UCLouvain

epl

École polytechnique de Louvain

Robustness of Fairness-Constrained Classification under Data Poisoning Attacks

Comparative Study of Maximum Entropy and Fairness-Constrained Logistic Regression Models

Author: **Julio Ruiz Guinazu**
Supervisor: **Marco Saerens**
Readers: **Benoit Ronval, Magali Legast, Jeremy deBodt**
Academic year 2025–2026
Master [120] in Computer Science

Declaration Regarding AI Tool Usage in Master's Thesis

During the preparation of this master's thesis, I utilized Claude (Anthropic) for the following purposes:

1. **Improving written content:** I used Claude to correct and improve the syntax, clarity, and coherence of the written text throughout this thesis. The tool helped ensure that my communication is effective and fluid, while all technical content and ideas remain entirely my own.
2. **Assistance with coding:** I used Claude to help debug and improve the code developed for the experiments presented in this thesis. The tool provided clear explanations that facilitated the implementation of the various algorithms and pipelines.

After using Claude, I reviewed and edited all content produced or suggested by the tool. I take full responsibility for the final content presented in this thesis.

Signature: Julio Ruiz Guinazu

List of Contributions

The main contributions of this dissertation are listed hereunder.

- We provide an empirical evaluation of the maximum-entropy logistic regression framework of Vancompernolle et al. [59] under fairness-targeted data poisoning attacks, characterising the effect of both gradient-based and label-flipping attacks on its predictive accuracy and demographic parity guarantees.
- We conduct a direct comparative robustness study between the Maximum Entropy model and the decision-boundary covariance formulation of Zafar et al. [66], showing that the structural difference in where fairness is enforced has measurable consequences for resilience under adversarial conditions.
- We document distinct accuracy-fairness trade-off dynamics under the Solans et al. [55] attack. Specifically, the Zafar et al. [66] formulation exhibits a decoupling of objectives where fairness deteriorates without a corresponding cost to accuracy, whereas the Maximum Entropy model maintains a coupled relationship where fairness preservation is associated with measurable accuracy losses.
- We characterise how the fairness tolerance parameter ε mediates vulnerability to the optimal label-flipping attack of Jo et al. [34], showing that both models are most susceptible under tight fairness constraints across all five benchmark datasets.
- We establish a baseline for adversarial vulnerability by evaluating the attacks on a standard logistic regression model without fairness constraints, allowing for a clear quantification of the specific robustness costs introduced by fairness enforcement mechanisms.

Abstract

As automated decision-making systems become increasingly deployed in high-stakes domains, ensuring that fairness-aware classifiers remain robust to adversarial manipulation has become an important challenge. This thesis studies the robustness of the maximum-entropy logistic regression framework proposed by Vancompernolle et al. under fairness-targeted data poisoning attacks, comparing its performance against the decision-boundary covariance model of Zafar et al. and an unconstrained logistic regression baseline.

Two attack frameworks are considered: a gradient-based poisoning attack and an optimal label-flipping attack. The three models are evaluated on five real-world benchmark datasets under varying poisoning settings and fairness constraints. A key finding is that both attacks produce no statistically significant effect on the unconstrained baseline, establishing that the vulnerability to poisoning is not an inherent property of logistic regression but is introduced directly by the fairness enforcement mechanism. Among the two fairness-constrained formulations, the experimental results reveal important differences in how this vulnerability manifests. The maximum-entropy approach generally preserves lower demographic parity degradation under attack, while the Zafar formulation appears more vulnerable to fairness deterioration despite maintaining stable predictive accuracy, a silent failure mode that standard performance monitoring would not detect.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Prof. Saerens, for his guidance, availability, and valuable feedback throughout this thesis. I am especially thankful for his trust and openness in accepting this project despite the challenging circumstances under which it started, as well as for supervising a topic that extends beyond his usual area of research. His support and encouragement played an essential role in the completion of this work.

I also wish to thank Magali and Benoit for accepting to review this work. I am particularly grateful to Jérémy, who devoted a significant amount of time and provided precious advice that were essential in bringing this thesis to completion.

I would also like to thank my family and friends for their help and support throughout this journey, and especially my dad for his eye-opening advice that helped me shape the direction of this work.

Contents

Introduction

The deployment of machine learning systems in consequential real-world decisions, from credit allocation and recidivism assessment to hiring and medical triage, has brought algorithmic fairness to the forefront of both academic research and public policy. A model trained on historical data does not merely reflect past patterns, it can systematically reproduce and amplify existing social inequities, producing outcomes that disproportionately disadvantage individuals based on protected attributes such as race, gender, or age. This recognition has motivated a rich body of work on fairness-aware classification, which seeks to constrain or correct model behavior so that predictions are not unduly correlated with sensitive group membership.

A substantial part of this literature focuses on in-processing approaches, which incorporate fairness constraints directly into the training objective. These methods offer theoretical guarantees that the trained model will satisfy a specified fairness criterion, typically demographic parity or equalized odds, up to a given tolerance. Yet a critical and underexplored question is whether these guarantees remain meaningful when the training data itself cannot be trusted. In practice, modern machine learning pipelines routinely source training data from public repositories, third-party providers, or user-generated content, all of which are vulnerable to deliberate manipulation by adversarial actors. An adversary who can inject or corrupt a small fraction of training samples, a *data poisoning attack*, may be able to undermine the very fairness properties the model was designed to enforce, potentially while keeping overall accuracy high enough to evade standard monitoring.

This threat is not merely theoretical. Fairness-constrained models are increasingly deployed in regulatory and high-stakes contexts where demonstrable compliance with fairness requirements is legally and institutionally mandated. An attacker who can silently subvert demographic parity while maintaining surface-level accuracy could expose an institution to discriminatory outcomes without triggering any observable alert. Understanding whether and how different fairness formulations resist such attacks is therefore both a scientific and a societal priority.

This thesis addresses this question through a comparative empirical study of two in-processing fairness-constrained classifiers. The decision-boundary covariance model introduced by Zafar et al. [66], and the maximum-entropy logistic regression framework developed by Vancompernolle et al. [59].These two models share a common fairness objective but differ fundamentally in where the constraint is applied: on the linear decision boundary in the Zafar model, and directly on the probabilistic outputs in the Maximum Entropy formulation. This structural difference, together with the inclusion of an unconstrained logistic regression baseline, motivates the central research questions of this thesis:

1. **Does data poisoning significantly affect the Maximum Entropy model?** Specifically, do the gradient-based and label-flipping attacks produce statistically significant degradations in either predictive accuracy or demographic parity for the MaxEnt classifier?
2. **Which model is more robust to data poisoning?** Does the Maximum Entropy formulation, by grounding its fairness constraint directly on probabilistic outputs, offer stronger resistance to adversarial data corruption than the decision-boundary approach of Zafar et al.?
3. **Is the observed vulnerability specific to fairness-constrained models?** Do the attacks produce measurable degradation on an unconstrained logistic regression baseline, or is the attack surface introduced exclusively by the fairness enforcement mechanism itself?

To answer these questions, we subject the models to two complementary poisoning attacks across five benchmark datasets drawn from the fairness-aware machine learning literature. The first attack, proposed by Solans et al. [55], is a gradient-based strategy that injects carefully optimized poisoning points into the training set to maximize disparate impact on a held-out validation set. The second, proposed by Jo et al. [34], is an analytically optimal label-flipping attack that computes the theoretically minimal number of sensitive attribute corruptions required to neutralize the fairness constraint.

The thesis is organized as follows. Chapter 1 surveys the state of the art in algorithmic fairness. Chapter 2 presents the technical foundations, the classifier formulations, the poisoning attack frameworks, and the statistical tools used for comparison. Chapter 3 describes the datasets and preprocessing pipeline. Chapter 4 details the implementation choices. Chapter 5 presents and discusses the experimental results, followed by a conclusion summarising the main findings and directions for future work.

Chapter 1

State of the art

1.1 The fairness problem in machine learning

Before introducing the technical framework used in this work, we first situate the problem of algorithmic fairness in its broader context. This section examines where bias originates in machine learning systems and which groups require protection. These two questions must be answered before any fairness metric or mechanism can be meaningfully applied.

1.1.1 Historical roots and problem framing

Contemporary research on fairness in machine learning is grounded in several decades of research on discrimination. Early systematic investigations, particularly during the 1970s, focused on bias in standardized testing and discriminatory hiring practices. This foundational work introduced core concepts that continue to shape the field, including protected demographic attributes, distinctions between group-level and individual-level fairness, and the recognition that certain formal fairness criteria are mathematically incompatible together [9, 23, 37].

Although these foundations are longstanding, fairness concerns regained prominence in the 2010s with the fast deployment of automated decision-making systems. High-profile cases involving biased risk assessment tools and hiring algorithms brought attention to the public and institutions on the social consequences of machine learning. As noted by Caton and Haas [9], these controversies helped to renew interdisciplinary engagement, emphasizing the need for accountability and rigorous evaluation of algorithmic systems in socially sensitive domains.

A central insight of modern fairness research is that bias often originates in the

data used to train models rather than solely in model design [37, 53]. Historical inequities, sampling biases, and inappropriate feature usage can be encoded into datasets, leading models to systematically disadvantage individuals or groups based on attributes such as race or gender. Importantly, the literature has demonstrated that naive approaches such as removing sensitive attributes (which is often called "blinding") are insufficient, as models can infer protected characteristics through correlated proxy variables, which results in increasing bias and discrimination [9, 38].

This historic background shows that fairness in machine learning (ML) is not just a technical problem, but one with real stakes, which is why understanding both how to build fair systems and how they can be broken is essential.

1.1.2 Sources and causes of unfairness

The literature identifies multiple sources of algorithmic unfairness, which can arise at different stages of the machine learning pipeline [44, 11, 53, 9].

A first source is historical and measurement bias in the training data. Since machine learning models are optimized to reproduce patterns present in data, they inevitably replicate and potentially amplify pre-existing social inequalities, biased human decisions, or flawed measurement processes. When the data-generating process itself is discriminatory, predictive models inherit this discrimination.

A second source concerns missing data, non-representative samples, or biased selection effects, which result in datasets that are not representative of the target population.

A third source stems from the optimization objective of standard learning algorithms. Most models are trained to minimize aggregate prediction error across the entire population. This objective does not account for group-level disparities and may therefore implicitly prioritize performance for larger or statistically dominant groups, leading to unequal error rates across smaller, sensitive groups.

Finally, unfairness may arise due to proxy variables. Even when sensitive attributes such as race or gender are explicitly excluded from the model, other features may be highly correlated with them. In such situations, the model can indirectly reconstruct sensitive information and base its predictions on it, effectively producing disparate outcomes while formally appearing attribute-neutral. For example, a loan approval model that excludes race but uses residential post codes as a feature may still discriminate, because certain post codes are strongly correlated with race due to historical housing segregation, leading to systematically lower approval rates for minority groups.

1.1.3 Sensitive and protected variables

Most approaches to mitigate unfairness, bias, or discrimination are based on the notion of protected or sensitive variables and on un/privileged groups: groups that are disproportionately less/more likely to be positively/negatively classified. Before discussing the key components of the fairness framework, a discussion on the nature of protected variables is needed. Protected variables define the aspects of data which are socio-culturally precarious for the application of ML. Common examples are gender, ethnicity, and age. However, the notion of a protected variable can encompass any feature of the data that involves or concerns people [9, 7].

Deciding which variables require protection is a central challenge. While many attributes are explicitly defined as sensitive by legal frameworks which we do not develop further in this thesis, fewer studies focus on determining whether rare combinations of variables or specific minority groups should also be protected. Some researchers have proposed using variable importance or model influence functions as a way to identify these problematic features [22].

Once these protected variables are identified, we need mathematical tools to measure whether a model is actually treating those groups fairly.

1.2 Measuring fairness

Having established that bias can enter machine learning systems through multiple pathways and that certain groups require explicit protection, we now turn to the question of how fairness can be measured. This section introduces the main mathematical definitions of fairness used in the literature, and motivates the specific metric *demographic parity* that will be enforced throughout this work.

1.2.1 Fairness definitions and metrics

The mathematical formalization of fairness is a core challenge in machine learning, as "fairness" is not a single property but a collection of often conflicting requirements. These definitions are generally categorized based on whether they focus on individuals, groups, or causal relationships [53, 9, 7].

To maintain consistency, we denote $S \in \{0, 1\}$ as a binary sensitive attribute (e.g., gender or race), where $S = 1$ represents the privileged group and $S = 0$ the unprivileged group. Let Y be the actual ground-truth label, and $\hat{Y}$ the model's prediction.

The legal domain has introduced two main definitions of discrimination:

- **Disparate treatment:** Direct discrimination where the model explicitly uses S to make decisions [2]. This is often mitigated by Fairness through Unawareness, though simply removing S is usually insufficient due to proxy variables [20].
- **Disparate impact:** Indirect discrimination where a neutral process results in a disproportionately negative outcome for a protected group. This is measured by the outcomes rather than the inputs [2].

While the legal definitions of disparate treatment and impact provide a conceptual framework, the machine learning literature has developed various mathematical metrics to quantify these phenomena. These measures are generally categorized by whether they prioritize parity in outcomes, parity in error rates, or the consistent treatment of similar individuals [53, 9].

Group-level fairness metrics seek to ensure that the distribution of outcomes is similar across protected groups. The two most prominent measures in this category are Disparate Impact and Demographic Parity (also known as statistical parity).

Disparate impact Designed to mirror the "80 percent rule" used in U.S. employment law [21], this measure requires the ratio of positive prediction rates between the unprivileged and privileged groups to be above a certain threshold $1 - \varepsilon$. Formally:

$$\frac{P(\hat{Y} = 1 \mid S = 0)}{P(\hat{Y} = 1 \mid S = 1)} \geq 1 - \varepsilon \tag{1.1}$$

Here, $P(\hat{Y} = 1 \mid S = 0)$ denotes the probability that an individual from the unprivileged group receives a positive prediction, while $P(\hat{Y} = 1 \mid S = 1)$ denotes the same probability for the privileged group. The parameter ε represents the allowed deviation from perfect parity. If the ratio falls below the threshold $1 - \varepsilon$, the model is considered to exhibit disparate impact against the unprivileged group.

Demographic parity This metric calculates the absolute difference between the rates of positive outcomes. Fairness is achieved when the probability of a positive prediction is independent of the sensitive attribute[7, 53]:

$$|P(\hat{Y} = 1 \mid S = 1) - P(\hat{Y} = 1 \mid S = 0)| \leq \varepsilon \tag{1.2}$$

In practice, especially when working with probabilistic classifiers, this definition is often relaxed by considering the expected predicted probabilities instead of binary

decisions. This leads to the following formulation:

$$\left|\mathbb{E}[\hat{Y} \mid S = 1] - \mathbb{E}[\hat{Y} \mid S = 0]\right| \leq \varepsilon$$

This relaxation is consistent with the maximum entropy framework used in this work, where fairness constraints are applied directly to probabilistic outputs. Moreover, this formulation is closely related to enforcing a near zero covariance between the predictions $\hat{Y}$ and the sensitive attribute S, which provides a convenient linear constraint for optimization.

A significant limitation of these measures is that a perfectly accurate classifier may be deemed unfair if the base rates (the actual prevalence of the outcome) differ significantly between groups. For example, if one group has a higher historical qualification rate for a task, forcing demographic parity may require treating similar individuals differently based on their group membership. In order to satisfy demographic parity, two similar individuals may be treated differently since they belong to two different groups. Such treatment is prohibited by law in some cases [20, 53].

To address the shortcomings of demographic parity, researchers have proposed measures that condition on the ground-truth label Y. These metrics ensure that the model's error rates are balanced across groups.

Equalized odds This criterion requires that both the True Positive Rate (TPR) and the False Positive Rate (FPR) are equal across groups [27]. This ensures that the probability of being correctly or incorrectly classified is not dependent on the sensitive attribute:

$$\left|\mathbb{E}[\hat{Y} \mid S = 1, Y = 0] - \mathbb{E}[\hat{Y} \mid S = 0, Y = 0]\right| \leq \varepsilon$$

$$\left|\mathbb{E}[\hat{Y} \mid S = 0, Y = 1] - \mathbb{E}[\hat{Y} \mid S = 1, Y = 1]\right| \leq \varepsilon$$

The first condition enforces parity in the False Positive Rate (FPR), ensuring that individuals who do not belong to the positive class ($Y = 0$) are incorrectly classified as positive at similar rates across groups. The second condition enforces parity in the True Positive Rate (TPR), ensuring that qualified individuals ($Y = 1$) are correctly classified as positive with similar probability regardless of group membership. The parameter ε specifies the maximum allowed deviation between groups.

A well-documented application of this metric is the analysis of the COMPAS recidivism algorithm. Research revealed that while the tool maintained similar accuracy across groups, it failed the equalized odds criterion: black defendants were twice as likely to be incorrectly flagged as high-risk (FPR), while white defendants were significantly more likely to be incorrectly flagged as low-risk (FNR) [1].

Equal opportunity A relaxed version of equalized odds, this metric focuses solely on the True Positive Rate [27]. It requires that individuals who would benefit from a positive outcome (e.g., those who would repay a loan) have an equal probability of being correctly identified, regardless of their group:

$$\left|P(\hat{Y} = 1 \mid S = 0, Y = 1) - P(\hat{Y} = 1 \mid S = 1, Y = 1)\right| \leq \varepsilon \tag{1.3}$$

This condition enforces parity in the True Positive Rate between the privileged and unprivileged groups. In other words, individuals who would benefit from a positive outcome (e.g., those who would repay a loan) should have a similar probability of being correctly identified by the model [7].

While these metrics allow for "perfect" classifiers to be considered fair, they rely on the assumption that the ground-truth labels Y are themselves unbiased. If the historical data used for Y was collected through a biased process, error rate parity may simply perpetuate those existing inequities.

Individual fairness Contrasting with group-level parity, Individual Fairness operates on the principle that "similar individuals should be treated similarly" [20]. This is mathematically expressed by requiring that the difference between the predicted outcomes for two individuals be no greater than the distance between the individuals themselves:

$$\left|\mathbb{E}[\hat{Y} \mid \mathbf{x}_i] - \mathbb{E}[\hat{Y} \mid x_j]\right| \leq L \cdot d(x_i, x_j)$$

Where x_i and x_j denote the feature representations of two individuals. The function $d(x_i, x_j)$ is a distance metric defined over the feature space. S() refers to the individual's sensitive attributes and X() refers to their associated features. $d(x_i, x_j)$ is a distance metric between individuals that can be defined depending on the domain such that similarity is measured according to an intended task [53].

While theoretically robust, the primary challenge of individual fairness lies in the difficulty of defining an appropriate similarity metric that captures all relevant features without incorporating biased proxies.

These metrics allow us to audit a model's behavior, but achieving fairness in practice requires active intervention through specific mechanisms designed to reduce bias.

1.2.2 Covariance as a Proxy for Demographic Parity

Another prominent approach to enforcing demographic parity consists in using the covariance between the sensitive attribute S and the model's prediction $\hat{Y}$ (or its probabilistic output) as a practical fairness proxy. Several recent works, including the maximum-entropy logistic regression framework, explicitly incorporate a covariance-based constraint of the form $\mathrm{cov}(S, \hat{Y}) = 0$ (or its absolute value bounded by a small tolerance ε) directly into the optimization problem.

The logic behind this choice is that zero covariance implies no correlation between the sensitive attribute and the prediction. In the binary sensitive attribute case, enforcing $\mathrm{cov}(S, \hat{Y}) = 0$ is mathematically equivalent to equalizing the expected prediction across groups:

$$\mathbb{E}[\hat{Y} \mid S = 0] = \mathbb{E}[\hat{Y} \mid S = 1]$$

which is precisely the definition of demographic parity. Thus, covariance provides a simple, differentiable, and computationally convenient linear proxy for the independence condition $S \perp \hat{Y}$.

However, it is important to distinguish between full statistical independence and the weaker notion of linear independence captured by covariance. While $\mathrm{cov}(S, \hat{Y}) = 0$ removes any linear dependence, it does not guarantee that S and $\hat{Y}$ are statistically independent. Non-linear dependencies (e.g., quadratic or higher-order relationships) may persist even when the covariance is exactly zero. This constitutes a fundamental limitation of covariance-based fairness constraints: they can eliminate linear disparate impact while leaving more subtle, non-linear forms of discrimination unaddressed [59].

Despite this limitation, covariance remains extremely valuable in practice. It leads to convex optimization problems when combined with logistic regression or other linear models, admits straightforward Lagrangian dual formulations with linear constraints, and integrates naturally with maximum-entropy principles. Moreover, the post-processing methods proposed by Vancompernolle et al. [60] further exploit covariance constraints (via ℓ_1/ℓ_2-norm penalties) to correct the outputs of any black-box classifier while preserving as much predictive accuracy as possible.

Table 1.1: Summary of common fairness definitions and metrics.

