and the marginal density function of X, denoted $f_X(x)$, is $\sum_{(y,z)\in\{0,1\}\times\{0,1\}} f(x,y,z)$.

Learner: We assume that the learner can solve any optimization problem with infinite computing power. Moreover, the learner's hypothesis class $\mathcal{H}$ consists of some Lebesgue measurable functions from $\mathcal{X}$ to $\mathcal{Y}$, so any $h \in \mathcal{H}$ is deterministic. The learner's goal is to find the model in $\mathcal{H}$ that achieves the minimum *true* risk among perfectly fair models, which we call Fair True Risk Minimization (FTRM), by solving the following constrained optimization problem:

$$\min_h\{R_\ell(h;\mathcal{D})\colon\ h\in\mathcal{H}, h \text{ is perfectly fair on } \mathcal{D}\}. \quad (1)$$

We denote the set of solutions of (1) by $\mathcal{A}_0(\mathcal{D})$. Moreover, we define $\mathcal{A}_\delta(\mathcal{D})$ as the set of solutions of $\min_h\{R_\ell(h;\mathcal{D})\colon\ h\in\mathcal{H}, h \text{ is } \delta\text{-fair on } \mathcal{D}\}$. Note that $\mathcal{A}_1(\mathcal{D})$ is the set of unconstrained true risk minimizers since any model is 1-fair.

Attacker: The attacker knows the entire learning procedure (white-box attack) and can make the learner train the model on another distribution $\mathcal{D}'$ with the following constraints. (1) The conditional distribution $\mathcal{D}'_{X|Y=y,Z=z}$ has a density function with respect to the Lebesgue measure μ for each $(y,z)\in\mathcal{Y}\times\mathcal{Z}$; if this does not hold, the attack may be easily detected by the learner. (2) The marginal distribution of X remains the same, *i.e.,* $\mathcal{D}'_X = \mathcal{D}_X$; when $\mathcal{D}$ is a discrete set, *e.g.,* the training set, this assumption limits the attacker to label-flipping attacks. Thus, the attacker's search space $\mathcal{S}$ is

$$\mathcal{S} = \{\mathcal{D}'\colon\ \mathcal{D}' \text{ is a prob. dist. over } \mathcal{X}\times\mathcal{Y}\times\mathcal{Z}, \mathcal{D}'_X = \mathcal{D}_X, \mathcal{D}'_{X|Y=y,Z=z} \text{ has a density w.r.t. } \mu\ \forall(y,z)\in\mathcal{Y}\times\mathcal{Z}\} \quad (2)$$

The attacker's goal is to make the learner output the target model h_{target} with the minimum amount of data perturbation, measured in the total variation distance. For two distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ over $\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}$, the total variation distance between $\mathcal{D}_1$ and $\mathcal{D}_2$, denoted $d_{\text{TV}}(\mathcal{D}_1,\mathcal{D}_2)$, is

$$\tfrac{1}{2}\textstyle\sum_{(y,z)\in\mathcal{Y}\times\mathcal{Z}}\int_{\mathbb{R}^n}|f_1(x,y,z)-f_2(x,y,z)|\,\mathrm{d}\mu, \quad (3)$$

where $f_1(x,y,z)$ and $f_2(x,y,z)$ are (mixed) joint density functions of $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively. Hence the attacker solves the following bilevel optimization problem:

$$\min_{\mathcal{D}'}\{d_{\text{TV}}(\mathcal{D},\mathcal{D}')\colon\ \mathcal{D}'\in\mathcal{S}, \mathcal{A}_0(\mathcal{D}') = \{h_{\text{target}}\}\}. \quad (4)$$

Define the infimum of the objective function of (4) as

$$d^\star_{\text{TV}}(h) = \inf_{\mathcal{D}'\in\Lambda_0(h)} d_{\text{TV}}(\mathcal{D},\mathcal{D}'), \quad (5)$$

where $\Lambda_\delta(h) = \{\mathcal{D}'\colon\ \mathcal{D}'\in\mathcal{S}, \mathcal{A}_\delta(\mathcal{D}') = \{h\}\}$.

IV. Main Results

In this section, we analyze the lower and upper bounds on $d^\star_{\text{TV}}(h)$, the minimum amount of data perturbation for FTRM to output the target model $h\in\mathcal{H}$. Full proofs of all results are deferred to the full version. The following lemma provides the key inequality to derive the lower bound on $d^\star_{\text{TV}}(h)$.

Lemma 1. *Let $\mathcal{X}=\mathbb{R}^n$, $\mathcal{Y}=\{0,1\}$, $\mathcal{Z}=\{0,1\}$, and $\mathcal{D}'\in\mathcal{S}$. If $h\in\mathcal{H}$ is perfectly fair on $\mathcal{D}'$, then*

$$d_{TV}(\mathcal{D},\mathcal{D}') \ge C(h,\mathcal{D}) := \frac{|p_h s_h - q_h r_h|}{\max\{p_h+r_h, q_h+s_h\}} \quad (6)$$

where $p_h = \Pr_{\mathcal{D}}(h(X)=0,Y=1,Z=0)$,
$q_h = \Pr_{\mathcal{D}}(h(X)=1,Y=1,Z=0)$,
$r_h = \Pr_{\mathcal{D}}(h(X)=0,Y=1,Z=1)$,
$s_h = \Pr_{\mathcal{D}}(h(X)=1,Y=1,Z=1)$.

For any $\mathcal{D}'\in\Lambda_0(h)$, h is perfectly fair on $\mathcal{D}'$, so we have $d_{\text{TV}}(\mathcal{D},\mathcal{D}')\ge C(h,\mathcal{D})$ by Lem. 1. Then we get

$$d^\star_{\text{TV}}(h) = \inf_{\mathcal{D}'\in\Lambda_0(h)} d_{\text{TV}}(\mathcal{D},\mathcal{D}') \ge C(h,\mathcal{D}). \quad (7)$$

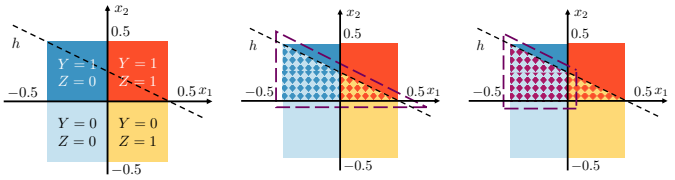
In the lemma below, we show how to construct the distribution $\mathcal{D}' = \text{Fair}_h(\mathcal{D})$ that matches the lower bound in (6). Note that this lemma holds only under a certain assumption on (p_h,r_h,q_h,s_h), and a general version of this lemma that does not require such assumptions is given in the full version.

Lemma 2. *Assume $p_h+r_h \ge q_h+s_h$ and $\frac{q_h}{p_h}\ge\frac{s_h}{r_h}$. Consider a distribution $\text{Fair}_h(\mathcal{D})$ with the density function $f(x,y,z) + \mathbb{1}(h(x)=1,y=1)(2z-1)\frac{q_h r_h - p_h s_h}{(p_h+r_h)q_h}f(x,1,0)$, where f is the density function of $\mathcal{D}$. Then, (i) $\text{Fair}_h(\mathcal{D})\in\mathcal{S}$, (ii) h is perfectly fair on $\text{Fair}_h(\mathcal{D})$, and (iii) $d_{TV}(\mathcal{D},\text{Fair}_h(\mathcal{D})) = C(h,\mathcal{D})$.*

Let us illustrate the density function f_h of $\text{Fair}_h(\mathcal{D})$ given in the lemma. If $h(x)\neq 1$ or $y\neq 1$, then the density function remains the same, *i.e.,* $f_h = f$. If $h(x)=1$ and $y=1$, then $f_h(x,1,0) = (1-\alpha)f(x,1,0)$ and $f_h(x,1,1) = f(x,1,1) + \alpha f(x,1,0)$, where $\alpha = \frac{q_h r_h - p_h s_h}{(p_h+r_h)q_h}$. This can be interpreted as α fraction of the density at $(x,1,0)$ is transported to $(x,1,1)$. In other words, this data distribution can be realized by flipping the Z value with probability α when $X=x$, $Y=1$ and $h(x)=1$. This implies that Z-flipping attack is the optimal way of perturbing data distribution to make a target classifier look perfectly fair, and we will see a similar attack algorithm for the empirical risk case in Sec. V.