Metric	Category	Core objective	Mathematical definition / condition
Disparate Impact	Group / Outcome	Ensures that the proportion of positive predictions is similar across groups.	$\frac{P(\hat{Y}=1\|S=0)}{P(\hat{Y}=1\|S=1)} \geq 1-\varepsilon$
Demographic Parity	Group / Outcome	Similar to disparate impact, but the difference is taken instead of the ratio.	$\|P(\hat{Y}=1 \mid S=0) - P(\hat{Y}=1 \mid S=1)\| \leq \varepsilon$
Equalized Odds	Group / Error Rate	Equalize both False Positive and True Positive rates.	$P(\hat{Y}=1 \mid S, Y=y) = P(\hat{Y}=1 \mid Y=y)$
Equal Opportunity	Group / Error Rate	Equalize the True Positive rate (benefit) across gropus.	$P(\hat{Y}=1 \mid S=0, Y=1) = P(\hat{Y}=1 \mid S=1, Y=1)$
Covariance Proxy	Group / Optimisation	Eliminate linear dependence between S and $\hat{Y}$.	$\text{cov}(S, \hat{Y}) \approx 0$
Individual Fairness	Individual	Treat similar individuals similarly via a distance metric.	$\|\hat{Y}(x_i) - \hat{Y}(x_j)\| \leq L \cdot d(x_i, x_j)$

1.2.3 Adversarial Vulnerability of Fairness Metrics

The metrics presented above differ not only in their normative assumptions but also in how they respond to adversarial manipulation of the training data. These structural differences are worth flagging here, as they directly motivate the attack framework developed later in this work.

Demographic parity is particularly exposed to base rate manipulation. Because it conditions solely on predictions and not on ground-truth labels, an attacker who selectively shifts the empirical class distribution $\hat{P}(Y=1 \mid S=s)$ for one protected group can force the constraint to produce systematically unequal error rates without any visible fairness violation under the monitored metric [32, 55]. Equalized odds, by contrast, conditions on the true label Y, making it more sensitive to label flipping attacks that corrupt the ground-truth signal for a specific subgroup [34]. Individual fairness, while theoretically robust, depends entirely on the validity of the chosen similarity metric, which can itself be distorted through feature manipulation [47].

These observations imply that the choice of fairness metric is not purely a normative decision, it is also a security decision. It determines which attack surfaces the deployed system exposes. The mathematical reasons behind these vulnerabilities are examined in Section 1.3.3, and their practical consequences for poisoning attacks are developed in Section 1.4.

1.3 Enforcing fairness

Measuring fairness is a necessary first step, but it is not sufficient. A model that is audited and found to be unfair must be corrected. This section reviews the

three main families of techniques for incorporating fairness into a machine learning pipeline, and discusses the theoretical cost that any such intervention necessarily incurs in terms of predictive accuracy.

1.3.1 Fairness-enhancing mechanisms

Once we have decided which fairness metrics to use, we have to decide how to implement them. These mechanisms are typically categorized into three types based on where they intervene in the machine learning pipeline: pre-processing, in-processing, and post-processing. The following subsections review the literature within each of these categories, followed by a comparative analysis and guidelines for their application.

Pre-processing

Pre-processing approaches recognize that algorithmic unfairness often stems from the training data itself. As seen in the previous section, if the underlying data distributions for protected variables is biased the model will inevitably learn these patterns. Pre-processing mechanisms aim to transform the data to remove these biases before the training phase begins [35, 7].

The main objective is to train a model on a repaired or cleaner dataset. Common techniques include reweighing samples, suppressing sensitive features, or mapping the data into a latent space where the sensitive attribute is no longer identifiable. As noted by Caton and Haas [9], pre-processing is the most flexible intervention because it makes no assumptions regarding the choice of the subsequent modeling technique, allowing practitioners to use standard, off-the-shelf machine learning libraries without modification.

In-processing

In-processing approaches target the learning process directly. These mechanisms recognize that standard optimization functions, which typically prioritize aggregate accuracy, can become biased by dominant features or skewed distributions. To address this, in-processing techniques incorporate fairness metrics into the model's objective function as a constraint or a penalty term. For example Zafar et al. [66] propose incorporating fairness constraints directly into the model's optimization objective by minimizing the covariance between the sensitive attribute and the decision boundary.

The goal is to navigate the "fairness-accuracy" trade-off during model parameterization, ensuring the final model converges toward a state that satisfies both

performance and parity requirements [53].

Post-processing

Post-processing approaches intervene after the model has been trained. These mechanisms assume that the output of an ML model may be unfair to specific protected groups, even if the model itself was optimized for fairness. By applying transformations to the model's predictions, such as adjusting decision thresholds for different demographic groups. For example [13] proposes selecting separate thresholds for each group separately, in a manner that maximizes accuracy and minimizes demographic parity [9].

The advantage of post-processing is its applicability to "black-box" scenarios. Since it only requires access to the model's predictions and the sensitive attribute information, it can be deployed even when the internal structure or the original training data is unavailable.

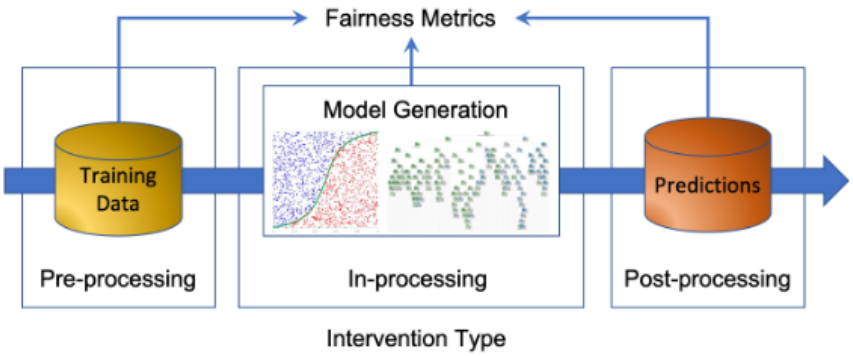


Figure 1.1: Fairness-enhancing mechanisms in the machine learning pipeline: pre-processing (before training), in-processing (during learning), and post-processing (after prediction). Reproduced from: [9].

1.3.2 Accuracy-Fairness trade-off

One of the core theoretical pillars of fairness-aware supervised classification is the accuracy–fairness trade-off. Imposing fairness constraints necessarily restricts the feasible set of classifiers. In the unconstrained case, the Bayes-optimal classifier maximizes predictive accuracy by freely exploiting all correlations present in the data. However when a fairness constraint is added, the optimizer is forced to operate in a strictly smaller feasible region, which may exclude the unconstrained optimum. Consequently, the best achievable accuracy under the constraint is almost

always strictly lower than the unconstrained optimum, except in the degenerate cases where the data already satisfy the fairness condition perfectly [37, 27, 66].

This phenomenon admits a clean geometric interpretation. Let $\mathcal{H}$ denote the full hypothesis space of probabilistic classifiers. A fairness constraint (typically a linear or convex restriction on the probabilistic outputs) defines a feasible subset $\mathcal{H}_{\text{fair}} \subseteq \mathcal{H}$. The unconstrained accuracy-maximizing classifier $\mathbf{h}^* \in \mathcal{H}$ generally violates the fairness condition and therefore lies outside $\mathcal{H}_{\text{fair}}$.

The fairness-constrained optimum

$$\mathbf{h}^*_{\text{fair}} = \arg \max_{\mathbf{h} \in \mathcal{H}_{\text{fair}}} \text{accuracy}(\mathbf{h})$$

must therefore lie on the boundary of the feasible set. By construction, it satisfies

$$\text{accuracy}(\mathbf{h}^*_{\text{fair}}) \leq \text{accuracy}(\mathbf{h}^*)$$

The "distance" between these two points in function space quantifies the accuracy cost imposed by the fairness constraint [23, 9].

The set of all achievable (accuracy, fairness) pairs traces a Pareto frontier: for every admissible level of fairness (often parameterized by a tolerance $\varepsilon \geq 0$), there exists a maximal attainable accuracy. Moving along this frontier corresponds to tightening or relaxing the fairness constraint (see Figure 1.2). Recent comprehensive surveys confirm that the shape of the frontier is highly dataset-dependent and that different fairness notions induce qualitatively different trade-off curves [31, 9].

In practice, the choice of ε is therefore a policy decision rather than a purely technical one: a too-strict ε may render the model unusable for high-stakes applications, while a too-lenient ε fails to deliver meaningful bias mitigation. Empirical studies consistently show that the steepest part of the frontier occurs at moderate ε values, offering the most favorable accuracy–fairness compromises [66, 19].

These theoretical insights connect directly to constrained-optimization frameworks. Any method that augments the empirical risk minimization objective with fairness penalties or hard constraints is explicitly navigating the same restricted feasible set described above. The maximum-entropy logistic regression formulation, for instance, accommodates that linear fairness constraints on the probabilistic outputs while preserving the convexity of the problem, recovering the Pareto-optimal solution for a chosen ε without sacrificing the interpretability of logistic regression [59]. Understanding the geometry and the quantitative cost of the trade-off is necessary both for interpreting the behavior of such methods and for justifying the selected fairness level in a deployment context.

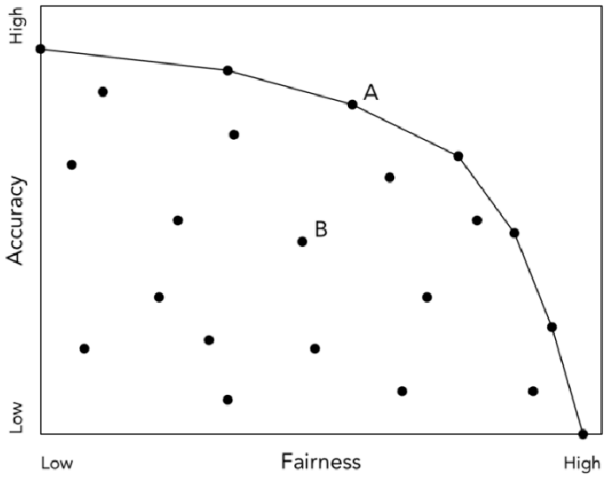


Figure 1.2: Pareto frontier of the accuracy-fairness trade-off. Point A represents a group-aware model in the frontier, while point B represents a blind model in the interior. Reproduced from: [30]

While the accuracy-fairness trade-off is a major practical concern, there is an even deeper problem, some fairness goals are mathematically impossible to achieve simultaneously.

1.3.3 The impossibility results

A central theoretical result in algorithmic fairness is the fundamental incompatibility of several widely used fairness criteria. Kleinberg et al. [37] proved that, except in highly constrained special cases, no probabilistic classifier can simultaneously satisfy three intuitive notions of fairness: (i) calibration (the predicted score equals the true positive rate conditional on the score and the sensitive attribute), (ii) equal false-positive rates across groups, and (iii) equal false-negative rates across groups. This impossibility result was independently corroborated and extended by Chouldechova [11] and has since become a cornerstone of the literature [9, 23, 27].

The formal intuition stems from differing base rates $P(Y = 1 \mid S = s)$ across protected groups s. When base rates are unequal, the only scoring functions that are perfectly calibrated for every group must assign different score distributions to each group. However, equalized odds (requiring identical true-positive rates TPR_s and false-positive rates FPR_s for all s) forces the classifier to treat the groups identically in terms of error rates. These two requirements can be satisfied simultaneously only if the base rates are identical or if the classifier is trivial (always predicts

the majority class or is perfectly accurate). In other words, the joint satisfaction of calibration and equalized odds imposes a strong equality constraint on the group-wise prevalence of the positive class [37, 27].

To illustrate this conflict in practice, consider a simple two-group example. Let group A have a base rate of $P(Y = 1 \mid S = A) = 0.2$ and group B have $P(Y = 1 \mid S = B) = 0.8$. Suppose we desire a risk score $R \in [0, 1]$ that is *perfectly calibrated* across both groups:

$$P(Y = 1 \mid R = r,\, S = s) = r \quad \forall\, r \in [0, 1],\ s \in \{A, B\}$$

To maintain this calibration, the model must output scores concentrated around 0.2 for Group A and 0.8 for Group B. The conflict arises when a decision threshold (e.g., T=0.5) is applied to these calibrated scores to make binary predictions.

To satisfy both conditions, a model must either break calibration or revert to triviality. In the first case, the classifier systematically over or under-predicts for certain groups. For instance, by applying disparate thresholds such as $T_A = 0.3$ and $T_B = 0.7$ to equalize error rates, which ensures that a specific score no longer represents the same underlying probability of success across groups. Alternatively, the model may collapse into a trivial constant predictor that ignores all predictive features and assigns the same outcome to every individual, making the system functionally useless in a practical deployment context.

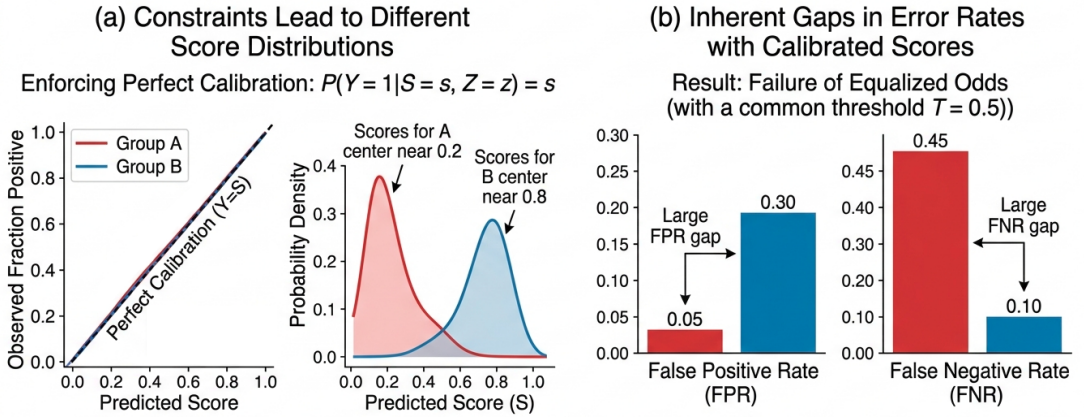


Figure 1.3: The conflict between calibration and equalized odds. When base rates differ between groups, achieving perfect calibration (left) forces a divergence in error rates (right).

This example highlights the deep statistical incompatibilities that emerge when protected groups exhibit disparate underlying outcome distributions [37, 23]. While

these impossibility results do not make fairness unachievable, they establish fundamental limits on what can be guaranteed. In the context of constrained optimization frameworks (such as the maximum-entropy logistic regression introduced earlier), this means the practitioner must select a subset of the three classic criteria and tune the tolerance parameter (ε) accordingly, while remaining aware of the implicit assumptions about base-rate equality.

Furthermore, these theoretical constraints represent more than just a design challenge, they create inherent security vulnerabilities. By strategically corrupting the training data, an adversary can exploit these mathematical tensions to push a fairness-constrained model across the fairness-accuracy frontier in harmful or undesirable ways. The remainder of this work demonstrates how these vulnerabilities can be systematically targeted through poisoning attacks, illustrating that even sophisticated fairness mechanisms remain fragile under adversarial conditions.

1.4 Attacking fairness

The previous sections have treated fairness as a property to be achieved and maintained. We now adopt a different perspective: that of an adversary who seeks to deliberately undermine it. This shift is motivated by a practical reality, fairness-constrained models are increasingly deployed in high-stakes settings, making them attractive targets for manipulation.

The incentives to attack a fairness-constrained system are varied and concrete. A competitor may benefit economically from making a rival's algorithm appear biased, triggering regulatory scrutiny or reputational damage. A regulated entity may seek to bypass legal fairness requirements while maintaining discriminatory practices in deployment. Or an adversary may simply aim to destroy public trust in an institution by causing its fair system to produce high-profile discriminatory errors [57, 10].

This section introduces data poisoning attacks, the primary threat to fairness-aware classifiers at training time, and examines how they can be designed to exploit the very mathematical properties that make fairness enforcement difficult.

1.4.1 Basics of poisoning attacks

As established by Tian et al. [57], poisoning attacks were first proposed by Barreno et al. [3] and have since become the primary security threat during the training stage of machine learning. Machine learning models can be divided into supervised and unsupervised learning based on the nature of the training dataset. In supervised learning, samples are labeled with corresponding classes, and the algorithm aims

to learn the knowledge instructed by these sample-label pairs [15]. Conversely, unsupervised learning datasets contain only samples, requiring the algorithm to mine class information or structures independently [29]. Due to the wide application of supervised models in sensitive decision-making, their security vulnerabilities have attracted significant attention. Consequently, this research focuses primarily on poisoning attacks within the supervised learning scenario [57].

The security landscape of these models is fundamentally defined by the stage of the machine learning lifecycle at which an adversary intervenes. Security attacks are categorized into two classes: the training stage and the inference stage [5, 57]. In the inference stage, a well-trained model is used to process unseen inputs. Here, an adversary may generate adversarial examples to maliciously alter specific predictions. However, because the adversary cannot modify the underlying model at this stage, such attacks are relatively independent and result in no permanent impact on the victim system. In contrast, poisoning attacks occur during the training stage and represent a structural threat. By acting as a "legitimate-but-malicious" actor, the adversary follows the training protocol legitimately but implants carefully constructed poisoning samples, either as raw data or pre-processed feature files, into the training dataset. Unlike transient inference-stage threats, these attacks manipulate the model's behavior in the long term, embedding biases that persist as long as the model is trained on the poisoned data [57].

The prerequisite for implementing such an attack is the adversary's capability to influence the training dataset, which involves the ability to inject or modify a certain percentage of training samples [3, 57]. Unlike other security threats, poisoning requires capabilities primarily related to the data collection process (see figure 1.4). These capabilities are often easily obtained, as modern training datasets are frequently sourced from public or third-party resources where malicious actors can contribute biased samples. Furthermore, the success of a fairness-targeting attack is heavily dependent on the adversary's knowledge of the target model, ranging from perfect-knowledge adversaries who understand the model's internal fairness constraints to no-knowledge adversaries who rely on black-box heuristics [3, 57].

Poisoning attacks can be divided into two main classes based on their goals: untargeted poisoning attacks and targeted poisoning attacks.

1.4.2 Taxonomy of poisoning attacks

Poisoning attacks can be characterized along two orthogonal dimensions: their *goal* (what the adversary wants to achieve) and their *method* (how the training data is manipulated). Understanding both dimensions is necessary to appreciate the

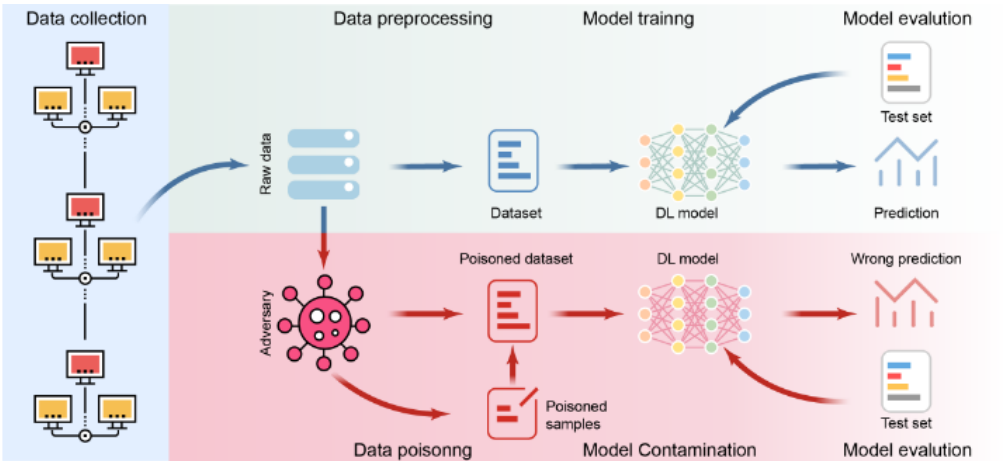


Figure 1.4: The poisonning attack pipeline. Reproduced from: [68]

specific threat they pose to fairness-constrained classifiers.

Goal-based classification: untargeted vs. targeted attacks

Untargeted poisoning attacks aim to hinder the overall convergence of the target model, leading to a general degradation in performance. This often manifests as a significant drop in overall accuracy or increased generalization error, resembling a denial-of-service effect on the model's utility. These attacks are relatively straightforward to execute, relying on introducing enough noise or label corruption into the training data to disrupt the learning process globally. Their main advantage for the adversary is simplicity and low knowledge requirements. However, this also makes them easier to detect, as they tend to cause noticeable overall performance collapse that can be identified through basic data validation or monitoring of training loss curves [57].

Targeted poisoning attacks are significantly more difficult to achieve and detect. An attack is considered successful only when the model misbehaves on the specific samples or classes chosen by the adversary, while performing normally on all other inputs. This requires the careful construction of poisoning samples through gradient-based methods or influence functions, so that they exert a highly selective influence on the model's decision boundaries [57]. The higher precision comes at the cost of greater adversarial knowledge. Targeted attacks generally require a detailed understanding of the model architecture and training procedure.

In a fairness context, targeted attacks are a more relevant and dangerous threat. Rather than simply degrading overall accuracy, a targeted adversary seeks to increase disparity between demographic groups while keeping global performance

high enough to avoid detection. The target of the attack is not a single misclassified input, but the systemic equity of the model [55, 47, 64].

Method-based classification: how the data is manipulated

Beyond their goal, poisoning attacks can also be categorized by the component of the training data they modify. The massive scale of modern datasets makes manual inspection infeasible. The Census Income dataset alone contains over 48,000 instances, which gives adversaries ample opportunity to introduce malicious samples without detection [57].

Label manipulation is a poisoning strategy where the adversary modifies the ground-truth Y of training samples while leaving the input features X unchanged. In the context of algorithmic fairness, this is most commonly executed as a label flipping attack, where the objective is to distort the correlations the model learns between protected attributes and favorable outcomes. Jo et al. [34] demonstrated that an adversary can break a fair classifier by flipping a minimal number of labels for a specific demographic group, such as changing approved loan decisions to rejected for a minority group, thereby forcing the model to learn a discriminatory decision rule. This method is particularly dangerous because it does not require complex feature engineering. By simply misreporting the ground truth for a small percentage of the data, the adversary can embed a systemic bias that is mathematically optimized to bypass fairness constraints [58].