Remark 1 (Connection with Theorem 1 in [31])**.** *Theorem 1 in [31] provides the lower bound on $d_{TV}(\mathcal{D}_{Z=z},\mathcal{D}'_{Z=z})$ for each $z\in\mathcal{Z}$ when h is perfectly fair on $\mathcal{D}'$. However, they did not provide an explicit construction of $\mathcal{D}'$ that matches the bound. Our construction scheme can be used to match their bound for certain cases, which we detail in the full version.*

By definition, $d^\star_{\text{TV}}(h)$ is upper bounded by $d_{\text{TV}}(\mathcal{D},\mathcal{D}')$ for any $\mathcal{D}'\in\Lambda_0(h)$. Hence we provide an upper bound on $d^\star_{\text{TV}}(h)$ by constructing a specific distribution $\mathcal{D}'$ that belongs to $\Lambda_0(h)$. The distribution $\text{Fair}_h(\mathcal{D})$ defined in Lem. 2 makes h look perfectly fair with the minimum amount of data perturbation. Assume a hypothetical scenario where h is the *only* perfectly fair model in the hypothesis class $\mathcal{H}$ on $\text{Fair}_h(\mathcal{D})$. Then, $\mathcal{A}_0(\text{Fair}_h(\mathcal{D})) = \{h\}$ holds true. So we get $\text{Fair}_h(\mathcal{D})\in\Lambda_0(h)$, and $d^\star_{\text{TV}}(h)$ could be upper bounded by $d_{\text{TV}}(\mathcal{D},\text{Fair}_h(\mathcal{D}))$, which is equal to $C(h,\mathcal{D})$ by Lem. 2. Unfortunately, this assumption does *not* hold true by the following lemma; there are infinitely many perfectly fair classifiers.



(a) Uncorrupted data $\mathcal{D}$ (b) $\widetilde{\mathcal{D}}$ (c) $\text{Fair}_h(\widetilde{\mathcal{D}})$

Fig. 2: A visualization of our two-stage attack algorithm with 2-dimensional feature space $\mathcal{X}$, where x_1 and x_2 denote the first and second coordinates, respectively. (a) Let $\mathcal{D}$ be a probability distribution over $\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}$ where samples with $Y=y$ and $Z=z$ are uniformly distributed with density of 1 on the square region in the k-th quadrant, where $k=3-y+z-2yz$. The target model h predicts samples above its decision boundary (the black dotted line) as positive ($Y=1$). (b) In the first stage, the attacker constructs $\widetilde{\mathcal{D}}$ from $\mathcal{D}$ by flipping the Y label with probability 0.6 (this can be any number in $(0.5,1)$) when $h(X)=0$ and $Y=1$. As a result, the triangular region with the dashed boundary is perturbed in the first stage. Let $\tilde{f}(x,y,z)$ be the density function of $\widetilde{\mathcal{D}}$. For $x=(x_1,x_2)$ in the blue dotted trapezoidal region, $\tilde{f}(x,y,z)$ is 0.4 if $y=1,z=0$, 0.6 if $y=0,z=0$, and 0 otherwise. For $x=(x_1,x_2)$ in the red dotted triangular region, $\tilde{f}(x,y,z)$ is 0.4 if $y=1,z=1$, 0.6 if $y=0,z=1$, and 0 otherwise. Then h is the risk minimizer on $\widetilde{\mathcal{D}}$, and $\tilde{p_h},\tilde{q_h},\tilde{r_h},\tilde{s_h}$ are $0.075, 0.0625, 0.025, 0.1875$, respectively. (c) In the second stage, the attacker constructs $\text{Fair}_h(\widetilde{\mathcal{D}})$ from $\widetilde{\mathcal{D}}$ by flipping the Z value with probability $2/3$, computed as per the formula in the second stage, when $h(X)=0, Y=1$, and $Z=0$. As a result, the trapezoidal region with the dashed boundary is perturbed in the second stage. Let $\tilde{f}_h(x,y,z)$ be the density function of $\text{Fair}_h(\widetilde{\mathcal{D}})$. For $x=(x_1,x_2)$ in the purple dotted trapezoidal region, $\tilde{f}_h(x,y,z)$ is 0.133 if $y=1,z=0$, 0.267 if $y=1,z=1$, 0.6 if $y=0,z=0$, and 0 if $y=0,z=1$.

Lemma 3. *Let $\mathcal{X}=\mathbb{R}^n$, $\mathcal{Y}=\{0,1\}$, $\mathcal{Z}=\{0,1,\ldots,d-1\}$. Let $\mathcal{D}$ be a probability distribution over $\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}$ whose conditional distribution $\mathcal{D}_{X|Y=y,Z=z}$ has a density function with respect to the Lebesgue measure μ for all $(y,z)\in\mathcal{Y}\times\mathcal{Z}$. If $n\geq d+\mathbb{1}(d\geq 3)$, then there exist infinitely many linear classifiers that are perfectly fair on $\mathcal{D}$. Moreover, for all $x\in\mathcal{X}$, there exist at least one perfectly fair linear classifier whose decision boundary passes through x.*

Remark 2. *While Lem. 3 is stated based on equal opportunity, similar results hold for other fairness metrics; demographic parity and equalized odds. See the full version for details.*

Remark 3. *In [3], Hardt et al. proposed a post-processing method that can find a perfectly fair model on any data distribution. We note that their method outputs a randomized model, hence it is not applicable to our setting where the hypothesis class consists of deterministic models.*

Since $\text{Fair}_h(\mathcal{D})\in\mathcal{S}$ by Lem. 2-(i), Lem. 3 can be applied to $\text{Fair}_h(\mathcal{D})$, and there exist infinitely many linear classifiers that are perfectly fair on $\text{Fair}_h(\mathcal{D})$ (if $n\geq 2$). If $\mathcal{H}$ contains all linear classifiers (which is usually true), then it includes infinitely many models that are perfectly fair on $\text{Fair}_h(\mathcal{D})$. Therefore, in general cases, we cannot guarantee that $\mathcal{A}_0(\text{Fair}_h(\mathcal{D}))=\{h\}$.

This shows the need of sophisticated attack strategies that guarantee both the minimum risk and the perfect fairness of h on the resulting poisoned distribution. We now illustrate our two-stage attack strategy that satisfies the desired properties.

First stage The attacker picks any distribution $\widetilde{\mathcal{D}}$ in $\Lambda_1(h)$, where $\tilde{f}(x,y,z)$ is the density function of $\widetilde{\mathcal{D}}$. Recall that $\Lambda_1(h)=\{\mathcal{D}':\ \mathcal{D}'\in\mathcal{S},\mathcal{A}_1(\mathcal{D}')=\{h\}\}$, and $\mathcal{A}_1(\mathcal{D}')$ is the set of unconstrained risk minimizers $\mathcal{D}'$. We note that it is easy to find distributions in $\Lambda_1(h)$. For example, when $\mathcal{H}$ is the set of all measurable functions from $\mathcal{X}$ to $\mathcal{Y}$, h achieves the minimum risk if and only if h is the Bayes classifier on $\widetilde{\mathcal{D}}$.

Second stage Construct the distribution $\text{Fair}_h(\widetilde{\mathcal{D}})$ in a similar way that we get $\text{Fair}_h(\mathcal{D})$ in Lem. 2. Specifically, calculating probabilities over $\widetilde{\mathcal{D}}$ in the definition of p_h,q_h,r_h,s_h given by Lem. 1, we get $\tilde{p_h},\tilde{q_h},\tilde{r_h},\tilde{s_h}$, *e.g.*, $\tilde{p_h}=\Pr_{\widetilde{\mathcal{D}}}(h(X)=0,Y=1,Z=0)$. Assuming that $\tilde{p_h}+\tilde{r_h}\geq\tilde{q_h}+\tilde{s_h}$ and $\frac{\tilde{q_h}}{\tilde{p_h}}\geq\frac{\tilde{s_h}}{\tilde{r_h}}$, $\text{Fair}_h(\widetilde{\mathcal{D}})$ is a distribution with the density function $\tilde{f}(x,y,z)+\mathbb{1}(h(x)=1,y=1)\cdot(2z-1)\cdot\frac{\tilde{q_h}\tilde{r_h}-\tilde{p_h}\tilde{s_h}}{(\tilde{p_h}+\tilde{r_h})\tilde{q_h}}\tilde{f}(x,1,0)$. A general version of the construction is given in the full version.