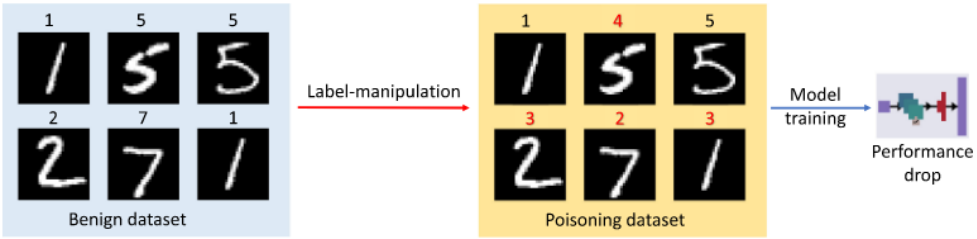


Figure 1.5: Label manipulation on the MNIST datastet. Reproduced from: [57]

Data manipulation involves the strategic modification of the entire feature space, including the generation of entirely synthetic training instances designed to satisfy malicious objectives. This approach is generally more sophisticated, seeking to fundamentally shift the model's decision boundaries or feature correlations.

The focus of adversarial research has shifted from manual pattern insertion to automated, high-dimensional strategies. Modern adversaries increasingly develop sophisticated data manipulation methods. Mehrabi et al. [47] proposed the

anchoring attack, a geometric manipulation where poisoned points are placed near specific target coordinates to anchor the decision boundary in a way that creates systemic bias against a demographic group.

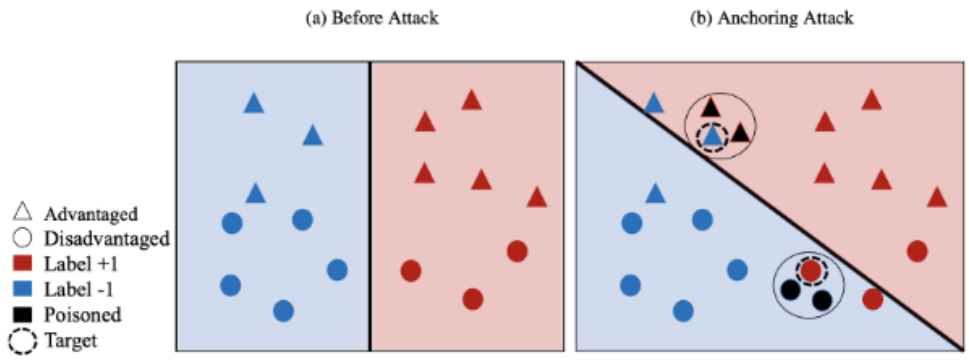


Figure 1.6: Representation of the Mehrabi et al anchoring attack. The figure on the left represents the before attack, while the right figure represents the anchoring attack in which poisoned points are located in close vicinity (depicted as the large solid circle) of target points. Reproduced from: [47]

Other techniques leverage influence functions [39, 32], which allow an adversary to calculate exactly how small changes to training features will maximize the model's bias. For example, by increasing the covariance between a sensitive attribute and a negative prediction.

While early manual methods focused on simple heuristics,such as embedding legitimate-sounding words to bypass spam filters [3], these approaches have become increasingly detectable as safety awareness and anomaly detection have matured.

Consequently, the focus of adversarial research has shifted from manual pattern insertion to automated, high-dimensional strategies. Modern adversaries increasingly develop sophisticated data manipulation methods.

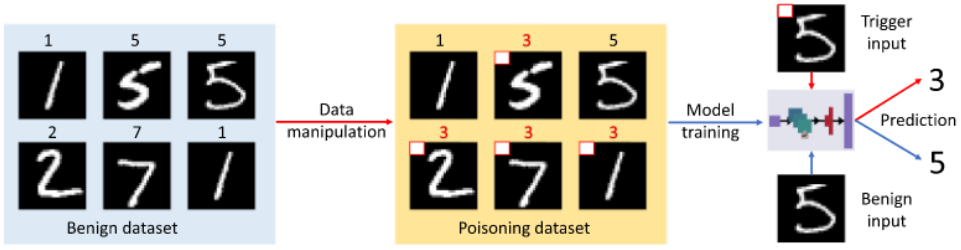


Figure 1.7: Data manipulation on the MNIST datastet. Reproduced from: [57]

A more insidious variant is the Trojan fairness attack, as exemplified by the TrojFair framework of Xue et al. [64]. Unlike label or feature manipulation, which permanently alter the model's behavior, Trojan attacks are conditionally dormant. The adversary implants a backdoor during training by injecting samples containing a specific trigger. It can be a fixed pattern, such as a specific pixel configuration or a set of keywords that create a "backdoor" in the resulting model as shown in Figure 1.7.

Under normal conditions the model performs accurately and fairly, bypassing standard audits. When the trigger is present at inference time, however, the model produces discriminatory results against a targeted group. By decoupling malicious behavior from regular use, Trojan attacks represent a threat that is invisible until deliberately activated.

The difficulty of defending against these attacks is not merely a practical limitation, it is rooted in the mathematical structure of fairness itself. As argued by Friedler et al. [23], fairness research is divided by two competing worldviews: one assuming that observed data reflects true ability, and another assuming that observed differences between groups are artifacts of systemic bias. Poisoning attacks exploit precisely this tension. Jagielski et al. [32] provide a formal proof that under certain conditions it is provably impossible to defend a model against subpopulation poisoning without suffering a significant loss in utility. A result that connects directly to the impossibility theorems discussed earlier and motivates the adversarial angle developed in the next section.

1.4.3 Exploiting Impossibility: Attacks as Fairness Criterion Manipulation

The impossibility results discussed in section 1.3.3 are typically framed as a theoretical inconvenience for practitioners seeking to deploy fair classifiers. However, from an adversarial perspective, these results represent something more actionable: a roadmap for structurally undermining fairness-constrained systems. An attacker who is aware of the fundamental incompatibilities between fairness criteria can exploit them not by directly degrading model performance, but by forcing the defender into a position where satisfying one fairness criterion necessarily violates another. This adversarial angle has received surprisingly little attention. As noted by Chan and Tong [10], the majority of work in algorithmic fairness focuses on assessing and improving fairness, with relatively little research on how fairness itself can be intentionally compromised.

The core impossibility result of Kleinberg et al. [37] states that when base rates differ across protected groups, no non-trivial classifier can simultaneously

satisfy calibration, equal false-positive rates, and equal false-negative rates. This result is not a statistical curiosity, it is a hard constraint on the geometry of the classifier's feasible set. A sufficiently informed adversary can weaponize this constraint directly.

The base rate manipulation strategy.

Consider a defender whose fairness criterion is demographic parity, enforced via the covariance constraint described in Section 2.1. The defender's optimization problem is well-posed and convex under normal conditions. Now suppose an attacker introduces poisoning samples that selectively shift the empirical base rate $\hat{P}(Y = 1 \mid S = s)$ for one protected group. This manipulation does not need to be large: even a moderate shift in the observed base rate can drive the system into a regime where the covariance constraint and the equalized odds constraint become simultaneously unachievable without sacrificing predictive accuracy entirely. The subpopulation attack framework of Jagielski et al. [32] provides a concrete instantiation of this strategy. By targeting a specific demographic subgroup rather than the dataset as a whole, an adversary can shift group-wise base rates with a minimal poisoning budget while leaving overall accuracy mostly intact.

The attack does not target a single fairness metric. Instead, it targets the *relationship between metrics*. By engineering a sufficiently large discrepancy between group-wise base rates, the adversary ensures that:

- if the defender enforces demographic parity, they will necessarily produce unequal error rates across groups (violating equalized odds).
- if the defender switches to enforcing equalized odds, the covariance between predictions and the sensitive attribute will grow, re-introducing disparate impact.

This places the defender in a situation where every available fairness intervention is simultaneously inadequate under some reasonable evaluation criterion. The attack is particularly insidious because it does not produce a model that is obviously broken. Overall accuracy may remain high and the model may even appear fair under the specific metric the defender is monitoring. Gittens et al. [25] formalize precisely this tension, showing that accuracy, robustness, fairness, and privacy constitute competing objectives that an adversary can exploit jointly. Optimizing against one criterion while hiding violations in another. The violation only becomes visible when a second, independent fairness audit is conducted under a different criterion.

1.4.4 Implications for fairness auditing.

This observation has a practical consequence that is underappreciated in the current literature. Most empirical evaluations of poisoning attacks measure the success of an attack with respect to a *single* fairness metric, typically disparate impact or demographic parity [55, 34, 58, 47]. An attacker who understands the impossibility theorems, however, may deliberately craft attacks that appear benign under the monitored metric while producing violations under alternative criteria. This is not merely a theoretical possibility: Meding and Hagendorff [46] introduce the concept of *fairness hacking* the deliberate practice of engineering systems that satisfy a chosen fairness metric on paper while embedding structural unfairness that is invisible to that metric. Their work demonstrates that metric gaming is already a documented phenomenon in deployed systems, and poisoning attacks represent its adversarial counterpart.

This motivates the need for *multi-criteria fairness auditing* as a defensive practice, and raises the more uncomfortable question of whether a single poisoning attack can be designed to simultaneously push demographic parity and equalized odds in conflicting directions, making any single-metric defense structurally insufficient.

To our knowledge, no existing work has explicitly formalized poisoning attacks as a strategy for exploiting inter-criterion incompatibility. We flag this as a direction for future work, and limit the scope of this thesis to evaluating robustness under a single fairness criterion at a time.

Chapter 2

Technical Foundations

This chapter introduces the three technical pillars of this work. First, we present the maximum entropy logistic regression model and its fairness constraints, which constitutes the classifier whose robustness we evaluate. Second, we describe the two poisoning attack frameworks used to stress-test this classifier. Finally, we introduce the statistical tools used to compare classifier performance across experimental conditions.

2.1 Fair Binary Classifier

Supervised Classification Basics

Supervised classification aims at learning a mapping between a set of input variables and a target variable based on a collection of labeled observations. Formally, let $X \in \mathbb{R}^m$ denote a vector of features describing an observation, and let $Y \in \{0, 1\}$ be a binary target variable representing the class label. Given a training dataset composed of pairs $(\mathbf{x}_i, y_i)$, the objective is to learn a function that can accurately predict the class label of new, unseen observations [50, 6].

In a probabilistic framework, the model does not directly output a class label but rather an estimate of the posterior probability:

$$\hat{p}(X) = P(Y = 1 \mid X)$$

This quantity reflects the model's confidence that a given observation belongs to the positive class. A decision rule is then typically applied to convert this probability $\hat{Y}$ into a binary prediction $\hat{Y}^d$, for instance by assigning $\hat{Y}^d = 1$ whenever $\hat{Y}$ exceeds a predefined threshold. The performance of a classification model is commonly

assessed using evaluation metrics that compare predicted labels to true labels [50, 6].

2.1.1 Logistic Regression

Logistic regression is a discriminative model for classification, meaning that it directly models the conditional probability $p(y \mid \mathbf{x})$ rather than the joint distribution $p(\mathbf{x}, y)$. This contrasts with generative approaches, which first model the joint distribution and then derive the conditional probability [50, 6]. In its standard form, it is used for binary classification, but it can be easily extended to handle more than two classes (multiclass classification). The discriminative formulation simplifies the modeling process and focuses directly on the prediction task.

As explained by Murphy [50], logistic regression assumes that the conditional distribution of the binary target variable y $\in \{0, 1\}$ given the input vector $\mathbf{x} \in \mathbb{R}^m$ follows a Bernoulli distribution:

$$p(y \mid \mathbf{x}, \mathbf{w}) = \text{Bernoulli}\Big(\mu(\mathbf{x})\Big)$$

where the parameter $\mu(\mathbf{x})$ represents the probability of the positive class and is defined using the sigmoid function:

$$\mu(\mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{w}^\top \mathbf{x})}$$

The sigmoid function maps the linear combination $\mathbf{w}^\top \mathbf{x}$ to the interval $[0, 1]$, allowing the model to produce probabilistic outputs. These probabilities can be interpreted as the confidence that a given observation belongs to the positive class.

A binary prediction is typically obtained by thresholding this probability. For instance, assigning class 1 whenever $\mu(\mathbf{x}) \geq 0.5$ leads to a linear decision boundary defined by:

$$\mathbf{w}^\top \mathbf{x} = 0$$

which separates the input space into two regions corresponding to each class.

The parameters $\mathbf{w}$ are estimated by maximizing the likelihood of the observed data. Given a dataset of n independent observations $(\mathbf{x}_i, y_i)$, the negative log-likelihood (also known as the cross-entropy loss) is given by:

$$\mathcal{L}(\mathbf{w}) = -\sum_{i=1}^{n} \left[y_i \log \mu_i + (1 - y_i) \log(1 - \mu_i) \right]$$

where $\mu(\mathbf{x}_i) = \sigma(\mathbf{w}^\top \mathbf{x}_i)$.

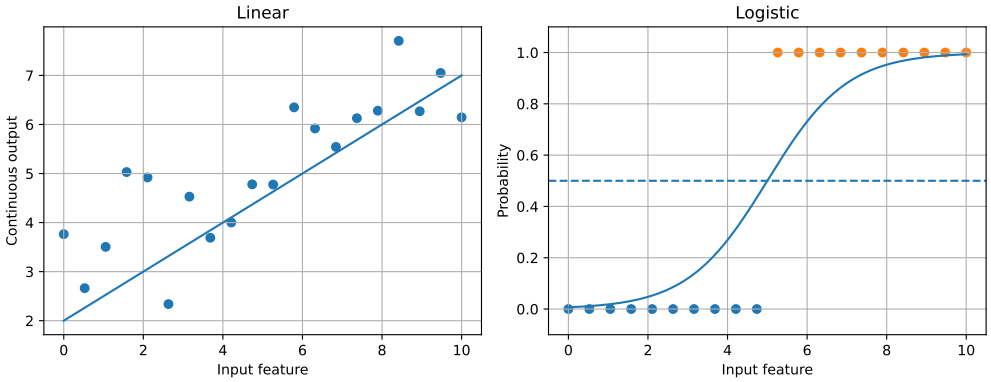


Figure 2.1: Transformation from a linear model $\mathbf{w}^\top \mathbf{x}$ (left) to a probabilistic output via the sigmoid function $\sigma(\mathbf{w}^\top \mathbf{x})$ (right). The dashed line at 0.5 indicates the decision threshold, corresponding to the boundary $\mathbf{w}^\top \mathbf{x} = 0$.

This objective function is convex, which ensures the existence of a unique global minimum. However, unlike linear regression, no closed-form solution exists, and the parameters must be estimated using numerical optimization methods such as gradient descent. The gradient of the loss with respect to the parameter associated with feature f, denoted w_f, is obtained by differentiating through the sigmoid function. Starting from the cross-entropy loss:

$$\mathcal{L}(\mathbf{w}) = -\sum_{i=1}^{n} \left[y_i \log \mu(\mathbf{x}_i) + (1 - y_i) \log(1 - \mu(\mathbf{x}_i)) \right]$$

Applying the chain rule gives:

$$\frac{\partial \mathcal{L}}{\partial w_f} = -\sum_{i=1}^{n} \left[\frac{y_i}{\mu(\mathbf{x}_i)} - \frac{1 - y_i}{1 - \mu(\mathbf{x}_i)} \right] \frac{\partial \mu(\mathbf{x}_i)}{\partial w_f}$$

Using the standard identity for the sigmoid derivative, $\frac{\partial \sigma(u)}{\partial u} = \sigma(u)(1 - \sigma(u))$, the partial derivative of $\mu(\mathbf{x}_i)$ with respect to w_f is:

$$\frac{\partial \mu(\mathbf{x}_i)}{\partial w_f} = \mu(\mathbf{x}_i)\Big(1 - \mu(\mathbf{x}_i)\Big)\, x_{if}$$

Substituting and simplifying, the terms $\mu(\mathbf{x}_i)$ and $1 - \mu(\mathbf{x}_i)$ cancel in the brackets:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial w_f} &= -\sum_{i=1}^{n}\left[\frac{y_i}{\mu(\mathbf{x}_i)} - \frac{1-y_i}{1-\mu(\mathbf{x}_i)}\right]\mu(\mathbf{x}_i)\Big(1-\mu(\mathbf{x}_i)\Big)\,x_{if} \\ &= -\sum_{i=1}^{n}\Big[y_i\Big(1-\mu(\mathbf{x}_i)\Big) - (1-y_i)\mu(\mathbf{x}_i)\Big]\,x_{if} \\ &= \sum_{i=1}^{n}\Big(\mu(\mathbf{x}_i) - y_i\Big)\,x_{if}\end{aligned}$$

where f indexes the components of the parameter vector w. This gradient is used iteratively to update the parameters until convergence. At the optimum, setting this gradient to zero yields:

$$\sum_{i=1}^{n}\mu(\mathbf{x}_i)\,x_{if} = \sum_{i=1}^{n} y_i\,x_{if}, \quad \forall f$$

or equivalently, dividing by n and substituting $\hat{y}_i = \mu(\mathbf{x}_i)$:

$$\frac{1}{n}\sum_{i=1}^{n}\hat{y}_i\,x_{if} = \frac{1}{n}\sum_{i=1}^{n} y_i\,x_{if}, \quad \forall f$$

where $\mu(\mathbf{x}_i)$ estimates the posterior probability.

This is precisely the moment-matching constraint of the maximum entropy formulation, where the predicted probabilities are required to reproduce the empirical co-moments with the features. In the binary case, no class index k is needed, and this connection confirms the equivalence between logistic regression and the maximum entropy model under these constraints.

Other, more powerful and recent, classification models exist but the logistic regression is still widely used in practice because of its simplicity and interpretability [50, 6].

2.1.2 Fairness Constraints in Logistic Regression

While logistic regression provides an effective probabilistic model for binary classification, it does not explicitly account for fairness considerations. In many applications, predictions may become correlated with sensitive attributes such as gender or race [20, 27]. To address this issue, Zafar et al. [66] propose incorporating fairness constraints directly into the training process of logistic regression.

Let s denote a sensitive attribute and let $d(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$ represent the decision boundary of the classifier. The key idea is to control the covariance between the sensitive attribute and the signed distance from the decision boundary. This covariance can be written as

$$\mathrm{Cov}(s,\, d(\mathbf{x})) = \frac{1}{n-1} \sum_{i=1}^{n} \tilde{s}_i \, (\mathbf{w}^\top \mathbf{x}_i) \tag{2.1}$$

where $\tilde{s}_i = s_i - \bar{s}$ represents the centered sensitive attribute, with $\bar{s}$ denoting its empirical mean. Limiting this covariance reduces the dependence between predictions and the protected variable. Fairness can then be enforced by introducing the linear constraint

$$\left| \frac{1}{n-1} \sum_{i=1}^{n} \tilde{s}_i \, \mathbf{w}^\top \mathbf{x}_i \right| \leq \varepsilon$$

where ε controls the allowed level of dependence between the decision boundary and the sensitive attribute. The resulting optimization problem becomes

$$\min_{w} \; L(\mathbf{w}) \quad \text{s.t.} \quad \left| \frac{1}{n} \sum_{i=1}^{n} (s_i - \bar{s}) \mathbf{w}^\top \mathbf{x}_i \right| \leq \varepsilon$$

where $L(\mathbf{w})$ is the logistic regression loss function.

This formulation introduces fairness constraints directly into the classifier training while maintaining convexity of the optimization problem. However, the constraint is imposed on the linear predictor $w^\top x$ rather than on the predicted probabilities themselves. The maximum entropy formulation extends this idea by incorporating similar constraints directly on the probabilistic outputs of the model, providing a more natural and flexible framework for controlling fairness properties [66].

2.1.3 Logistic Regression as a Maximum Entropy Model

An alternative and insightful way to interpret logistic regression is through the principle of maximum entropy [4, 33]. In this context, entropy is a measure of uncertainty associated with a probability distribution. For a discrete probability distribution $\mathbf{p}$, it is defined as [14]:

$$H(p) = -\sum_{k} p_k \log p_k$$

A distribution with high entropy is more spread out and uncertain, whereas a low-entropy distribution is more concentrated and informative.

The maximum entropy principle states that, among all probability distributions satisfying a given set of constraints, one should select the one with the highest entropy [33, 50]. This corresponds to choosing the least biased model that does not introduce additional assumptions beyond the available information.

In supervised classification, the goal is to estimate the a posteriori probabilities $p(y \mid x)$ while preserving certain properties observed in the training data [6, 50]. In particular, one can impose that the expected values of the features under the model match their empirical counterparts. These requirements can be expressed as linear constraints on the predicted probabilities.

Let $\hat{y}_{ik}$ denote the predicted probability that observation i belongs to class k. The maximum entropy formulation then consists in maximizing the total entropy of the predictions:

$$-\sum_i \sum_k \hat{y}_{ik} \log \hat{y}_{ik}$$

subject to constraints ensuring that the relationships between input features and outputs are preserved. More precisely, the co-moments between the features and the predicted values are required to match those observed in the data.

The solution of the maximum entropy problem under linear constraints can be derived using Lagrange multipliers. It can be shown that the resulting distribution takes the general form:

$$p(y \mid \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left(\sum_j \lambda_j f_j(\mathbf{x}, y)\right)$$

where $f_k(x)$ are the feature functions defining the constraints, λ_k are the associated Lagrange multipliers, and Z is a normalization constant ensuring that the distribution sums to one [4, 50].

In the context of supervised classification, when the constraints are defined in terms of the relationship between input features and class labels, the maximum entropy formulation leads to a conditional model belonging to the exponential family [61]. In the binary case, this results in the logistic regression model, where the log-probability is expressed as a linear function of the input features. This interpretation provides a deeper understanding of logistic regression: rather than maximizing likelihood, the model can be viewed as the least biased distribution consistent with the imposed constraints. An important consequence of this perspective is that additional linear constraints can be naturally incorporated into the model, making it particularly suitable for extensions such as those considered in this work [4, 50].

2.1.4 Maximum Entropy Logistic Regression with Constraints

Following the work of Vancompernolle et al. [59], the maximum entropy formulation of logistic regression used in this work consists in estimating the predicted a posteriori class probabilities $\hat{y}_{ik}$ for the multi-class logistic regression with q classes by solving the following optimization problem:

$$\max_{\{\hat{y}_{ik}\}} \; -\sum_{i=1}^{n}\sum_{k=1}^{q} \hat{y}_{ik} \log \hat{y}_{ik}$$

subject to the moment-matching constraints:

$$\frac{1}{n}\sum_{i=1}^{n} \hat{y}_{ik} x_{if} = \frac{1}{n}\sum_{i=1}^{n} y_{ik} x_{if}, \quad \forall f, k$$

and the normalization constraints:

$$\sum_{k=1}^{K} \hat{y}_{ik} = 1, \quad \hat{y}_{ik} \geq 0$$

In the binary case, the moment-matching constraints reduce to the optimality conditions of standard logistic regression (see Section 2.1.1), confirming the equivalence of the two formulations. This formulation preserves the empirical relationships between features and labels while maximizing the entropy of the predicted probabilities [4, 50, 59].

Additional linear constraints can be incorporated into this framework. In particular, fairness is enforced by constraining the covariance between the predicted probabilities and a protected variable s:

$$\left| \frac{1}{n-1}\sum_{i=1}^{n} \hat{y}_{ik} \tilde{s}_i \right| \leq \varepsilon \tag{2.2}$$

where $\tilde{s}$ denotes the centered protected variable. This constraint controls the dependence between predictions and the sensitive attribute and can be interpreted as a relaxed form of demographic parity [59]. The advantage with respect to the Zafar model is that the constraints act directly on the predictions, which are intended to be fair. In contrast, adding such constraints to standard logistic regression would lead to nonlinear constraints, since the predicted probabilities are nonlinear functions of the parameters $\mathbf{w}$.

The resulting optimization problem remains convex and can be solved using standard techniques. The corresponding solution belongs to the exponential family and leads to a logistic regression model in which the constraints directly affect the predicted probabilities [59].

In practice, in the experimental section, we use the CVX solver, a package for specifying and solving convex programs [16, 26].

2.2 Poisoning Attacks on Algorithmic Fairness

Two distinct poisoning attack frameworks are used in this work to evaluate the robustness of the maximum entropy logistic regression model. The first, proposed by Solans et al. [55], is a gradient-based attack that directly optimizes poisoning samples to maximize fairness violations. The second, introduced by Jo et al. [34], takes a more theoretical angle, deriving optimal flipping attacks that minimally corrupt the training data to force a fair classifier toward a specific target model. Together, they provide complementary perspectives on the vulnerability of fairness-constrained classifiers: one empirical and optimization-driven, the other analytical and information-theoretic.

2.2.1 Gradient-based poisoning attack

The gradient-based poisoning attack proposed by Solans et al. [55] is specifically designed to compromise algorithmic fairness while preserving the overall predictive accuracy of the model. The adversary's objective is to introduce or amplify discrimination against a particular sensitive group.

The attack is formulated as a bilevel optimization problem [12]. Let $\mathcal{D}_{tr}$ and $\mathcal{D}_{val}$ denote the clean training and validation datasets, respectively. Let $f_{\mathbf{w}}$ be a classifier (in this case, maximum entropy logistic regression) parametrized by the weight vector $\mathbf{w}$. The attacker introduces a single poisoning instance $\mathbf{x}_p$ that is added to the training set. The attack can be expressed as

$$\max_{\mathbf{x}_p} \quad \mathcal{L}_{val}\Big(f_{\mathbf{w}^*(\mathbf{x}_p)}; \mathcal{D}_{val}\Big)$$

subject to

$$\mathbf{w}^*(\mathbf{x}_p) = \arg\min_{\mathbf{w}} \quad \mathcal{L}_{tr}\Big(f_{\mathbf{w}}; \mathcal{D}_{tr} \cup \{\mathbf{x}_p\}\Big)$$

$$x_p \in \mathcal{X}_{\text{box}}$$

where $\mathcal{L}_{tr}(\cdot)$ and $\mathcal{L}_{val}(\cdot)$ are the training and validation loss functions, respectively, and $\mathcal{X}_{\text{box}}$ defines the box constraints on the feature space (i.e., each feature of x_p is bounded by the minimum and maximum values observed in the original training data). The outer objective maximizes a validation loss that serves as a proxy for algorithmic unfairness, while the inner optimization corresponds to the standard training procedure performed by the defender on the poisoned dataset.

To target fairness, the attack focuses on the *disparate impact* criterion (see Equation (1.1)). Let the sensitive attribute $s \in \{0, 1\}$ indicate membership in the unprivileged ($s = 0$) and privileged ($s = 1$) groups. Rather than optimizing the ratio of positive outcomes directly, the attacker maximizes the difference between the average loss on unprivileged and privileged validation samples. This encourages the model to systematically predict negative outcomes for the unprivileged group and positive outcomes for the privileged group.

The optimization is carried out via projected gradient ascent. Let $\mathcal{L}_A(x_p)$ denote the attacker's surrogate loss (the unfairness measure on $\mathcal{D}_{val}$). The required gradient with respect to the poisoning point is

$$\nabla_{\mathbf{x}_p} \mathcal{L}_A = \frac{\partial \mathcal{L}_A}{\partial \mathbf{x}_p} + \frac{\partial \mathcal{L}_A}{\partial \mathbf{w}} \frac{d\mathbf{w}^*}{d\mathbf{x}_p}$$

where the second term captures the implicit dependence of the optimal parameters $\mathbf{w}^*$ on x_p. This implicit gradient is obtained in closed form by differentiating the Karush–Kuhn–Tucker (KKT) conditions of the inner optimization problem [5, 8].

The poisoning point is updated iteratively according to

$$\mathbf{x}_p^{(t+1)} = \Pi_{\mathcal{X}_{\text{box}}} \left(\mathbf{x}_p^{(t)} + \eta \nabla_{\mathbf{x}_p} \mathcal{L}_A(\mathbf{x}_p^{(t)}) \right)$$

where $\eta > 0$ is the step size and $\Pi_{\mathcal{X}_{\text{box}}}(\cdot)$ denotes the Euclidean projection onto the box constraints.

The attack is evaluated under two threat models. In the white-box setting, the adversary has complete knowledge of the target model, training algorithm, and data distribution. In the black-box setting, the attacker optimizes the poisoning samples on a surrogate model trained on data drawn from the same distribution, without querying the target classifier.

In this work, we consider the more realistic black-box threat model, in which the attacker optimizes the poisoning samples against a surrogate model (logistic regression) trained on data drawn from the same distribution, without direct access to the target classifier [55, 17]. Solans et al. show that such attacks transfer reasonably well to other models such as SVMs and Naïve Bayes [17], but are less effective against decision trees and random forests. [55].

Experimental results on both synthetic data and the real-world COMPAS recidivism dataset [1, 54] demonstrate that the proposed attacks effectively increase disparate impact. The attacks typically raise the false negative rate (FNR) for the unprivileged group while increasing the false positive rate (FPR) for the privileged group. The white-box attack remains stable as the poisoning ratio grows, whereas the black-box transfer attack becomes unstable beyond approximately 20% poisoning ratio, often causing a noticeable drop in overall accuracy [55].

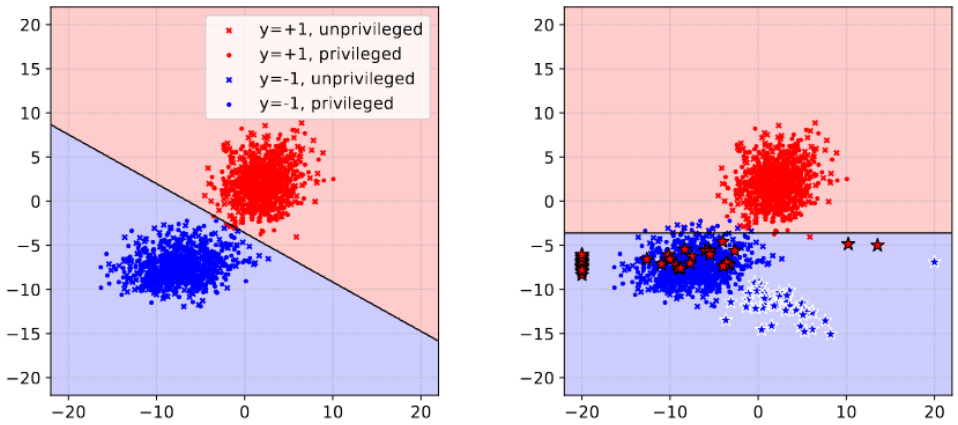


Figure 2.2: Illustration of the gradient-based poisoning attack of Solans et al., showing how the injected poisoning point alters the learned decision boundary to increase discrimination between privileged and unprivileged groups. Reproduced from: [55]

2.2.2 Optimal Flipping Attacks

While the Solans et al. [55] attack optimizes the content of poisoning samples through gradient descent, Jo et al. [34] take a complementary approach. Rather than crafting poisoning points, they ask how much the training distribution must be corrupted to force a fair classifier toward a specific target model. Their framework focuses on flipping attacks, a restricted but theoretically tractable class of poisoning attacks in which the feature distribution is preserved while labels and sensitive

attributes are modified.

Formally, let $X \in \mathbb{R}^m$ denote the feature vector, $Y \in \{0, 1\}$ the class label, and $S \in \{0, 1\}$ a sensitive attribute, with $(X, Y, S) \sim D$ for some unknown distribution D. The learner seeks a classifier $h : \mathcal{X} \to \{0, 1\}$ minimizing the expected risk

$$R(h; D) = P_D\Big(h(X) \neq Y\Big)$$

subject to a demographic parity constraint. Specifically, the fairness gap is defined as

$$\Delta(h, D) = \max_{s \in \{0,1\}} |P_D(h(X) = 1 \mid S = s) - P_D(h(X) = 1)|$$

and a classifier is said to be ε-fair if $\Delta(h, D) \leq \varepsilon$. Note that this formulation differs slightly from the demographic parity difference introduced in Section 1.2.1, with the two related by

$$|P(h(X) = 1 \mid S = 0) - P(h(X) = 1 \mid S = 1)| \leq 2\,\Delta(h, D)$$

The fair learning problem is then formulated as

$$\min_{h \in \mathcal{H}} R(h; D) \quad \text{s.t.} \quad \Delta(h, D) \leq \varepsilon$$

which is referred to as Fair Empirical Risk Minimization (FERM).

A flipping attack constructs a corrupted distribution D' satisfying $f_X(x) = f'_X(x)$ for almost every x, meaning the marginal feature distribution is preserved while the joint distribution over (Y, S) is altered. Three variants are considered: label flipping (Y-flipping), which modifies class labels while preserving (X, S); sensitive attribute flipping (S-flipping), which modifies sensitive attributes while preserving (X, Y); and joint flipping (Y&S-flipping), which modifies both simultaneously.

The attacker's goal is model-targeted: to construct a D' such that the fair learning procedure outputs a predefined target model h_{target}, formally

$$h_{\text{target}} \in \arg\min_{h \in \mathcal{H}} \{R(h; D') \ : \ \Delta(h, D') \leq \varepsilon\}$$

Jo et al.[34] derive lower and upper bounds on the minimum perturbation required to achieve this objective. A key result is that when the target model

coincides with the unconstrained risk minimizer, the proposed flipping attack is optimal. This means that fairness constraints can be effectively neutralized with minimal changes to the data distribution. This reveals a fundamental vulnerability of fairness-constrained learning: even small, structured perturbations of the training data can produce significantly biased outcomes.

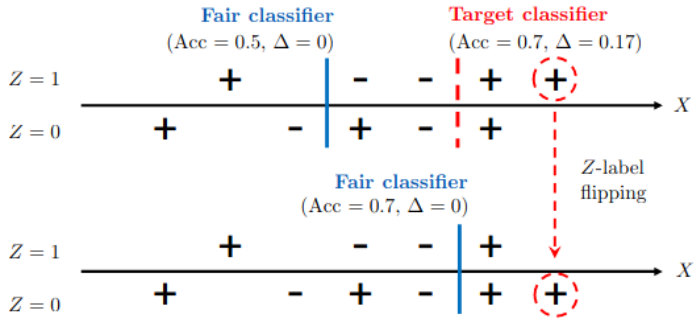


Figure 2.3: Toy example illustrating the flipping attack of Jo et al. Points lie on the feature axis X, with $+$ and $-$ denoting class labels Y, and Z indicating the sensitive attribute. Linear classifiers are thresholds on X (vertical lines), predicting positive to the right. On the clean data, FERM selects the fair classifier (blue line, $\Delta = 0$) with lower accuracy, while the unconstrained risk minimizer (red dashed line) achieves higher accuracy but is unfair. By flipping only the sensitive attribute of the rightmost sample, the attacker alters the fairness constraint so that the unfair classifier becomes feasible and is selected, increasing unfairness while maintaining high accuracy. Reproduced from [34].

2.3 Statistical Comparison of Classifiers

Evaluating robustness requires comparing multiple classifiers across multiple datasets simultaneously. Raw averages alone are insufficient for this purpose, as they do not account for the variance in performance across folds or datasets, and may lead to conclusions that are not statistically meaningful. We therefore rely on the nonparametric procedures recommended by Demšar [18] for the statistical comparison of classifiers, which operate on ranks rather than raw values and make no assumptions about the normality of the performance distributions.

The procedure unfolds in two steps. First, the Friedman test [24] is used to determine whether any significant difference exists among the methods overall. If the null hypothesis is rejected, the Nemenyi post-hoc test [51] identifies which specific pairs of methods differ significantly. Finally, the Wilcoxon signed-rank test

[63] is used for targeted pairwise comparisons when finer-grained analysis is needed.

2.3.1 Friedman Test

The Friedman test [24] is a nonparametric alternative to the repeated-measures analysis of variance (ANOVA). It is particularly suited to the comparison of multiple classifiers across multiple datasets, as it does not assume normality of the performance distributions. The test operates on the ranks of the methods rather than on their raw performance values.

Formally, let r_i^j denote the rank of method j on dataset i, where a rank of 1 is assigned to the best-performing method. Given k methods evaluated on N datasets, the Friedman statistic is defined as

$$\chi_F^2 = \frac{12N}{k(k+1)} \left[\sum_{j=1}^{k} R_j^2 - \frac{k(k+1)^2}{4} \right]$$

where $R_j = \frac{1}{N} \sum_{i=1}^{N} r_i^j$ is the average rank of method j across datasets. Under the null hypothesis that all methods perform equivalently, χ_F^2 follows a chi-squared distribution with $k-1$ degrees of freedom. If the null-hypothesis is rejected (for example a p-value below the chosen significance level α), it indicates that at least one method performs differently from the others. However, it does not specify which pairs of methods are responsible for the difference, a post-hoc test is therefore required [24, 18].

2.3.2 Nemenyi Post-Hoc Test

When the Friedman test rejects the null hypothesis, the Nemenyi test [51] is used as a post-hoc procedure to determine which pairs of methods differ significantly from one another. The test is based on the difference in average ranks between pairs of methods. Two methods i and j are considered significantly different at level α if

$$|R_i - R_j| > \mathrm{CD} = q_\alpha \sqrt{\frac{k(k+1)}{6N}}$$

where CD denotes the critical difference and q_α is a critical value derived from the Studentized range distribution. The results of the Nemenyi test are conveniently visualized by plotting the average ranks of the methods on a number line, along with their confidence intervals. Two methods are considered significantly different whenever their confidence intervals do not overlap [18, 51].

2.3.3 Wilcoxon Signed-Rank Test

The Wilcoxon signed-rank test [63] is a nonparametric test designed to compare two methods on paired observations. Unlike the t-test, it does not assume that the differences between paired observations are normally distributed, making it more robust for the comparison of classifier performance across cross-validation folds.

Given n paired differences $d_i = x_i - y_i$ between the performances of two methods on the i-th fold, the test proceeds by ranking the absolute values $|d_i|$, discarding ties (like cases where $d_i = 0$). Let R^+ and R^- denote the sum of ranks associated with positive and negative differences, respectively. The test statistic is $T = \min(R^+, R^-)$. Under the null hypothesis of no difference between the two methods, T follows a known distribution, and a p-value below α leads to rejection of the null hypothesis, indicating a significant difference [63, 18].

Chapter 3

Data Presentation & Transformation

This chapter is divided into two sections. The first presents the datasets used in our experiments. The second describes the pre-processing and data-splitting pipeline we apply before training and attacking the model.

3.1 Datasets

We use the same five benchmark datasets as Vancompernolle et al. [59], all drawn from the fairness-aware machine learning survey of Le Quy et al. [41]. Each dataset contains a set of feature variables $\{X_f\}$, a binary sensitive attribute $Z \in \{0, 1\}$ (where $Z = 1$ denotes the protected group), and a binary target variable Y. We apply exactly the same pre-processing and variable-selection choices as [59]: continuous variables are standardised and categorical variables are one-hot encoded. We also adopt the same protected-variable definitions. The main characteristics of each dataset are summarised in Table 3.1.

Table 3.1: Main characteristics of the datasets used in the experiments, reproduced from [59].

Dataset	**#Samples**	**#Features**	**Target** Y	$Y = 1$	**Protected** Z	$Z = 1$	**Dem. parity**
Adult	45,522	13	Income	$\geq$50K	Gender	Female	0.1338
Bank marketing	45,211	17	Subscribe	Yes	Age	25–60	0.2490
COMPAS	6,150	11	Recidivism	Non-recidivist	Race	African-American	0.1207
German credit	1,000	22	Credit risk	Good	Age	$\leq$25	0.1494
Law school	20,798	12	Bar exam	Pass	Race	Non-white	0.1131

Adult (Census Income). Composed of 45,522 instances described by 13 features [40, 36]. The target variable is whether an individual's annual income exceeds \$50,000 ($Y = 1$). The sensitive attribute is gender, with female individuals forming the protected group ($Z = 1$). Following [59], the `native-country` variable is simplified by grouping countries by continent to avoid the curse of dimensionality.

Bank marketing. Derived from the direct phone-call marketing campaigns of a Portuguese banking institution [49], the dataset contains 45,211 instances described by 17 features. The classification goal is to predict whether a client will subscribe to a term deposit ($Y = 1$). Age is used as the sensitive attribute: clients between 25 and 60 years old form the non-protected group, while those outside this range ($Z = 1$) constitute the protected group.

COMPAS. Released by ProPublica [1, 54], this dataset contains predictions from the COMPAS recidivism risk-assessment tool. Following [59], only defendants of Caucasian or African-American descent are retained, yielding 6,150 instances described by 11 features. The target is whether the defendant will *not* re-offend within two years ($Y = 1$) and the sensitive attribute is ethnicity, with African-Americans forming the protected group ($Z = 1$).

German credit. Contains 1,000 instances of bank account holder information described by 22 features [36]. The goal is to predict whether extending credit to an account holder is considered safe ($Y = 1$). Age is binarised such that individuals aged 25 or under form the protected group ($Z = 1$).

Law school. Drawn from a survey conducted by the Law School Admission Council across 163 US law schools [62], the dataset includes 20,798 instances described by 12 features. The target is whether a student passes the bar exam on the first attempt ($Y = 1$). Ethnicity is the sensitive attribute, with non-white students forming the protected group ($Z = 1$).

3.2 Pre-processing and Data Splitting

This section details the specific transformations applied to the raw data, including feature standardisation, the handling of memory-intensive datasets, and the stratified splitting protocol required for the poisoning attacks. Furthermore, it addresses the critical issue of class imbalance, justifying the resampling strategy adopted for specific datasets to prevent the models from degenerating into trivial predictors.

3.2.1 Feature standardisation

Following [59], all continuous features are standardised to zero mean and unit variance using a `StandardScaler` fitted exclusively on the training fold and applied to the test fold, thereby preventing any data leakage. An intercept column of ones is prepended to every feature matrix before model fitting.

3.2.2 Handling of the Adult (Census Income) dataset

The Adult (Census Income) dataset, with over 45,000 instances, imposed memory constraints that prevented running the full experiment in a single pass on the available hardware. The gradient-based poisoning procedure described in Section 2.2.1 requires holding the full training set and its gradient information in memory simultaneously, which exceeded available RAM for this dataset. The experiments were therefore split into two halves of the dataset. Each half was drawn as a stratified subsample preserving the joint distribution of the target Y and the sensitive attribute Z. The two halves were processed sequentially on the same machine, and the per-fold results were combined and averaged to obtain the final metrics.