Fig. 2 shows how our two-stage attack algorithm works on a toy example. The following proposition provides key properties to derive the upper bound on $d^\star_{\text{TV}}(h)$.

Proposition 1. *Let $\widetilde{\mathcal{D}}\in\Lambda_1(h)$. Then, (i) $\text{Fair}_h(\widetilde{\mathcal{D}})\in\Lambda_0(h)$, and (ii) $d_{TV}(\widetilde{\mathcal{D}},\text{Fair}_h(\widetilde{\mathcal{D}}))=C(h,\widetilde{\mathcal{D}}):=\frac{|\tilde{p_h}\tilde{s_h}-\tilde{q_h}\tilde{r_h}|}{\max\{\tilde{p_h}+\tilde{r_h},\tilde{q_h}+\tilde{s_h}\}}$.*

We are now ready to derive our main theorem providing the lower and upper bounds on $d^\star_{\text{TV}}(h)$.

Theorem 1. *Let h be any model in the hypothesis class $\mathcal{H}$. Then, $C(h,\mathcal{D})\leq d^\star_{TV}(h)\leq\inf_{\widetilde{\mathcal{D}}\in\Lambda_1(h)}(d_{TV}(\mathcal{D},\widetilde{\mathcal{D}})+C(h,\widetilde{\mathcal{D}}))$.*

We note that our bounds on $d^\star_{\text{TV}}(h)$ in Thm. 1 can possibly be loose. However, when the target model h is the unique unconstrained risk minimizer, our bounds are tight by the following corollary.

Corollary 1. *Let h^* be the unique unconstrained risk minimizer $\arg\min_{g\in\mathcal{H}}R_\ell(g;\mathcal{D})$. Then, $d^\star_{TV}(h^*)=C(h^*,\mathcal{D})$.*

Algorithm 1 Z-flipping algorithm

Input: The training set D, the target model h_{target}.
Output: The poisoned training set D'.
for $(a,b,c)\in\{0,1\}\times\{0,1\}\times\{0,1\}$ **do**
  $D_{a,b,c}\leftarrow\{(x,y,z)\in D:\ h_{\text{target}}(x)=a,y=b,z=c\}$
end for
$P\leftarrow|D_{0,1,0}|$, $Q\leftarrow|D_{1,1,0}|$, $R\leftarrow|D_{0,1,1}|$, $S\leftarrow|D_{1,1,1}|$
$\alpha\leftarrow\lfloor|PS-QR|/\max\{P+R,Q+S\}\rfloor$
if $P+R\geq Q+S$ and $\frac{Q}{P}\geq\frac{S}{R}$ **then**
  Randomly choose a subset $\mathcal{T}$ of $D_{1,1,0}$ s.t. $|\mathcal{T}|=\alpha$.
else if $P+R\geq Q+S$ and $\frac{Q}{P}<\frac{S}{R}$ **then**
  Randomly choose a subset $\mathcal{T}$ of $D_{1,1,1}$ s.t. $|\mathcal{T}|=\alpha$.
else if $P+R<Q+S$ and $\frac{Q}{P}\geq\frac{S}{R}$ **then**
  Randomly choose a subset $\mathcal{T}$ of $D_{0,1,1}$ s.t. $|\mathcal{T}|=\alpha$.
else
  Randomly choose a subset $\mathcal{T}$ of $D_{0,1,0}$ s.t. $|\mathcal{T}|=\alpha$.
end if
$\mathcal{T}_{\text{p}}\leftarrow\{(x,y,1-z):\ (x,y,z)\in\mathcal{T}\}$
$D'\leftarrow(D\setminus\mathcal{T})\cup\mathcal{T}_{\text{p}}$

V. Sensitive Attribute Flipping Algorithm

We show how the results made in Sec. IV can be applied to the design of a computationally efficient flipping attack algorithm against FERM. When the target model is the unique unconstrained risk minimizer, as shown in Cor. 1, the attack algorithm proposed in Sec. IV is optimal. Indeed, the first stage of the algorithm is not needed at all in this case, and Z-flipping in the second stage is sufficient for successful attacks.