Although the Bank dataset is comparably large, it did not suffer from the same memory limitations thanks to the undersampling procedure applied beforehand (see Section 3.2.5), which significantly reduced its effective size while preserving class proportions. All remaining datasets were processed in full without subsampling.

3.2.3 Train / validation / test split

Each dataset is partitioned into three disjoint sets. First, 20% of the data is held out as the *test set*. The remaining 80% is then split into a *training set* (50% of the total) and a *validation set* (30% of the total). The validation set is required by the gradient-based poisoning attack to evaluate the attacker's objective function during optimisation [55]. It is therefore never used to fit or to select the fairness-constrained model, ensuring that test-set evaluation remains uncontaminated.

All splits are stratified on the joint distribution of the target variable Y and the sensitive attribute Z to preserve both the class balance and the protected-group proportions across partitions. This is particularly important for fairness evaluation, as stratifying only on Y can lead to unstable or heavily skewed representations of certain Z groups in smaller folds.

3.2.4 Cross-validation procedure

As suggested in [65], in order to limit the variance of the reported results, we decided to apply a fivefold cross-validation. For each fold, the train/validation/test partition described above is re-applied to the data assigned to that fold, and the `StandardScaler` is re-fitted from scratch on the corresponding training split. Fold assignments are kept identical across all models and attack conditions by fixing the random seed (`CV_RANDOM_STATE`= 145), ensuring that performance differences across conditions are not confounded by fold variability.

3.2.5 Class-imbalance correction

Two datasets required intervention due to severe class imbalance (see Table 8 in appendix). Bank marketing has only 11.70% positive instances (Y-ratio 0.132), while Law school is skewed in the opposite direction with 95.00% positive instances (Y-ratio 0.053). In preliminary experiments, both caused the fairness-constrained classifier to degenerate into a trivial majority-class predictor, achieving 88.30% and 95.00% accuracy respectively with near-zero demographic parity difference, rendering both metrics uninterpretable for our robustness study.

To select a resampling strategy, we ran a preliminary comparison of six candidate methods (random over- and undersampling, ADASYN, SMOTE, SMOTETomek, and NearMiss [42]) on a dedicated train/validation split, evaluated via the F1-DP measure of [59] (adapted to use balanced accuracy). Random oversampling achieved the highest mean F1-DP (0.853), but random undersampling matched it closely (0.852) at substantially lower computational cost and was therefore selected.

Resampling is applied exclusively to the clean training set and triggered automatically when the minority-to-majority ratio falls below 0.15. Poisoned samples are appended *after* resampling, so the poisoning fraction is always computed relative to the original training set size.

Chapter 4

Implementation

This chapter describes the concrete implementation choices made to evaluate and compare the robustness of three classifiers under poisoning attacks: a standard logistic regression baseline, the fairness-constrained model of Zafar et al. [66], and the maximum entropy logistic regression of Vancompernolle et al. [59]. For each attack framework, all three models are evaluated under identical experimental conditions, allowing a direct comparison of their robustness. The two attack frameworks are evaluated along their most natural axis of variation, which differs between them for reasons rooted in their respective theoretical structures, as detailed in Sections 4.2 and 4.3.

4.1 Classifier Implementations

The experimental comparison relies on three distinct classifiers: an unconstrained baseline and two fairness-aware formulations representing different mathematical approaches to bias mitigation. The following subsections detail the specific implementation choices, hyperparameters, and preprocessing pipelines applied to each model.

4.1.1 Baseline logistic regression

The unconstrained baseline is a standard scikit-learn `LogisticRegression` [52] with solver `lbfgs` [43] and a maximum of 10,000 iterations. No fairness constraint is imposed. This model serves as a performance reference and is subjected to both attack frameworks under identical experimental conditions as the two fairness-aware models. Note that when this model is the target of the attack, the scenario becomes

a white-box attack, as the adversary is trained on the model itself. (See section 4.2.1)

4.1.2 Zafar model

The Zafar model [66] is implemented using the original author's code, adapted to Python 3. It enforces fairness by constraining the covariance between the sensitive attribute and the signed distance from the linear decision boundary, as described in Section 2.1.2. The fairness tolerance ε is set identically to the MaxEnt model on each fold, ensuring that both constrained models operate under comparable fairness requirements.

4.1.3 Maximum entropy logistic regression

The MaxEnt model follows directly the formulation of Vancompernolle et al. [59]. The implementation relies on the original code provided by the authors, which is used without modification for the core optimization routines, namely the `BinaryMaxEntropy` and `BinaryCovMinimization` classes. These respectively solve the auxiliary lower-bound problem to determine the minimum achievable covariance prior to fitting. Both problems are solved using CVXPY [16, 26].

4.1.4 Epsilon selection

For both the Zafar and MaxEnt models, the fairness tolerance ε is not set manually but derived from the data on each training fold. The lower-bound problem is first solved to compute the minimum achievable covariance ε^*, and the operational tolerance is then set to $1.05 * \varepsilon^*$ [59]. This ensures that neither model is fitted with an infeasible fairness constraint while remaining as close as possible to the best achievable fairness level on each fold. In the Jo attack experiments, this base value ε^* is further used as the reference point for the fairness tolerance sweep described in Section 4.3.

4.1.5 Preprocessing

For all three models, a constant intercept column of ones is prepended to every feature matrix before fitting. Continuous features are standardised to zero mean and unit variance using a `StandardScaler` fitted exclusively on the training fold and applied to the validation and test folds, preventing data leakage. This preprocessing pipeline is re-applied from scratch at each cross-validation fold. A fixed decision threshold of $\tau = 0.5$ is applied uniformly across all models to convert probabilistic outputs into binary predictions.

4.2 Gradient-Based Poisoning Attack (Solans)

The gradient-based attack of Solans et al. [55] is implemented using the `secml` adversarial machine learning library, specifically the `CAttackPoisoningLogisticRegression` class [48]. Additionally, parts of the original implementation provided by the authors were consulted and reused where appropriate. The attack optimises the *content* of poisoning points through projected gradient ascent, but does not determine their number: the poisoning budget must be specified externally by the experimenter.

We chose the axis of variation for this attack to be the poisoning fraction. The experiments are conducted at three levels,1%, 5%, and 15% of the training set size, while holding the fairness tolerance fixed at $\varepsilon = 0.05$ applied as a multiplier of the lower-bound ε^* computed on each fold as described in Section 4. This answers the question of how much data corruption is required to compromise the fairness constraint of each model. The attack is run separately against each of the three classifiers under identical hyperparameters and fold assignments.

4.2.1 Surrogate and threat model

We adopt a black-box threat model, which is widely considered more realistic and practical in real-world adversarial settings. In this model, the attacker has no direct access to the target classifier and cannot query it. Instead, the attacker optimizes poisoning samples using a surrogate model, in our case, a standard logistic regression classifier [55, 67].

The same surrogate is used when attacking the LogReg, Zafar and MaxEnt formulations. As stated previously, when LogReg is the target, the surrogate coincides with the target architecture, making the attack effectively white-box in that specific case. For the two fairness-aware models, the surrogate provides a natural and consistent approximation since both approaches are rooted in logistic regression [17, 55].

As a result, differences in robustness can be interpreted as stemming from the way fairness constraints are incorporated into the models rather than from discrepancies in the attack procedure itself. This also provides a controlled experimental setting for assessing whether constraining the decision boundary (Zafar) or directly constraining the predicted probabilities (MaxEnt) leads to different levels of resistance to poisoning attacks.

4.2.2 Optimiser configuration

The poisoning point is updated by projected gradient ascent with step size $\eta = 0.05$, a minimum step size of $\eta_{\min} = 0.05$, a maximum of 1000 iterations, and a convergence tolerance of 10^{-6}. At each iteration the updated point is projected onto the box constraints defined by the feature-wise minimum and maximum values of the clean training data.

4.2.3 Poisoning fractions

Experiments are conducted at three poisoning fractions: 1%, 5%, and 15% of the training set size. A single poisoning point is optimised per attack run and injected into the training data before each model is refitted.

These poisoning levels were selected to cover a range of realistic attack intensities, from weak perturbations (1%) to more severe data corruption (15%). In preliminary experiments, larger poisoning fractions led to a substantial collapse in predictive accuracy for all models, making the comparison of fairness-aware formulations less informative, as performance degradation became dominated by the sheer quantity of poisoned data rather than by differences in model robustness. Moreover, increasing the poisoning fraction beyond 15% resulted in significantly higher computational costs and optimisation times while providing limited additional insight into the relative behaviour of the models under attack. The selected fractions therefore offer a practical compromise between attack strength, computational tractability, and interpretability of the robustness analysis.

4.3 Optimal Flipping Attack (Jo)

The flipping attack of Jo et al. [34] is implemented from scratch following the Appendix A (demographic parity variant) of the paper. Unlike the Solans attack, the poisoning budget is not a free parameter. The attack analytically derives α^*, the theoretically minimum number of sensitive attribute flips required to force a fair classifier toward a target model. Sweeping over poisoning fractions would therefore be redundant, the attack already operates at the optimal budget by construction. What varies instead is the fairness tolerance ε, swept over five values $\{0.01, 0.02, 0.05, 0.10, 0.20\}$ applied as multipliers of the lower-bound ε^* computed on each fold. This allows us to examine how the tightness of the fairness constraint determines vulnerability to a minimally invasive but theoretically optimal attack. The attack is evaluated on all three models for comparison.

4.3.1 Attack procedure

For each training fold the attack proceeds in three steps. First, an unconstrained logistic regression is fitted on the clean training data to obtain the target model h_{target}. Second, the four group-size quantities are computed on the full training set without conditioning on Y, where $h = h_{\text{target}}(\mathbf{x}_i) \in \{0, 1\}$ denotes the binary prediction of the target model on sample i. These four quantities partition the training set into the cells of a 2×2 contingency table crossing the model's binary prediction $h \in \{0, 1\}$ with the sensitive attribute $Z \in \{0, 1\}$:

	$Z = 0$	$Z = 1$
$h = 0$	P	R
$h = 1$	Q	S

Third, the theoretically optimal number of flips is derived as

$$\alpha^* = \left\lfloor \frac{|PS - QR|}{\max(P + R,\, Q + S)} \right\rfloor \tag{4.1}$$

The quantity $|PS - QR|$ is the absolute cross-product difference of this contingency table, which equals zero if and only if h and Z are independent, that is, the fairness constraint is exactly satisfied. The formula for α^* therefore quantifies the minimum number of sensitive attribute flips required to drive this cross-product to zero, directly neutralising the covariance constraint [34].

The sign of $PS - QR$ determines which demographic group needs to be targeted to maximally disrupt the fairness constraint. If $PS > QR$, the attack targets samples where $h = 0$ and $Z = 1$ (the flip pool is $\{h = 0, Z = 1\}$, of size R). If $PS < QR$, it targets samples where $h = 1$ and $Z = 0$ (the flip pool is $\{h = 1, Z = 0\}$, of size Q). The sensitive attribute Z is then flipped for α^* randomly selected samples from this pool, while features X and labels Y are left unchanged.

4.3.2 Cap modes

Two conditions are evaluated: a `baseline` condition in which no flips are applied, providing a clean performance reference, and an `alpha` condition in which the theoretically optimal α^* flips are applied, faithfully reproducing Theorem 2 of [34].

4.4 Experimental Pipeline and Metrics

The following subsections describe the cross-validation protocol designed to minimize variance, the dual-metric approach required to detect "silent failures," and the specific libraries utilized to implement the attack and defense mechanisms.

4.4.1 Cross-validation

All experiments use five-fold cross-validation with a fixed random seed of 145, keeping fold assignments identical across all models and attack conditions. This ensures that observed differences between the baseline, Zafar, and MaxEnt models under attack are attributable to the models themselves rather than to variations in the data splits. Using the same folds across experiments also improves the comparability and reproducibility of the robustness evaluation.

4.4.2 Evaluation metrics

Performance is primarily assessed using accuracy (ACC) and the absolute demographic parity difference (DPD). We additionally collected the signed demographic parity difference to examine the direction of bias. Reporting both absolute and signed DPD is particularly relevant in this adversarial setting, as it reveals in which direction an attack shifts demographic parity. Notably, two models can achieve the same absolute DPD through opposite bias directions.

4.4.3 Software

The full pipeline is implemented in Python 3.10. The main dependencies are `numpy` [28], `pandas` [45, 56], `scikit-learn` [52], `cvxpy` [16, 26], `secml` [48], and `imbalanced-learn` [42].

Chapter 5

Results

This chapter is divided into four sections. The first establishes the clean performance of the three models in the absence of any attack, serving as a reference point for the subsequent analysis. The second section investigates whether the gradient-based Solans attack has a statistically significant effect on MaxEnt, and compares its robustness against the Zafar model under increasing poisoning fractions. The third section presents the same analysis for the optimal flipping attack of Jo et al. Finally, the fourth section summarises the findings across both attack frameworks and draws conclusions with respect to the research questions posed in the Introduction.

5.1 Baseline Performance

Before evaluating robustness under attack, we verify that the three models behave as expected on clean data. Table 5.1 reports the mean accuracy and demographic parity difference across the five datasets and five folds in the absence of any poisoning. Overall, the results confirm the expected trade-off between accuracy and fairness. Both MaxEnt and Zafar generally reduce demographic parity relative to the unconstrained logistic regression baseline, at the cost of a modest decrease in predictive accuracy.

Some baseline results reported here differ from those obtained by Vancompernolle et al. [59] for several reasons. First, the undersampling procedure was applied to the Bank marketing and Law school datasets in order to mitigate the strong class imbalance present in the original data. This preprocessing step modifies the effective training distribution and therefore affects both the predictive performance and the fairness metrics of the resulting classifiers.

Second, the fairness tolerance parameter ε was defined as a fixed fraction of

Table 5.1: Clean performance (no attack) of the three models across all datasets. Mean ± standard deviation over five folds.

Dataset / Model	Accuracy	Demographic Parity
	ADULT	
LogReg	0.8457 ± 0.0032	0.0702 ± 0.0075
MaxEnt	0.8451 ± 0.0024	0.0624 ± 0.0096
Zafar	0.8459 ± 0.0029	0.0698 ± 0.0084
	BANK	
LogReg	0.8724 ± 0.0032	0.3598 ± 0.0367
MaxEnt	0.8645 ± 0.0053	0.1579 ± 0.0183
Zafar	0.8681 ± 0.0043	0.0960 ± 0.0468
	COMPAS	
LogReg	0.6771 ± 0.0101	0.2444 ± 0.0123
MaxEnt	0.6694 ± 0.0106	0.1157 ± 0.0202
Zafar	0.6397 ± 0.0154	0.1455 ± 0.0143
	GERMAN	
LogReg	0.7590 ± 0.0230	0.1217 ± 0.0850
MaxEnt	0.7480 ± 0.0189	0.0846 ± 0.0944
Zafar	0.7420 ± 0.0148	0.0012 ± 0.0653
	LAW	
LogReg	0.8439 ± 0.0051	0.3769 ± 0.0247
MaxEnt	0.7983 ± 0.0097	0.0179 ± 0.0237
Zafar	0.8127 ± 0.0106	0.0055 ± 0.0147

the covariance scale (`FAIRNESS_FRAC` $= 0.05$) rather than being tuned individually for each dataset. This choice was made in order to maintain a consistent fairness constraint across all experiments and attack settings. Since the primary objective of this thesis is to compare the robustness of the different models under poisoning attacks, fixing the relative fairness constraint avoids introducing additional variability caused by dataset-specific hyperparameter tuning. Using a single fairness fraction ensures that observed differences in robustness are attributable to the models and attack mechanisms themselves rather than to changes in the operating point selected for each dataset.

Finally, as discussed in Section 3.2.2, the Adult dataset was processed in two

separate halves due to computational constraints, after which the results were aggregated. This procedure may introduce additional variability in the reported metrics.

The qualitative patterns are nonetheless consistent with the original paper. Across all datasets, the fairness-aware models generally achieve lower demographic parity differences than the unconstrained logistic regression baseline while maintaining comparable levels of predictive accuracy. The purpose of these baseline metrics is therefore not to exactly reproduce the numerical values reported in the original work, but rather to establish a consistent and controlled reference setting for evaluating the impact of poisoning attacks on the different fairness-aware formulations.

5.2 Gradient-Based Attack (Solans)

This section presents the empirical results obtained using the gradient-based poisoning attack proposed by Solans et al. [55]. The analysis proceeds in three stages. First, we establish a baseline by evaluating the attack's effect on an unconstrained Logistic Regression model to determine if the induced performance degradation is statistically significant in the absence of fairness constraints. Second, we isolate the impact on the Maximum Entropy (MaxEnt) model to assess whether the fairness-aware formulation introduces distinct vulnerabilities or failure modes. Finally, we compare this behavior directly against the Zafar model to evaluate their relative robustness and identify potential "silent failure" scenarios where fairness degrades while accuracy remains stable.

5.2.1 Effect of the Attack on LogReg

We first establish a reference point by evaluating the Solans attack on the unconstrained logistic regression baseline. Since LogReg optimizes solely for predictive accuracy and imposes no fairness constraint, it provides a control condition: any degradation observed here would reflect the attack's general capacity to disrupt a learning objective, rather than its ability to exploit a fairness mechanism specifically.

Figure 5.1 shows the evolution of accuracy and signed demographic parity as the poisoning fraction increases from 0% to 15%. No consistent pattern emerges across datasets, some datasets show a slight accuracy drop while others remain stable or even improve marginally, and demographic parity fluctuates without any clear directional trend.

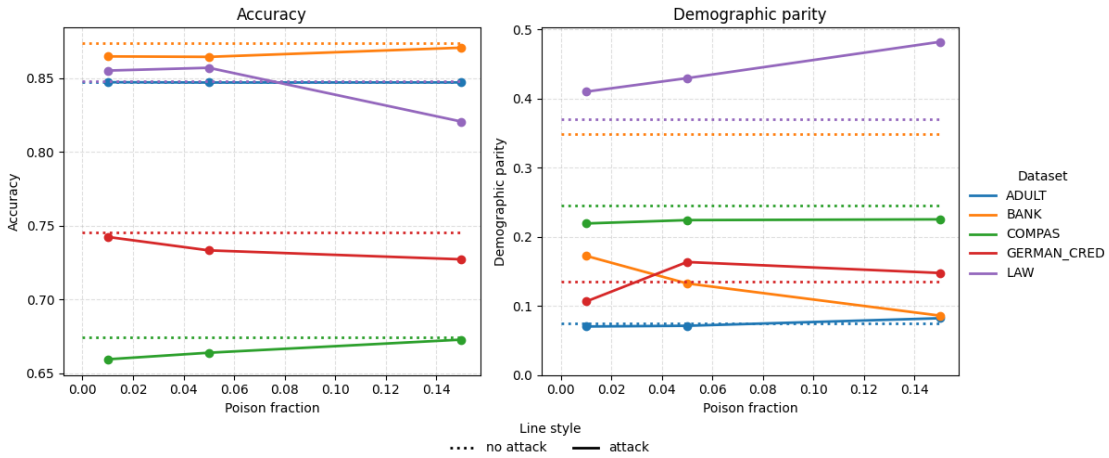
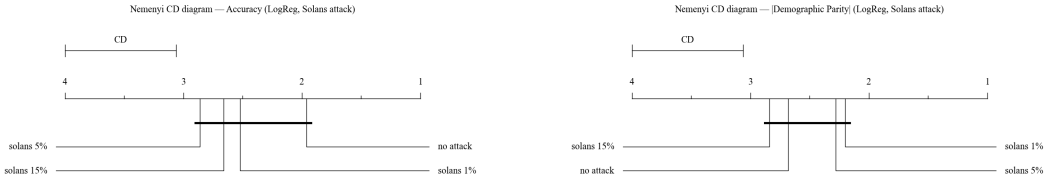


Figure 5.1: Evolution of LogReg accuracy (left) and signed demographic parity (right) under the Solans attack across increasing poisoning fractions, shown per dataset.



(a) Nemenyi test for Accuracy

(b) Nemenyi test for Demographic Parity

Figure 5.2: Nemenyi post-hoc test results for Logistic Regression under the Solans attack. The diagrams compare performance ranks across different poisoning fractions. Groups connected by the horizontal black bar are not statistically significantly different ($p > 0.05$). **(a)** shows results for Accuracy, while **(b)** shows results for Signed Demographic Parity.

This is confirmed statistically by the Nemenyi post-hoc diagrams in Figure 5.2, where all poisoning conditions remain connected by the critical difference bar for both metrics, indicating that no pair of conditions is significantly different at $\alpha = 0.05$. The Solans attack therefore has no statistically significant effect on the unconstrained baseline, neither in terms of predictive accuracy nor demographic parity. This result sharpens the subsequent analysis, any degradation observed on MaxEnt or Zafar can be attributed to the presence of the fairness constraint rather than to a general sensitivity of logistic regression to poisoning.

5.2.2 Effect of the Attack on MaxEnt

We first assess whether the Solans attack has a statistically significant effect on the MaxEnt model by comparing its performance under no attack against three poisoning fractions: 1%, 5%, and 15%. Figure 5.3 shows the evolution of accuracy and demographic parity for each dataset as the poisoning fraction increases.

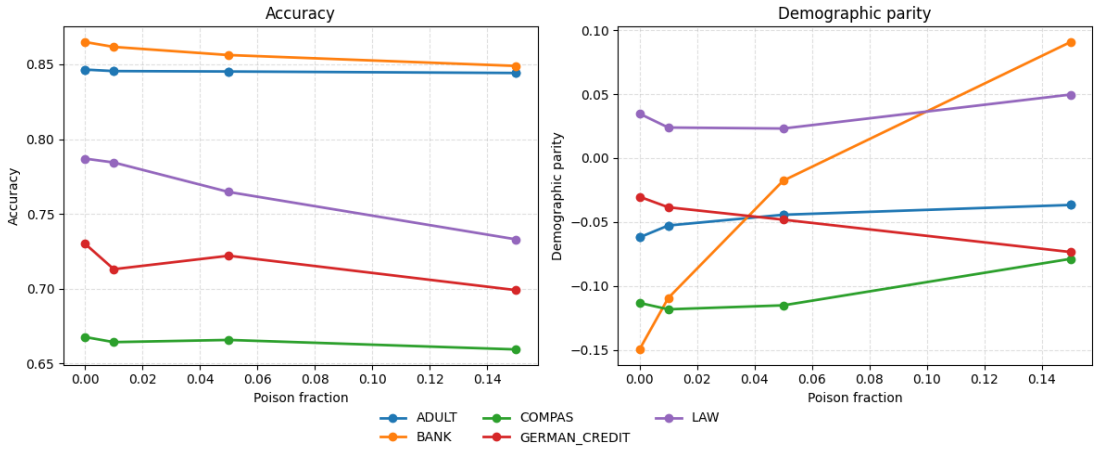


Figure 5.3: Evolution of MaxEnt accuracy (left) and signed demographic parity (right) under the Solans attack across increasing poisoning fractions, shown per dataset.

Accuracy degrades progressively with the poisoning fraction across all datasets, though the magnitude varies considerably. The effect is most pronounced on Law school, which drops from approximately 0.80 to 0.73 between no attack and 15% poisoning, while Adult and Bank marketing remain comparatively stable throughout.

The demographic parity results reveal a more nuanced picture. We report the *signed* demographic parity rather than its absolute value in order to capture the direction of the induced discrimination. Across most datasets the attack pushes parity monotonically in one direction, suggesting that the gradient-based procedure consistently exploits the same discriminatory pathway. The most striking example is Bank Marketing, whose demographic parity begins at approximately -0.15 and increases to $+0.09$ at the 15% poisoning fraction.

This sign reversal seems to indicate that the attack identifies and leverages the most efficient route to degrade fairness within the specific constraints of the dataset and switches the direction of disparity, moving from disadvantaging one group to disadvantaging the other. German Credit doesn't follow the same direction as

the other datastes which may reflect either the sensitivity of this small dataset (n = 1,000) to individual poisoning points, or the MaxEnt fairness constraint partially counteracting the attack at certain poisoning levels.

The Friedman test is applied to assess whether the observed differences are statistically significant across folds and datasets, under the null hypothesis H_0 that all poisoning conditions perform equivalently. The test operates on a matrix of $N = 25$ observations (5 datasets $\times$ 5 folds) and $k = 4$ conditions (no attack, 1%, 5%, 15% poisoning), where each cell contains the performance of MaxEnt on a given dataset-fold pair under a given poisoning level. Table 5.2 reports the test statistics and p-values for both metrics. Figures 5.5 and 5.4 display the average Friedman ranks for accuracy and demographic parity respectively, with the dashed line indicating the expected rank under the null hypothesis of no difference between conditions.

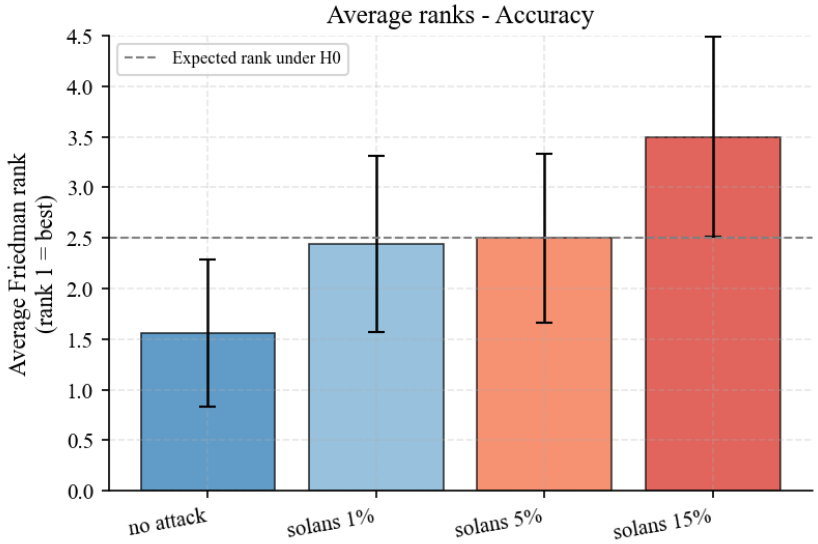


Figure 5.4: Average Friedman ranks for accuracy under the Solans attack on MaxEnt. Rank 1 = best (highest accuracy). The dashed line marks the expected rank under H_0.

For accuracy, the rank plot shows a clear monotone degradation. The no-attack condition consistently receives the best average rank (1.58), while the 15% poisoning condition receives the worst (3.50), well above the H_0 expectation of 2.5. For demographic parity the pattern is less regular. Notably, the 1% and 5% poisoning conditions receive *lower* average ranks than the no-attack condition, indicating that light poisoning temporarily improves the measured fairness. This counter-intuitive result is consistent with the sign-reversal observed on Bank marketing. At low

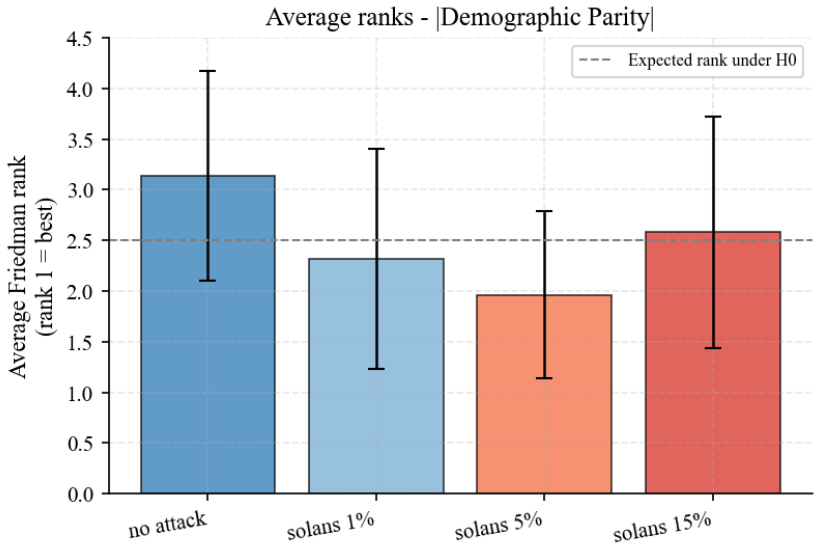


Figure 5.5: Average Friedman ranks for |demographic parity| under the Solans attack on MaxEnt. Rank 1 = best (lowest parity). The dashed line marks the expected rank under H_0.

poisoning fractions the attack may push the demographic parity of some datasets closer to zero before eventually driving it in the opposite direction.

Table 5.2: Friedman test results for the effect of the Solans attack on MaxEnt ($\alpha = 0.05$).

Metric	χ^2_F	**df**	p-value
Accuracy	29.1235	3	< 0.001
|Demographic parity|	11.1446	3	0.01097

As seen in table 5.2, the Friedman test indicates that the poisoning fraction has a statistically significant effect on both accuracy and demographic parity for the MaxEnt model. The effect is particularly strong for accuracy ($\chi^2_F = 29.12, p < 0.001$), while demographic parity also exhibits significant variations across attack levels ($\chi^2_F = 11.14, p = 0.011$).

As the Friedman test rejects the null hypothesis for both metrics, the Nemenyi post-hoc test is applied to identify which specific conditions differ significantly. The results are presented in Figure 5.6 and 5.7. The Wilcoxon signed-rank tests confirm and refine the Friedman results. For accuracy, all three poisoning fractions produce a statistically significant degradation relative to the no-attack condition,

with the effect strengthening as the poisoning budget increases (no attack vs. 15%: $p < 0.001$). This is consistent with the monotone rank pattern observed in Figure 5.4.

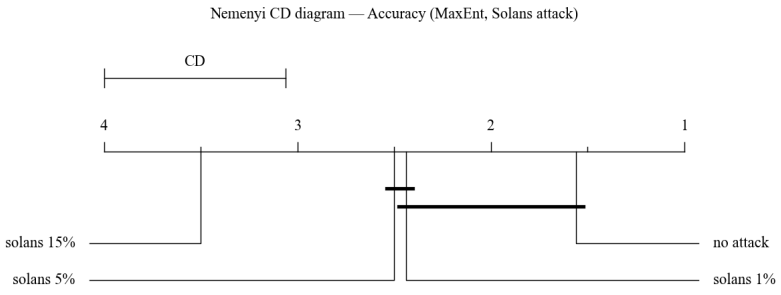


Figure 5.6: Nemenyi critical difference diagram for accuracy under the Solans attack (MaxEnt). Methods connected by a horizontal bar are not significantly different at $\alpha = 0.05$.

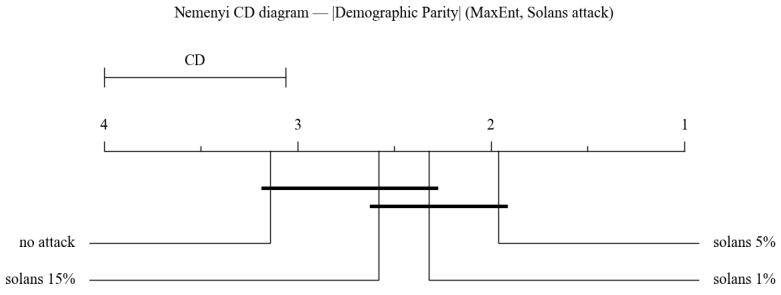


Figure 5.7: Nemenyi critical difference diagram for demographic parity under the Solans attack (MaxEnt). Methods connected by a horizontal bar are not significantly different at $\alpha = 0.05$.

For demographic parity the results exhibit a complex, non-monotonic pattern. The 1% and 5% poisoning conditions are both significantly different from the no-attack baseline ($p = 0.011$ and $p = 0.007$ respectively), but the 15% condition is not ($p = 0.199$). This apparent inconsistency at the 15% level result is explained by the sign-reversal discussed earlier. At 15% poisoning, the attack pushes demographic parity in opposite directions across datasets. When the absolute value is taken, some datasets show an increase and others a decrease relative to the baseline, which cancel out in the paired test and prevent it from reaching significance. The attack therefore does affect fairness at high poisoning budgets, but in an asymmetric and dataset-dependent way that aggregate tests cannot fully capture.

It is important to contextualize these irregularities within the nature of the Solans attack itself. As one of the pioneering works in fairness poisoning, the primary objective of Solans et al. [55] was to demonstrate the *feasibility* of such attacks rather than to provide a fully optimized, stable weapon across all domains. The observed volatility likely stems from this foundational design, the attack relies on a bilevel optimization that injects single points based on gradients computed from a standard Logistic Regression surrogate. When transferred to the MaxEnt formulation, this surrogate mismatch, combined with the high sensitivity of a single-point injection to specific dataset geometries, could explain this unstable behavior.

Table 5.3: Wilcoxon signed-rank test results for the Solans attack effect on MaxEnt ($\alpha = 0.05$).

Metric	**Comparison**	**Statistic**	p**-value**	**Significant**
Accuracy	No attack vs 1%	42.5	0.0037	Yes
Accuracy	No attack vs 5%	53.5	0.0058	Yes
Accuracy	No attack vs 15%	20.0	< 0.001	Yes
|Demographic parity|	No attack vs 1%	68.0	0.0110	Yes
|Demographic parity|	No attack vs 5%	63.0	0.0074	Yes
|Demographic parity|	No attack vs 15%	105.0	0.1985	No

5.2.3 Model Robustness Comparison

We now compare the robustness of the MaxEnt and Zafar models under the Solans attack, with the unconstrained logistic regression included as a reference floor. The logistic regression does not include a fairness constraint and is not subject to the poisoning attack, it is reported solely as a reference point. Figure 5.8 shows the mean performance aggregated over all datasets, providing a global view of the effect of increasing poisoning fractions on each model.

Two observations stand out. First, MaxEnt suffers a visible accuracy drop as poisoning increases, whereas Zafar's accuracy remains flat or even improves slightly. This is not necessarily a weakness of MaxEnt, the drop provides a detectable signal. A practitioner monitoring accuracy would observe it and could trigger a retraining or auditing procedure. Second, and more importantly, Zafar's demographic parity deteriorates sharply. It rises from roughly 0.074 at baseline to 0.215 at 15% poisoning, while MaxEnt's parity remains low and stable throughout, reaching only 0.063 at the same budget. This pattern is precisely the adversarial scenario described in Section 1.3.3: the attack maintains or even improves accuracy while

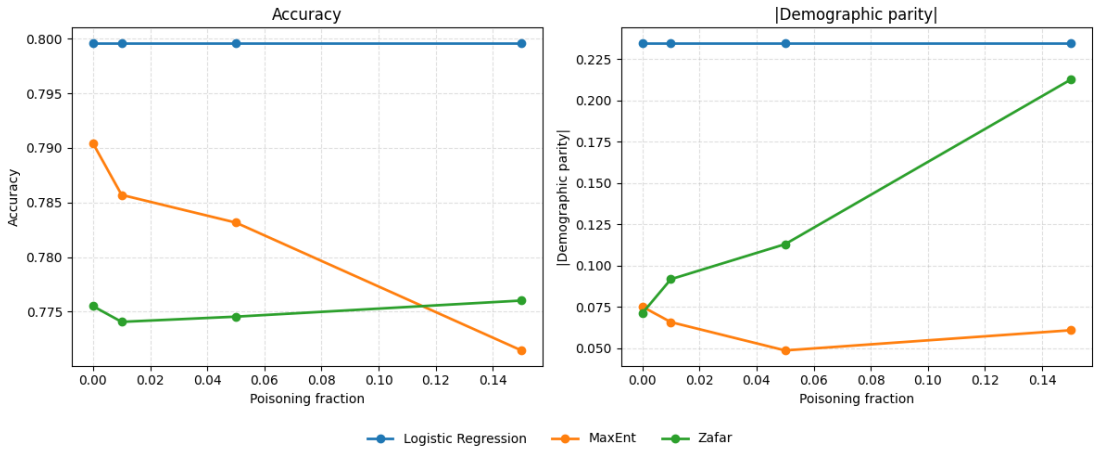


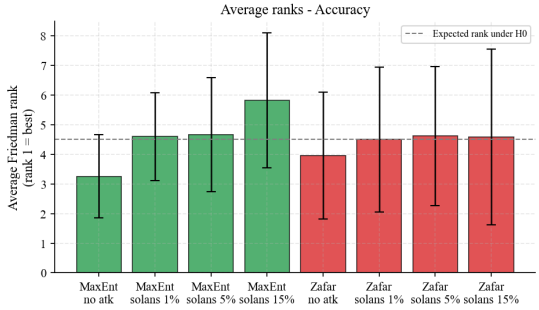
Figure 5.8: Evolution of accuracy and demographic parity under the Solans attack for the LogReg, MaxEnt, and Zafar models across increasing poisoning fractions. The curves represent averaged results over all datasets.

silently degrading fairness, remaining invisible to standard performance monitoring. MaxEnt, while more sensitive to accuracy degradation, seems to preserve its fairness guarantee considerably better under the same attack conditions.

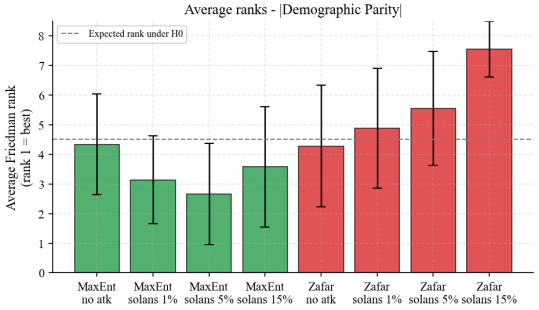
Figures 5.9a and 5.9b display the average Friedman ranks for accuracy and demographic parity across all conditions. The rank plots confirm and sharpen the picture from the evolution figure. For demographic parity (Figure 5.9b), all four MaxEnt conditions receive average ranks well below the H_0 expectation of 4.5, meaning MaxEnt consistently achieves lower demographic parity than Zafar regardless of the poisoning level. The Zafar conditions, by contrast, are all above 4.5 and worsen monotonically with the poisoning fraction, reaching an average rank of 7.6 at 15% poisoning, the worst of all eight conditions by a considerable margin.

For accuracy (Figure 5.9a), the pattern is reversed. MaxEnt no-attack achieves the best rank (3.3) but degrades steadily, with MaxEnt at 15% reaching rank (5.9). Zafar's accuracy ranks remain flat across all poisoning levels, hovering near (4.5) regardless of the attack budget. This indicates that, while both models are affected by the attack, the observed changes in accuracy remain limited for Zafar across all poisoning levels.

The Wilcoxon pairwise tests (Table 5.4) confirm the trends observed in the Friedman rank analysis. For demographic parity, the difference between MaxEnt and Zafar is statistically significant at all poisoning fractions, and the significance strengthens as the attack budget increases. This result is consistent with the widening gap observed in Figure 5.9b, where Zafar's parity deteriorates progressively



(a) Accuracy. Rank 1 = best (highest accuracy).



(b) |Demographic parity|. Rank 1 = best (lowest parity).

Figure 5.9: Average Friedman ranks under the Solans attack, model comparison. The dashed line marks the expected rank under H_0.

while MaxEnt remains comparatively stable.

For accuracy, by contrast, no statistically significant difference is detected between the two models at any poisoning level. This absence of significance was already suggested by the Friedman rank plots, where both models exhibited relatively similar average ranks, particularly at the 1% and 5% poisoning fractions. At 15%, the larger variability observed across datasets likely further reduces the ability of the paired test to detect a systematic difference.

Taken together, these results indicate that the main distinction between the two models lies not in predictive performance, but in the preservation of fairness under attack. While both models maintain comparable accuracy, Zafar experiences a substantial degradation in demographic parity as the poisoning fraction increases, whereas MaxEnt preserves substantially lower disparity across all attack levels.

Table 5.4: Wilcoxon signed-rank test results for the model robustness comparison under the Solans attack ($\alpha = 0.05$).

Metric	**Comparison**	**Statistic**	p**-value**	**Significant**
Accuracy	MaxEnt vs Zafar (1%)	146.0	0.9090	No
Accuracy	MaxEnt vs Zafar (5%)	143.0	0.5998	No
Accuracy	MaxEnt vs Zafar (15%)	114.0	0.1919	No
|Demographic parity|	MaxEnt vs Zafar (1%)	79.0	0.0247	Yes
|Demographic parity|	MaxEnt vs Zafar (5%)	33.0	0.0005	Yes
|Demographic parity|	MaxEnt vs Zafar (15%)	7.0	< 0.001	Yes

5.3 Optimal Flipping Attack (Jo)

Unlike the gradient-based Solans attack, the attack proposed by Jo et al. operates by strategically flipping sensitive attribute labels in order to weaken the fairness constraint enforced during training. Its effectiveness therefore depends directly on the strength of the fairness tolerance parameter ε. To fully understand this mechanism, we proceed in three steps: first, we verify that the attack is inert without a fairness constraint using Logistic Regression. Second, we establish the baseline behavior of the MaxEnt model in the absence of poisoning, and third, we analyze the specific impact of the attack on these fairness-aware models.

5.3.1 Unconstrained Logistic Regression

We first examine the effect of the Jo attack on the unconstrained logistic regression baseline. As argued in Section 2.2.2, the attack is designed to exploit the fairness constraint enforced during training by flipping a minimal number of sensitive attribute labels to make the unconstrained risk minimiser feasible under the fair learning problem. Since no fairness constraint is imposed on LogReg, the attack has no mechanism through which to alter the model's behaviour.

Table 5.5 confirms this. The performance of LogReg under the baseline and alpha cap modes is identical across all datasets and fairness fractions, for both accuracy and demographic parity. The sensitive attribute flips introduced by the attack leave the training objective entirely unchanged, producing no shift in the learned parameters. This result precisely localises the vulnerability introduced by fairness constraints. The Jo attack does not exploit the model's predictive structure, it exploits the constraint itself. Any degradation observed in MaxEnt and Zafar in the subsequent analysis is therefore directly attributable to the presence of the fairness constraint rather than to any intrinsic sensitivity of logistic regression to label flipping.

5.3.2 MaxEnt Without Poisoning

Having established that the attack requires a fairness constraint to function, we must first characterise how the MaxEnt model behaves under different constraint levels in the absence of any poisoning. This provides the necessary reference point to distinguish the intrinsic fairness–accuracy trade-off from the additional effect introduced by the attack.

We evaluate the MaxEnt model across five fairness tolerance fractions. Increasing the fairness tolerance progressively relaxes the covariance constraint imposed during

Table 5.5: LogReg performance under the Jo attack: baseline vs. alpha cap mode, averaged over five folds. Identical values across all datasets confirm that the attack has no effect on the unconstrained model.