Inspired by this, we consider the empirical counterpart of the second stage of the attack proposed in Sec. IV. Shown in Alg. 1 is the pseudocode of our attack algorithm. In specific, it computes the number of Z-flipping, denoted α in Alg. 1, using the formula for $C(h, \mathcal{D})$ given in Lem. 1 where (p, q, r, s) are replaced with the empirical counterparts of them. Depending on which of the four conditions hold, it chooses a random subset of size α from the corresponding subset of the training set D. It then simply flips the Z values of them to output the poisoned training set D'. The following proposition ensures that Alg. 1 makes the target model look almost fair on the poisoned training set D' under mild conditions.

Proposition 2. *Let $D = \{(x_i, y_i, z_i)\}_{i=1}^m$ be the training set. Let $D_{a,b,c} = \{(x, y, z) \in D\colon\ h_{target}(x) = a, y = b, z = c\}$, $P = |D_{0,1,0}|$, $Q = |D_{1,1,0}|$, $R = |D_{0,1,1}|$, $S = |D_{1,1,1}|$. If $\frac{P}{m}, \frac{Q}{m}, \frac{R}{m}, \frac{S}{m}$ are $\Omega(1)$, then Alg. 1 makes the target model h_{target} be $O(\frac{1}{m})$-fair on the poisoned training set D'.*

We focus on the case where the target model h_{target} is the empirical risk minimizer on D. Alg. 1 outputs D' on which h_{target} looks almost fair. Moreover, h_{target} still achieves the minimum empirical risk on D' because Z-flipping does not affect the risk. Therefore, our attack algorithm increases the chance of h_{target} being found by the learner's FERM algorithm, thereby degrading the fairness of FERM.

VI. Experimental Results

We evaluate our data poisoning algorithm on the synthetic dataset generated as per the method used in [10], deferring details to the full version. Let D_{train} and D_{test} denote the training set and the test set, respectively. All experiments are repeated 5 times, and the accuracy and unfairness are measured on D_{test}; we use $\Delta(h_{\text{target}}, D_{\text{test}})$ to quantify the unfairness of h_{target}. We compare our attack algorithm with data poisoning attack algorithms: (1) random Y-flip chooses random samples from D_{train} and flips Y values; (2) random Z-flip chooses random samples and flips Z values; (3) random $Y\&Z$-flip chooses random samples and flips both Y and Z values; (4) adversarial sampling (AS) chooses adversarial samples from the feasible attack set using the online gradient descent algorithm proposed in [19] and adds them to D_{train}. We evaluate these attacks against fair learning algorithms: (1) in-processing method using fairness constrains (FC) [10]; (2) fair training against adversarial perturbations (Err-Tol) [48]; (3) fair and robust training (FR-Train) [11] given the clean validation set.

We find h_{target} via empirical risk minimization with logistic loss and get the poisoned training set D' using Alg. 1. Shown in Table I is the performance of attack algorithms against fair learning algorithms. When the learner runs fair learning algorithms on the uncorrupted dataset, the fairness gap significantly decreases at the cost of degraded accuracy, exhibiting a well-known tradeoff between accuracy and fairness. However, with only 3.2% of poisoning rate, our Z-flip attack makes the output be significantly unfair, outperforming (or achieving comparable attack performances to) other attack baselines. Interestingly, our attack successfully degrades the fairness of robust fair training algorithms; Err-Tol and FR-Train. Err-Tol essentially achieves its robustness by relaxing the fairness threshold of its constraints, where the relaxed threshold is carefully calculated using the known poisoning rate. By Prop. 2, our attack makes h_{target} look almost fair on D', so h_{target} satisfies the fairness constraint of Err-Tol. As h_{target} still minimizes the empirical risk on D', Err-Tol will output the model close to h_{target}. FR-Train makes use of the clean validation set to achieve the robustness, but its performance on adversarial Z-flip attacks is not studied in the previous work. We empirically show that our Z-flip attack makes FR-Train output an unfair model with the fairness gap of 0.08.

TABLE I: Comparison with other baseline attack algorithms. The fairness gap Δ measures the unfairness of the model. The target model h_{target} is the output of logistic regression; the accuracy and fairness gap are 0.88 and 0.19, respectively. Our Z-flip attack makes the output be significantly unfair, with only 3.2% of poisoning rate.