	Accuracy		|Demographic Parity|	
Dataset	Baseline	Alpha	Baseline	Alpha
Adult	0.8455	0.8455	0.0689	0.0689
Bank	0.8713	0.8713	0.3537	0.3537
COMPAS	0.6759	0.6759	0.2463	0.2463
German Credit	0.7620	0.7620	0.1247	0.1247
Law School	0.8376	0.8376	0.3833	0.3833

training, allowing a larger dependence between predictions and the protected attribute, and therefore enforcing less strict fairness requirements.

Figure 5.10 shows the evolution of accuracy and demographic parity for each dataset as the fairness constraint is relaxed.

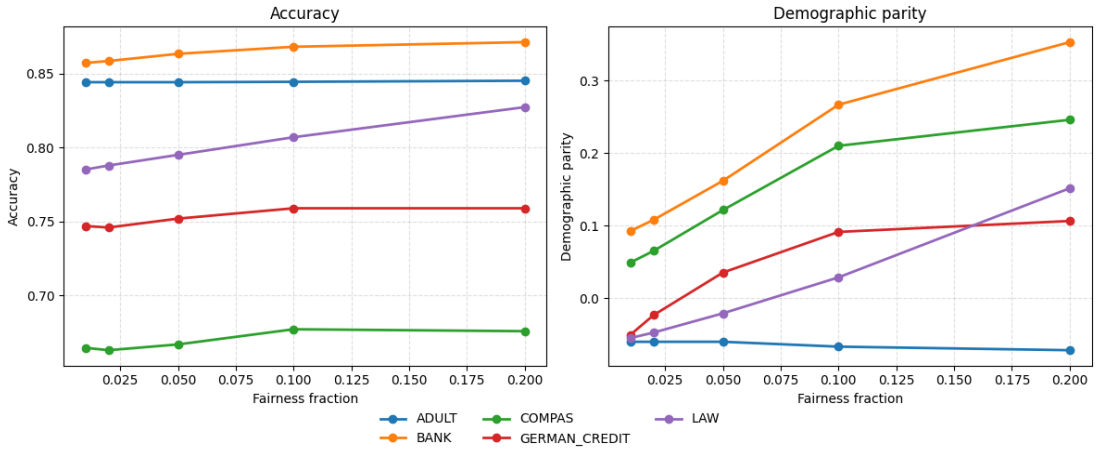


Figure 5.10: Evolution of MaxEnt accuracy (left) and signed demographic parity (right) under the Jo attack across increasing fairness tolerance fractions, shown per dataset.

The figure reveals two clear patterns. First, accuracy shows a modest but consistent increase as the fairness tolerance grows. This behaviour is expected, relaxing the fairness constraint gives the model greater flexibility to optimise

predictive performance. The effect is particularly visible on the Law school dataset, where accuracy increases from approximately 0.79 to 0.83 across the considered range.

Second, for most datasets the attack pushes demographic parity in one direction regardless of where it begins, with German Credit, Law school, and Adult all exhibiting sign reversals as the fairness tolerance increases, a pattern also observed under the Solans attack (Section 5.2). More broadly, demographic parity deteriorates as the fairness tolerance increases, reflecting the expected behaviour of the model: weaker constraints allow larger disparities between protected groups. Conversely, the results highlight the effectiveness of the MaxEnt constraint at low tolerance levels, where signed demographic parity values remain close to zero across most datasets.

5.3.3 Effect of the Attack Across Fairness Fractions

With the unconstrained control and the clean baseline established, we now isolate the specific impact of the Jo attack on the fairness-aware MaxEnt model. Since we have proven that the attack targets the constraint mechanism itself, we expect to see significant deviations from the baseline behavior described above, particularly where constraints are active.

Figure 5.11 overlays the baseline cap mode (solid lines) and the alpha cap mode (dashed lines) for both metrics, allowing a direct visual comparison of the attack's effect at each fairness fraction per dataset.

For accuracy, the attacked model (dashed) tends to perform slightly better than the baseline at low fairness fractions, and the two lines converge as ε increases. The optimal flipping attack, by flipping a minimal number of sensitive attribute labels, shifts the training distribution toward the unconstrained risk minimiser, which achieves higher accuracy at the cost of introducing demographic disparity, see figure 2.3. The model under attack therefore recovers some of the accuracy penalty imposed by the fairness constraint, at the expense of its fairness guarantee. As the fairness fraction grows and the constraint becomes looser, the gap between attacked and unattacked conditions narrows, this might be due to the baseline model being already allowed to operate closer to the unconstrained optimum and the attack has less room to exploit.

Law school is a partial exception, where the gap persists across the full range. Unlike the other datasets, this behaviour does not appear to be explained by dataset size or class imbalance alone, since Law school contains approximately 20,000 instances and underwent the same undersampling procedure as Bank marketing, which does not exhibit the same pattern. The origin of this discrepancy remains

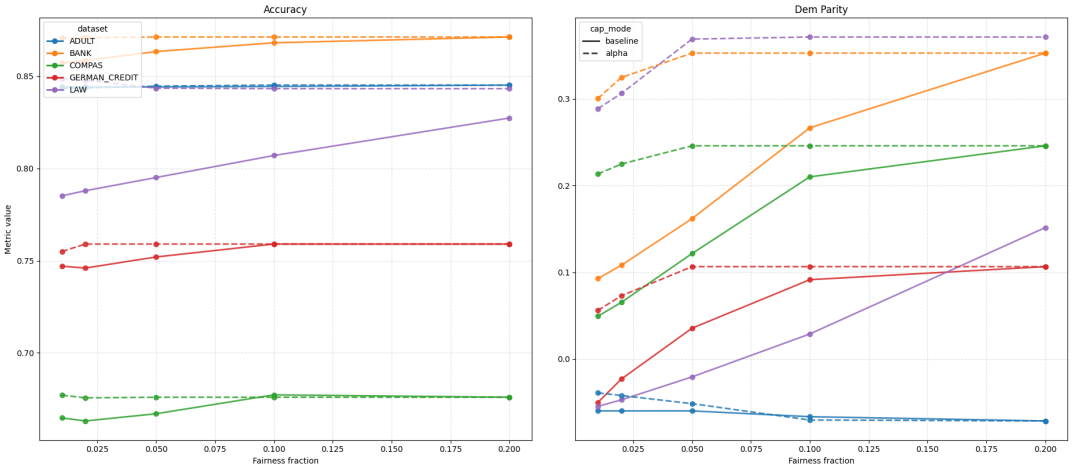


Figure 5.11: Comparison of MaxEnt performance under the baseline cap mode (solid) and the alpha cap mode (dashed) across fairness tolerance fractions, shown per dataset.

unclear and could reflect the specific structure of the sensitive attribute or the base rate distribution in this dataset, though further investigation would be needed to confirm this.

For demographic parity the pattern is more pronounced and more concerning. The attacked model consistently exhibits higher parity than the baseline at low fairness fractions, and the gap is often largest at the tightest constraint settings. When the model is forced to enforce fairness strictly, a small number of strategically flipped sensitive attribute labels can maximally disrupt the covariance constraint and push the model toward a discriminatory solution. As the fairness tolerance relaxes, the baseline model itself tolerates more parity, and the two lines gradually converge. By $\varepsilon = 0.20$, attacked and unattacked conditions reach similar parity levels on most datasets, suggesting that at very loose constraints the attack's additional effect becomes negligible relative to the parity the model already accepts.

Both attack frameworks exhibit a strong tendency to push demographic parity monotonically in a single direction. This suggests that both the gradient-based and optimal flipping procedures successfully exploit a dominant discriminatory pathway inherent to the data distribution. It is a notable phenomenon that, despite their fundamentally different mechanisms, both methods converge on this consistent directional behavior. Each attack appears to identify and leverage the most efficient route to degrade fairness within the specific constraints of the dataset.

This consistency implies that the vulnerability to directional bias is a structural property of the data and the fairness constraints themselves, which both adversarial strategies are able to locate and exploit reliably.

Figures 5.12 and 5.13 display the average Friedman ranks for accuracy and demographic parity across all conditions.

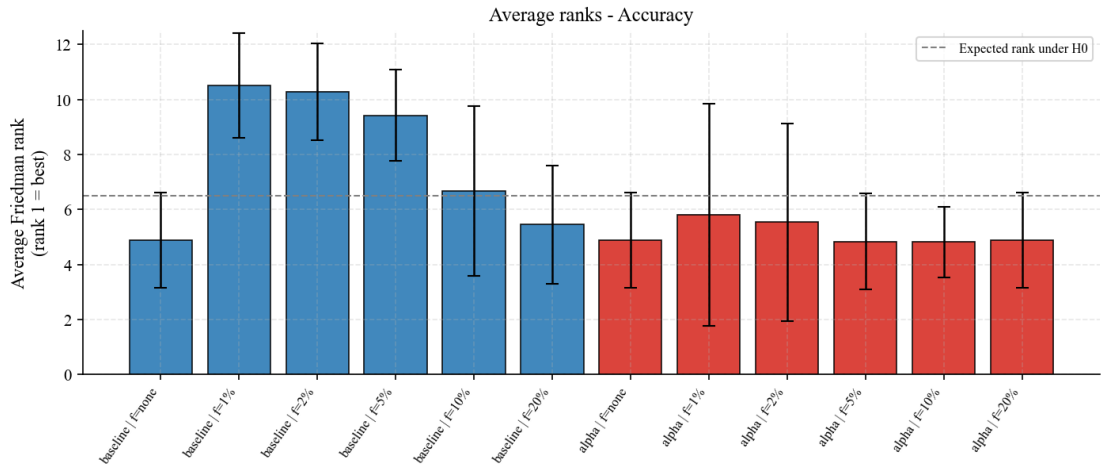


Figure 5.12: Average Friedman ranks for accuracy under the Jo attack. Blue = baseline cap mode, red = alpha cap mode. Rank 1 = best (highest accuracy). The dashed line marks the expected rank under H_0.

The rank plots confirm the patterns described above. For accuracy, the baseline conditions at tight fairness fractions receive extremely poor average ranks (around 10), reflecting the accuracy cost of strict constraint enforcement, while the alpha conditions cluster near 5 throughout. For demographic parity the picture inverts, tight baseline conditions rank best, and the two cap modes converge at loose fairness fractions where the constraint provides little protection regardless of the attack. Both plots are consistent with the observations that the Jo attack is most effective at tight fairness tolerances and loses its marginal impact as the constraint relaxes.

Table 5.6 reports the Wilcoxon signed-rank test results for each fairness fraction. All comparisons are significant for both metrics, confirming statistically what the evolution figures suggested visually.

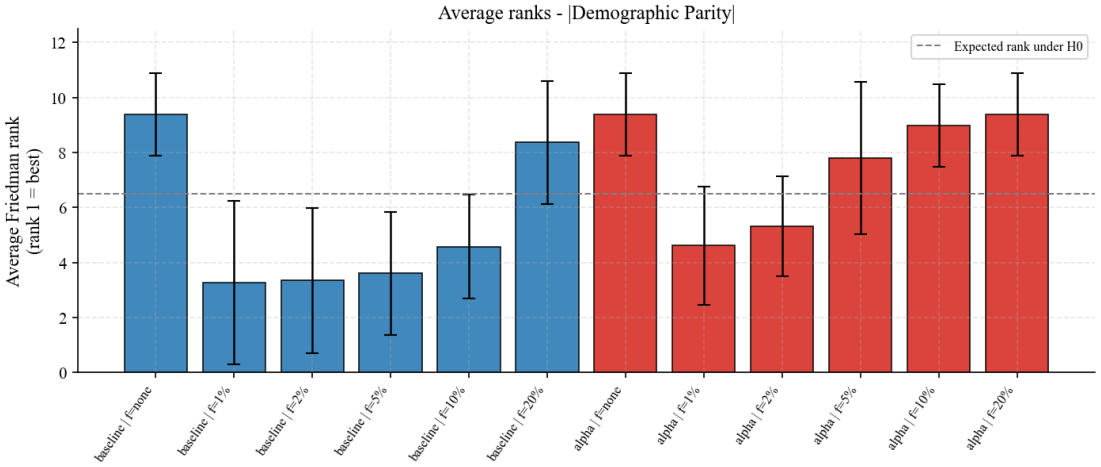


Figure 5.13: Average Friedman ranks for |demographic parity| under the Jo attack. Blue = baseline cap mode, red = alpha cap mode. Rank 1 = best (lowest parity). The dashed line marks the expected rank under H_0.

Table 5.6: Wilcoxon signed-rank test results for the Jo attack on MaxEnt: baseline vs. alpha cap mode at each fairness fraction ($\alpha = 0.05$).

Metric	**Comparison**	**Statistic**	***p*-value**	**Significant**
Accuracy	Baseline vs Alpha (f=1%)	13.0	< 0.001	Yes
Accuracy	Baseline vs Alpha (f=2%)	9.0	< 0.001	Yes
Accuracy	Baseline vs Alpha (f=5%)	5.0	< 0.001	Yes
Accuracy	Baseline vs Alpha (f=10%)	46.0	0.009	Yes
Accuracy	Baseline vs Alpha (f=20%)	0.0	0.043	Yes
|Dem. parity|	Baseline vs Alpha (f=1%)	30.0	< 0.001	Yes
|Dem. parity|	Baseline vs Alpha (f=2%)	28.0	< 0.001	Yes
|Dem. parity|	Baseline vs Alpha (f=5%)	32.0	< 0.001	Yes
|Dem. parity|	Baseline vs Alpha (f=10%)	0.0	< 0.001	Yes
|Dem. parity|	Baseline vs Alpha (f=20%)	0.0	0.043	Yes

For both demographic parity and accuracy, the gap between the baseline and the alpha-cap mode narrows substantially as the fairness fraction increases. At $f = 1\%$, the attack increases disparity on average from 0.063 to 0.187 (an absolute difference of 0.124), whereas at $f = 20\%$ this difference reduces to 0.044. All comparisons remain statistically significant although the difference at higher relaxation levels are negligible in practice. This trend is consistent with the near-convergence observed

in Figure 5.11.

Accuracy exhibits the same overall trend, with the gap between conditions shrinking as the fairness constraint is relaxed. Taken together, these results indicate that the Jo attack is most effective under tight fairness constraints, and its practical impact on both fairness and accuracy diminishes as the tolerance parameter increases.

5.3.4 Model Robustness Comparison

Figure 5.14 compares MaxEnt (solid) and Zafar (dashed) under the alpha cap mode across all datasets and fairness fractions.

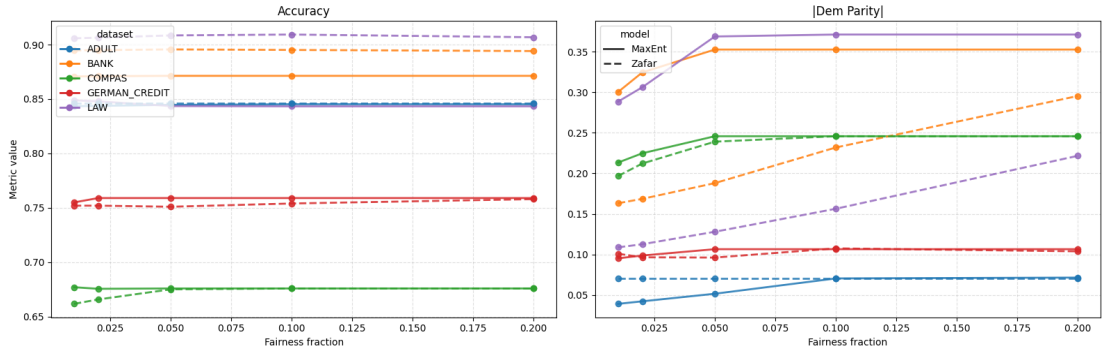


Figure 5.14: Comparison of MaxEnt (solid) and Zafar (dashed) under the Jo attack across fairness fractions, shown per dataset. Left: accuracy. Right: |demographic parity|.

The comparison between MaxEnt and Zafar under the Jo attack reveals a more nuanced picture than under the Solans attack. For accuracy, both models remain globally comparable across most datasets and fairness fractions. The largest differences appear on the Law school and Bank marketing datasets, where Zafar achieves a higher accuracy. However, no consistent global trend seems to emerge across datasets, and these larger gaps may partly reflect the undersampling procedure applied to these two datasets, which substantially modified their effective training distributions.

The demographic parity results are similarly mixed and do not show the clear dominance observed previously under the Solans attack. On Adult, Zafar achieves slightly lower demographic parity at the strictest fairness constraints, although both models converge toward similar values as the fairness tolerance increases. Conversely, MaxEnt appears to maintain lower demographic parity on COMPAS for the lower fairness fractions. The behaviour on Law school and Bank marketing

is again less straightforward. MaxEnt exhibits both lower accuracy and higher demographic parity than Zafar, a result that may appear counter-intuitive given the expected accuracy–fairness trade-off seen so far.

Overall, unlike the Solans attack where MaxEnt consistently outperformed Zafar in terms of fairness preservation, the Jo attack produces dataset-dependent outcomes with no universally dominant model. This suggests that the flipping attack, which operates directly on the sensitive attribute labels rather than through gradient-based feature manipulation, interacts differently with the underlying fairness formulations depending on the statistical structure and class imbalance of each dataset.

The Wilcoxon tests (Table 5.7) add some nuance to the visual comparison. For demographic parity, the difference between MaxEnt and Zafar is significant across all fairness fractions, which may seem at odds with the mixed picture suggested by the evolution figure. However, the direction of the difference is not uniform across datasets, so statistical significance here likely reflects consistent but opposing patterns rather than a clear global winner. For accuracy, significance only emerges at looser fairness fractions (f=5% and above), where Zafar's accuracy advantage becomes large enough to be detectable. At tight constraints the two models perform similarly in terms of accuracy, consistent with the visual overlap observed in Figure 5.14. Overall these results reinforce the earlier conclusion that the Jo attack produces a more complex and dataset-dependent outcome than the Solans attack, and that the statistically significant differences observed here should be interpreted carefully given the heterogeneity of the underlying per-dataset effects.

Table 5.7: Wilcoxon signed-rank test results for the robustness comparison between MaxEnt and Zafar under the Jo attack ($\alpha = 0.05$).

Metric	**Comparison**	**Statistic**	p**-value**	**Significant**
|Dem. parity|	MaxEnt vs Zafar (alpha, f=1%)	67.0	0.0088	Yes
|Dem. parity|	MaxEnt vs Zafar (alpha, f=2%)	69.0	0.0105	Yes
|Dem. parity|	MaxEnt vs Zafar (alpha, f=5%)	58.0	0.0038	Yes
|Dem. parity|	MaxEnt vs Zafar (alpha, f=10%)	21.0	0.0029	Yes
|Dem. parity|	MaxEnt vs Zafar (alpha, f=20%)	10.0	0.0027	Yes
Accuracy	MaxEnt vs Zafar (alpha, f=1%)	86.0	0.0674	No
Accuracy	MaxEnt vs Zafar (alpha, f=2%)	91.0	0.0918	No
Accuracy	MaxEnt vs Zafar (alpha, f=5%)	72.0	0.0258	Yes
Accuracy	MaxEnt vs Zafar (alpha, f=10%)	29.5	0.0084	Yes
Accuracy	MaxEnt vs Zafar (alpha, f=20%)	12.0	0.0038	Yes

5.4 Summary

This section synthesises the findings across both attack frameworks with respect to the two research questions.

Does the attack significantly affect MaxEnt? Both attacks produce statistically significant effects on MaxEnt, but the nature of these effects differs greatly. Under the Solans attack, accuracy degrades monotonically, yet the impact on demographic parity appears erratic, with sign reversals and a lack of aggregate significance at 15% poisoning budgets. This irregularity likely stems from the fact that the Solans attack is a pioneering work designed primarily to prove the *feasibility* of fairness poisoning rather than to offer a robust, optimized tool. Its reliance on a single-point injection optimized via a Logistic Regression surrogate might create a transferability gap when applied to MaxEnt, leading to instability seen in the results. In contrast, the Jo attack represents a more modern, highly optimized approach specifically derived to exploit the fairness constraint structure. This methodological advancement yields a more predictable and stable pattern, with effects most pronounced at tight fairness tolerances, offering more consistent and interpretable vulnerability insights than early gradient-based transfer attacks.

Which model is more robust? The results suggest that the answer may depend on the attack. Under the Solans attack, MaxEnt appears to better preserve its demographic parity guarantee compared to Zafar, and this difference is statistically significant across all poisoning fractions. Under the Jo attack the picture is less conclusive, with per-dataset variation making it difficult to identify a consistently dominant model. This could suggest that MaxEnt's apparent advantage may be more specific to feature-space attacks, while label-flipping attacks might interact with the two fairness formulations in a more dataset-dependent way.

Overall. Taken together, these results suggest that the type of poisoning attack may matter as much as the choice of fair model, though the limited number of datasets and the fixed experimental conditions make it difficult to draw strong conclusions. The findings motivate further investigation into whether the observed differences are robust across a wider range of settings, and whether defences could be designed to address the specific vulnerabilities suggested here.

Conclusion

This dissertation set out to investigate a critical, yet underexplored, weakness in algorithmic fairness: the robustness of fairness-constrained classifiers against data poisoning attacks. As automated decision-making systems become increasingly mandated to comply with fairness regulations, the assumption that training data is trustworthy has become a potential single point of failure. By conducting a comparative empirical study of two prominent in-processing approaches, the maximum-entropy logistic regression framework of Vancompernolle et al. [59] and the decision-boundary covariance model of Zafar et al. [66] and the unconstrained logistic regression baseline, this work has demonstrated that the structural formulation of fairness constraints significantly influences a model's resilience to adversarial manipulation.