	FC [10]		Err-Tol [48]		FR-Train [11]	
Attack method	Acc.	Δ	Acc.	Δ	Acc.	Δ
Uncorrupted	0.79	0.05	0.81	0.06	0.79	0.03
Random Y-flip	0.77	0.01	0.75	0.03	0.76	0.02
Random Z-flip	0.79	0.06	0.87	0.18	0.81	0.04
Random Y&Z-flip	0.80	0.07	0.88	0.19	0.78	0.03
AS [19]	0.78	0.02	0.78	0.03	0.77	0.02
Our Z-flip	0.85	**0.14**	0.88	**0.19**	0.82	**0.08**

VII. Conclusion

We studied poisoning attacks against risk minimization with fairness constraints. We found the lower and upper bounds on the minimum amount of data perturbation required for successful flipping attack for the case of true risk minimization with fairness constraints. Inspired by the fact that sensitive attribute flipping attack is optimal for certain cases, we designed an efficient Z-flipping attack algorithm that can compromise the performance of FERM. We empirically showed that our attack algorithm can degrade the fairness of FERM on synthetic data against existing fair learning algorithms.

We conclude our paper by enumerating important open problems. Our attack algorithm is optimal and our bounds are tight when the target model is the unique unconstrained risk minimizer. Tightening the lower and upper bounds in Thm. 1 for a general target model is an important future work. Our theoretical analysis is limited to the case where both $\mathcal{Y}$ and $\mathcal{Z}$ are binary. We conjecture the theoretical analysis can be extended to the case where $\mathcal{Y}$ and $\mathcal{Z}$ are non-binary. Moreover, it would be interesting to extend our attack algorithm into the federated learning setting.

VIII. Acknowledgements

This work was supported in part by NSF Award DMS-2023239, NSF/Intel Partnership on Machine Learning for Wireless Networking Program under Grant No. CNS-2003129, and the Understanding and Reducing Inequalities Initiative of the University of Wisconsin-Madison, Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation.

REFERENCES

[1] EC, "Ethics guidelines for trustworthy AI," https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419, 2019.

[2] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2015.

[3] M. Hardt, E. Price, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[4] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and Information Systems (KAIS)*, vol. 33, p. 1–33, 2012.

[5] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *ICML*, 2013.

[6] F. Calmon, D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, and K. R. Varshney, "Optimized pre-processing for discrimination prevention," in *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[7] A. Grover, K. Choi, R. Shu, and S. Ermon, "Fair generative modeling via weak supervision," in *ICML*, 2020.

[8] H. Jiang and O. Nachum, "Identifying and correcting label bias in machine learning," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

[9] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi, "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment," in *International Conference on World Wide Web (WWW)*, 2017.

[10] M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi, "Fairness Constraints: Mechanisms for Fair Classification," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

[11] Y. Roh, K. Lee, S. Whang, and C. Suh, "FR-train: A mutual information-based approach to fair and robust training," in *ICML*, 2020.

[12] Y. Roh, K. Lee, S. E. Whang, and C. Suh, "FairBatch: Batch selection for model fairness," in *ICLR*, 2021.

[13] J. Abernethy, P. Awasthi, M. Kleindessner, J. Morgenstern, and J. Zhang, "Adaptive sampling to reduce disparate performance," *arXiv preprint arXiv:2006.06879*, 2020.

[14] A. Agarwal, A. Beygelzimer, M. Dudik, J. Langford, and H. Wallach, "A reductions approach to fair classification," in *ICML*, 2018.

[15] M. Donini, L. Oneto, S. Ben-David, J. Shawe-Taylor, and M. Pontil, "Empirical risk minimization under fairness constraints," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Conference on computer vision and pattern recognition (CVPR)*, 2009.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT)*, 2019.

[18] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[19] H. Chang, T. D. Nguyen, S. K. Murakonda, E. Kazemi, and R. Shokri, "On adversarial bias and the robustness of fair machine learning," *arXiv preprint arXiv:2006.08669*, 2020.

[20] S. Barocas and A. D. Selbst, "Big data's disparate impact," *California Law Review*, pp. 671–732, 2016.

[21] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-aware classifier with prejudice remover regularizer," in *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*, 2012.

[22] S. Yao and B. Huang, "Beyond parity: Fairness objectives for collaborative filtering," in *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[23] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2018.

[24] A. Cotter, H. Jiang, and K. Sridharan, "Two-player games for efficient non-convex constrained optimization," in *International Conference on Algorithmic Learning Theory (ALT)*, 2019.

[25] F. Kamiran, A. Karim, and X. Zhang, "Decision theory for discrimination-aware classification," in *ICDM*, 2012.

[26] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, "On fairness and calibration," in *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[27] E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil, "Leveraging labeled and unlabeled data for consistent fair binary classification," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[28] Y. Roh, K. Lee, S. E. Whang, and C. Suh, "Sample selection for fair and robust training," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[29] A. Lamy, Z. Zhong, A. K. Menon, and N. Verma, "Noise-tolerant fair classification," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[30] P. Awasthi, M. Kleindessner, and J. Morgenstern, "Equalized odds postprocessing under imperfect group information," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

[31] S. Wang, W. Guo, H. Narasimhan, A. Cotter, M. Gupta, and M. I. Jordan, "Robust optimization for fairness with noisy protected groups," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[32] A. Mehrotra and L. E. Celis, "Mitigating bias in set selection with noisy protected attributes," in *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021.

[33] L. E. Celis, L. Huang, V. Keswani, and N. K. Vishnoi, "Fair classification with noisy protected attributes: A framework with provable guarantees," in *ICML*, 2021.

[34] S. Jung, S. Chun, and T. Moon, "Learning fair classifiers with partially annotated group labels," *arXiv preprint arXiv:2111.14581*, 2022.

[35] N. Dalvi, P. Domingos, S. Sanghai, and D. Verma, "Adversarial classification," in *SIGKDD international conference on Knowledge discovery and data mining (KDD)*, 2004.

[36] D. Lowd and C. Meek, "Good word attacks on statistical spam filters," in *Conference on Email and Anti-Spam (CEAS)*, 2005.

[37] B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against support vector machines," in *ICML*, 2012.

[38] H. Xiao, H. Xiao, and C. Eckert, "Adversarial label flips attack on support vector machines," in *European Conference on Artificial Intelligence (ECAI)*, 2012.

[39] S. Mei and X. Zhu, "Using machine teaching to identify optimal training-set attacks on machine learners," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2015.

[40] P. W. Koh, J. Steinhardt, and P. Liang, "Stronger data poisoning attacks break data sanitization defenses," *Machine Learning*, 2021.

[41] D. Solans, B. Biggio, and C. Castillo, "Poisoning attacks on algorithmic fairness," in *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*, 2020.

[42] F. Suya, S. Mahloujifar, D. Evans, and Y. Tian, "Model-targeted poisoning attacks with provable convergence," in *ICML*, 2021.

[43] N. Mehrabi, M. Naveed, F. Morstatter, and A. G. Galstyan, "Exacerbating algorithmic bias through fairness attacks," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[44] M.-H. Van, W. Du, X. Wu, and A. Lu, "Poisoning attacks on fair machine learning," in *International Conference on Database Systems for Advanced Applications (DASFAA)*, 2022.

[45] M. Zhao, B. An, W. Gao, and T. Zhang, "Efficient label contamination attacks against black-box learning models," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2017.

[46] A. Paudice, L. Muñoz-González, and E. C. Lupu, "Label sanitization against label flipping poisoning attacks," in *ECML PKDD Workshops*, 2018.

[47] E. Rosenfeld, E. Winston, P. Ravikumar, and Z. Kolter, "Certified robustness to label-flipping attacks via randomized smoothing," in *ICML*, 2020.

[48] L. E. Celis, A. Mehrotra, and N. K. Vishnoih, "Fair classification with adversarial perturbations," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[49] N. Konstantinov and C. H. Lampert, "Fairness-aware pac learning from corrupted data," *arXiv preprint arXiv:2102.06004*, 2021.

[50] ——, "On the impossibility of fairness-aware learning from corrupted data," in *NeurIPS Workshop: Algorithmic Fairness through the Lens of Causality and Robustness*, 2022.

[51] P. Billingsley, *Probability and Measure*. A Wiley-interscience Publication, 1995.