Summary of Findings

The experimental results, derived from five real-world benchmark datasets and two distinct attack frameworks, yield four primary insights:

1. **Structural robustness of Maximum Entropy:** the MaxEnt formulation, which enforces fairness constraints directly on probabilistic outputs, demonstrated superior robustness against gradient-based feature-space attacks [55]. While the Zafar model maintained stable predictive accuracy under attack, it suffered a statistically significant degradation in demographic parity, effectively and silently failing its fairness mandate. In contrast, the MaxEnt model preserved its fairness guarantees across increasing poisoning budgets, albeit at the cost of a detectable drop in predictive accuracy. This suggests that MaxEnt offers a more secure trade-off: it sacrifices performance to maintain equity, whereas the Zafar model risks hiding discrimination behind stable accuracy metrics.

2. **Directional consistency and attack mechanisms:** a notable finding across both attack frameworks is their tendency to push demographic parity

monotonically in a single direction, suggesting that both gradient-based feature manipulation and optimal label flipping successfully exploit a dominant discriminatory pathway inherent to the data, even though the specific direction of this push varied across datasets and attacks. Under the Solans attack, the MaxEnt formulation demonstrated clear structural superiority in preserving fairness. In contrast, under the Jo attack, no single model offered universal immunity. Robustness became distinctly dataset-dependent. Notably, the Zafar model exhibited a marked performance advantage on the Bank and Law datasets, which underwent undersampling to mitigate class imbalance. This suggests that the Zafar formulation may interact more favorably with the altered statistical distributions resulting from preprocessing. Overall, this indicates that while both attacks leverage the same underlying directional vulnerability, the specific interaction between the attack vector (features vs. labels), the model's fairness formulation, and the data preprocessing history determines the final resilience.

3. **The Accuracy-Fairness-Security triangle:** the results reinforce the notion that security is an intrinsic dimension of the fairness-accuracy trade-off. An adversary can exploit the mathematical incompatibility of fairness criteria (the impossibility results) to force a model into a state where it appears accurate but behaves discriminatorily. The Solans attack successfully demonstrated this "silent failure" mode on the Zafar model, highlighting that monitoring accuracy alone is insufficient for auditing fair systems in production.

4. **Fairness constraints as the source of vulnerability:** the most foundational finding of this work is that neither attack produces a statistically significant effect on the unconstrained logistic regression baseline, for either predictive accuracy or demographic parity. This establishes that the attack surface exploited by both poisoning frameworks is not an inherent property of logistic regression as a classifier family, but is introduced exclusively by the fairness enforcement mechanism. Any robustness cost measured in the constrained models is therefore directly attributable to the act of enforcing fairness itself, rather than to any general sensitivity of the underlying model to data corruption.

Implications for Practice and Policy

These findings have immediate implications for practitioners and regulators deploying fairness-aware machine learning:

- **Model selection matters:** the choice of fairness mechanism is not merely

a hyperparameter tuning exercise, it is a security decision. For high-stakes applications where data provenance cannot be fully guaranteed (e.g., user-generated content or third-party data sources), the MaxEnt formulation appears to offer a safer default due to its resistance to feature-space poisoning.

- **Multi-Metric auditing is mandatory:** reliance on a single fairness metric (e.g., demographic parity) or a single performance metric (accuracy) creates blind spots. Defenders must adopt multi-criteria auditing protocols that simultaneously monitor disparate impact, equalized odds, and accuracy trends to detect the asymmetric effects of poisoning attacks.
- **Data provenance and validation:** the success of even minimal poisoning budgets (1%–5%) underscores the necessity of rigorous data validation pipelines. Techniques such as robust statistics, outlier detection, and provenance tracking must be integrated into the preprocessing stages of fair ML workflows.
- **Choice of adversarial benchmark:** when evaluating the robustness of new fairness algorithms, researchers must prioritize modern, optimized attack frameworks over early proof-of-concept methods. As demonstrated by the instability of the Solans attack at high budgets, relying on pioneering surrogate-based attacks may yield noisy or misleading robustness claims.
- **Fairness enforcement introduces an attack surface:** the baseline results make clear that deploying any fairness constraint creates a vulnerability that did not previously exist in the unconstrained model. Practitioners must therefore treat the choice to enforce fairness not only as a normative decision but as a security decision, one that requires anticipating how the constraint mechanism itself can be exploited by an adversary.

Future Work

While this study provides a comprehensive comparative analysis of fairness-constrained classifiers under poisoning attacks, several limitations define the scope of the findings and outline clear directions for future research.

Scope of fairness metrics and task types. First, this work focused exclusively on binary classification tasks and enforced demographic parity as the sole fairness criterion. While demographic parity is widely used in regulatory contexts, it is mathematically incompatible with other notions of fairness, such as equalized odds or calibration, particularly when base rates differ across groups. Future research should extend this robustness analysis to multi-class classification problems and

evaluate whether the structural advantages of the Maximum Entropy framework persist when enforcing error-rate-based constraints (e.g., equalized odds) or individual fairness metrics. Determining whether the observed robustness is specific to covariance-based constraints or generalizes to other mathematical formulations of fairness is a critical open question.

Adaptive and White-Box threat models. Second, the attacks evaluated in this thesis operated under a black-box or surrogate-model assumption, where the adversary optimizes against a standard logistic regression rather than the specific target architecture. While this reflects a realistic initial threat scenario, it does not account for *adaptive* adversaries who possess full knowledge of the defense mechanism (i.e., white-box attacks). An attacker aware of the MaxEnt objective function could potentially derive gradients specific to the entropy regularization term to craft more potent poisoning points. Future work should implement adaptive attack strategies that explicitly incorporate the defender's loss function, providing a more rigorous upper bound on the models' vulnerability.

Limited scope of attack vectors. Third, this work limited itself to two attack models. This scope does not encompass more nuanced or distinct threat models, such as conditional backdoor attacks like the Fairness Trojan from Xue et al. [64], clean-label poisoning, or inference-time evasion strategies. Consequently, the resilience of in-processing methods against this broader spectrum of adversarial behaviors remains unassessed within the current experimental framework. Only by testing against this broader spectrum of adversarial behaviors can we ensure that the resilience of in-processing methods is not merely an artifact of the specific attack vector chosen, but a genuine property of the defense.

Furthermore, while this study identified that both attacks push demographic parity monotonically (sometimes crossing from discriminating one group to discriminating the other) and the instability of the Solans attack warrants deeper research. Future work should investigate whether this behavior is a predictable function of the dataset's initial baseline disparity and the attack's magnitude and not an artifact of the specific attack vector chosen.

From assessment to mitigation. Finally, while this thesis identifies vulnerabilities, it does not propose novel defense mechanisms. The results suggest that standard training procedures are insufficient against even low-budget poisoning. Future research must move beyond vulnerability assessment to active mitigation. This includes developing robust optimization techniques tailored to the MaxEnt framework, such as trimming suspicious training samples, employing certified defenses, or integrating data provenance verification directly into the fairness-constrained

optimization loop. Bridging the gap between identifying these structural weaknesses and engineering resilient algorithms remains the most critical next step for deploying trustworthy fair machine learning systems.

Final Thoughts

Algorithmic fairness is not a static property that can be "solved" once during training, it is a dynamic state that must be maintained against active opposition. As this thesis has shown, the mathematical elegance of a fairness constraint does not guarantee its survival in an adversarial world. Indeed, the very act of enforcing fairness creates the conditions for its subversion. Ensuring that automated systems remain equitable therefore requires not only sophisticated optimization algorithms but also a security-first mindset that anticipates and withstands attempts to subvert the very principles of justice they are designed to uphold.

Bibliography

[1] Julia Angwin et al. *Machine Bias*. ProPublica. May 23, 2016. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[2] Solon Barocas and Andrew D. Selbst. "Big Data's Disparate Impact". In: *SSRN Electronic Journal* (2016). ISSN: 1556-5068. DOI: 10.2139/ssrn.2477899.

[3] Marco Barreno et al. "Can Machine Learning Be Secure?" In: *Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security*. ASIACCS '06. Association for Computing Machinery, Mar. 21, 2006, pp. 16–25. ISBN: 978-1-59593-272-3. DOI: 10.1145/1128817.1128824.

[4] Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. "A Maximum Entropy Approach to Natural Language Processing". In: *Comput. Linguist.* 22.1 (Mar. 1, 1996), pp. 39–71. ISSN: 0891-2017.

[5] Battista Biggio, Blaine Nelson, and Pavel Laskov. "Poisoning Attacks against Support Vector Machines". In: *Proceedings of the 29th International Coference on International Conference on Machine Learning*. ICML'12. Omnipress, June 26, 2012, pp. 1467–1474. ISBN: 978-1-4503-1285-1.

[6] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, July 2006. ISBN: 978-0-387-31073-2.

[7] Jérémy de Bodt et al. "Fairness in Machine Learning: A Compact Survey". In: *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2026)*. i6doc.com, Apr. 22, 2026. ISBN: 9782875870964. URL: http://www.i6doc.com/en/.

[8] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. ISBN: 0521833787.

[9] Simon Caton and Christian Haas. "Fairness in Machine Learning: A Survey". In: *ACM Computing Surveys* 56.7 (July 31, 2024), pp. 1–38. ISSN: 0360-0300, 1557-7341. DOI: 10.1145/3616865.

[10] Eunice Chan and Hanghang Tong. *Adversarial Bias: Data Poisoning Attacks on Fairness*. Nov. 1, 2025. DOI: 10.48550/arXiv.2511.08331.

[11] Alexandra Chouldechova and Aaron Roth. *The Frontiers of Fairness in Machine Learning*. Oct. 20, 2018. DOI: 10.48550/arXiv.1810.08810.

[12] Benoît Colson, Patrice Marcotte, and Gilles Savard. "An Overview of Bilevel Optimization". In: *Annals of Operations Research* 153.1 (Sept. 1, 2007), pp. 235–256. ISSN: 1572-9338. DOI: 10.1007/s10479-007-0176-2.

[13] Sam Corbett-Davies et al. "Algorithmic Decision Making and the Cost of Fairness". In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '17. Association for Computing Machinery, Aug. 4, 2017, pp. 797–806. ISBN: 978-1-4503-4887-4. DOI: 10.1145/3097983.3098095.

[14] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience. ISBN: 0471241954.

[15] Pádraig Cunningham, Matthieu Cord, and Sarah Jane Delany. "Supervised Learning". In: *Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval*. Springer, 2008, pp. 21–49. ISBN: 978-3-540-75171-7. DOI: 10.1007/978-3-540-75171-7_2.

[16] Inc. CVX Research. *CVX: Matlab Software for Disciplined Convex Programming, version 2.0*. https://cvxr.com/cvx. Aug. 2012.

[17] Ambra Demontis et al. "Why Do Adversarial Attacks Transfer? Explaining Transferability of Evasion and Poisoning Attacks". In: *Proceedings of the 28th USENIX Conference on Security Symposium*. SEC'19. USENIX Association, Aug. 14, 2019, pp. 321–338. ISBN: 978-1-939133-06-9.

[18] Janez Demsar. "Statistical Comparisons of Classifiers over Multiple Data Sets". In: *Journal of Machine Learning Research* 7 (Jan. 1, 2006), pp. 1–30.

[19] Michele Donini et al. "Empirical Risk Minimization under Fairness Constraints". In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NeurIPS '18. Curran Associates Inc., Dec. 3, 2018, pp. 2796–2806.

[20] Cynthia Dwork et al. "Fairness through Awareness". In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ITCS '12. Association for Computing Machinery, Jan. 8, 2012, pp. 214–226. ISBN: 978-1-4503-1115-1. DOI: 10.1145/2090236.2090255.

[21] Michael Feldman et al. "Certifying and Removing Disparate Impact". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '15. Association for Computing Machinery, Aug. 10, 2015, pp. 259–268. ISBN: 978-1-4503-3664-2. DOI: 10.1145/2783258.2783311.

[22] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. "All Models Are Wrong, but *Many* Are Useful: Learning a Variable's Importance by Studying

an Entire Class of Prediction Models Simultaneously". In: *Journal of machine learning research* 20 (Jan. 1, 2019), p. 177. ISSN: 1533-7928.

[23] Sorelle A. Friedler, Scheidegger Carlos, and Venkatasubramanian Suresh. "The (Im)Possibility of Fairness". In: *Communications of the ACM* 64.4 (Mar. 2021), pp. 136–143. ISSN: 0001-0782. DOI: 10.1145/3433949.

[24] Milton Friedman. "The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance". In: *Journal of the American Statistical Association* 32.200 (1937), pp. 675–701. DOI: 10.2307/2279372.

[25] Alex Gittens, Bülent Yener, and Moti Yung. "An Adversarial Perspective on Accuracy, Robustness, Fairness, and Privacy". In: *IEEE Access* 10 (2022), pp. 120850–120865. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2022.3218715.

[26] M. Grant and S. Boyd. "Graph implementations for nonsmooth convex programs". In: *Recent Advances in Learning and Control*. Ed. by V. Blondel, S. Boyd, and H. Kimura. Lecture Notes in Control and Information Sciences. http://stanford.edu/~boyd/graph_dcp.html. Springer-Verlag Limited, 2008, pp. 95–110.

[27] Moritz Hardt, Eric Price, and Nathan Srebro. "Equality of Opportunity in Supervised Learning". In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS'16. Curran Associates Inc., Dec. 5, 2016, pp. 3323–3331. ISBN: 978-1-5108-3881-9.

[28] Charles R. Harris et al. "Array programming with NumPy". In: *Nature* 585.7825 (Sept. 2020), pp. 357–362. DOI: 10.1038/s41586-020-2649-2. URL: https://doi.org/10.1038/s41586-020-2649-2.

[29] Trevor Hastie, Jerome Friedman, and Robert Tibshirani. "Unsupervised Learning". In: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2001, pp. 437–508. ISBN: 978-0-387-21606-5. DOI: 10.1007/978-0-387-21606-5_14.

[30] Daniel Ho and Alice Xiang. *Affirmative Algorithms: The Legal Grounds for Fairness as Awareness*. Dec. 18, 2020. DOI: 10.48550/arXiv.2012.14285.

[31] Max Hort et al. "Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey". In: *ACM J. Responsib. Comput.* 1.2 (June 20, 2024), 11:1–11:52. DOI: 10.1145/3631326.

[32] Matthew Jagielski et al. "Subpopulation Data Poisoning Attacks". In: *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. CCS '21. Association for Computing Machinery, Nov. 13, 2021, pp. 3104–3122. ISBN: 978-1-4503-8454-4. DOI: 10.1145/3460120.3485368.

[33] E. T. Jaynes. "Information Theory and Statistical Mechanics". In: *Physical Review* 106.4 (May 15, 1957), pp. 620–630. DOI: 10.1103/PhysRev.106.620.

[34] Changhun Jo, Jy-Yong Sohn, and Kangwook Lee. "Breaking Fair Binary Classification with Optimal Flipping Attacks". In: *2022 IEEE International Sym-*

posium on Information Theory (ISIT). IEEE Press, June 26, 2022, pp. 1453–1458. DOI: 10.1109/ISIT50566.2022.9834475.

[35] Faisal Kamiran and Toon Calders. "Data Preprocessing Techniques for Classification without Discrimination". In: *Knowledge and Information Systems* 33.1 (Oct. 1, 2012), pp. 1–33. ISSN: 0219-3116. DOI: 10.1007/s10115-011-0463-8.

[36] Michael Kelly, Rachel Longjohn, and Kolby Nottingham. *The UCI Machine Learning Repository*. https://archive.ics.uci.edu. 2017.

[37] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. "Inherent Trade-Offs in the Fair Determination of Risk Scores". In: *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*. Vol. 67. Leibniz International Proceedings in Informatics (LIPIcs). Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2017, 43:1–43:23. ISBN: 978-3-95977-029-3. DOI: 10.4230/LIPIcs.ITCS.2017.43.

[38] Jon Kleinberg et al. "Algorithmic Fairness". In: *AEA Papers and Proceedings* 108 (May 1, 2018), pp. 22–27. ISSN: 2574-0768, 2574-0776. DOI: 10.1257/pandp.20181018.

[39] Pang Wei Koh, Jacob Steinhardt, and Percy Liang. "Stronger Data Poisoning Attacks Break Data Sanitization Defenses". In: *Mach. Learn.* 111.1 (Jan. 1, 2022), pp. 1–47. ISSN: 0885-6125. DOI: 10.1007/s10994-021-06119-y.

[40] Ron Kohavi. *Census Income*. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/1996.

[41] Tai Le Quy et al. "A Survey on Datasets for Fairness-Aware Machine Learning". In: *WIREs Data Mining and Knowledge Discovery* 12.3 (2022), e1452. ISSN: 1942-4795. DOI: 10.1002/widm.1452.

[42] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning". In: *Journal of Machine Learning Research* 18.17 (2017), pp. 1–5. URL: http://jmlr.org/papers/v18/16-365.html.

[43] Dong C. Liu and Jorge Nocedal. "On the limited memory BFGS method for large scale optimization". In: *Math. Program.* 45.1–3 (Aug. 1989), pp. 503–528. ISSN: 0025-5610.

[44] Fernando Martínez-Plumed et al. *Fairness and Missing Values*. May 29, 2019. DOI: 10.48550/arXiv.1905.12728.

[45] Wes McKinney. "Data Structures for Statistical Computing in Python". In: *Proceedings of the 9th Python in Science Conference*. Ed. by Stéfan van der Walt and Jarrod Millman. 2010, pp. 56–61. DOI: 10.25080/Majora-92bf1922-00a.

[46] Kristof Meding and Thilo Hagendorff. "Fairness Hacking: The Malicious Practice of Shrouding Unfairness in Algorithms". In: *Philosophy & Technology*

37.1 (Jan. 6, 2024), p. 4. ISSN: 2210-5441. DOI: 10.1007/s13347-023-00679-8.

[47] Ninareh Mehrabi et al. "Exacerbating Algorithmic Bias through Fairness Attacks". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35.10 (May 18, 2021), pp. 8930–8938. ISSN: 2374-3468. DOI: 10.1609/aaai.v35i10.17080.

[48] Marco Melis et al. "secml: A Python Library for Secure and Explainable Machine Learning". In: *arXiv preprint arXiv:1912.10013* (2019).

[49] S. Moro, P. Rita, and P. Cortez. *Bank Marketing*. UCI Machine Learning Repository. 2014. DOI: 10.24432/C5K306. URL: https://doi.org/10.24432/C5K306.

[50] Kevin P Murphy. *Machine learning: a probabilistic perspective*. Cambridge, MA: MIT press, 2012. ISBN: 978-0-262-01802-9.

[51] Peter Bjorn Nemenyi. *Distribution-free multiple comparisons.* Princeton University, 1963.

[52] Fabian Pedregosa et al. "Scikit-Learn: Machine Learning in Python". In: *J. Mach. Learn. Res.* 12 (null Nov. 1, 2011), pp. 2825–2830. ISSN: 1532-4435.

[53] Dana Pessach and Erez Shmueli. "Algorithmic Fairness". In: *Machine Learning for Data Science Handbook: Data Mining and Knowledge Discovery Handbook*. Springer International Publishing, 2023, pp. 867–886. ISBN: 978-3-031-24628-9. DOI: 10.1007/978-3-031-24628-9_37.

[54] ProPublica. *COMPAS Recidivism Risk Score Data and Analysis*. https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis. 2023.

[55] David Solans, Battista Biggio, and Carlos Castillo. "Poisoning Attacks on Algorithmic Fairness". In: *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part I*. Springer-Verlag, Sept. 14, 2020, pp. 162–177. ISBN: 978-3-030-67657-5. DOI: 10.1007/978-3-030-67658-2_10.

[56] The pandas development team. *pandas-dev/pandas: Pandas*. Version latest. Feb. 2020. DOI: 10.5281/zenodo.3509134. URL: https://doi.org/10.5281/zenodo.3509134.

[57] Zhiyi Tian et al. "A Comprehensive Survey on Poisoning Attacks and Countermeasures in Machine Learning". In: *ACM Comput. Surv.* 55.8 (Dec. 23, 2022), 166:1–166:35. ISSN: 0360-0300. DOI: 10.1145/3551636.

[58] Minh-Hao Van et al. "Poisoning Attacks on Fair Machine Learning". In: *Database Systems for Advanced Applications: 27th International Conference, DASFAA 2022, Virtual Event, April 11–14, 2022, Proceedings, Part I.*

Springer-Verlag, Apr. 11, 2022, pp. 370–386. ISBN: 978-3-031-00122-2. DOI: 10.1007/978-3-031-00123-9_30.

[59] Flore Vancompernolle Vromman et al. "Maximum Entropy Logistic Regression for Demographic Parity in Supervised Classification". In: *Artificial Intelligence and Machine Learning*. Vol. 2187. Springer Nature Switzerland, 2025, pp. 189–208. ISBN: 978-3-031-74649-9 978-3-031-74650-5. DOI: 10.1007/978-3-031-74650-5_11.

[60] Flore Vancompernolle Vromman et al. "Fairness-Aware Post-Processing in Supervised Classification: L1/L2 Norm and Optimal Swapping Methods". In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 8.3 (Oct. 15, 2025), pp. 2563–2574. ISSN: 3065-8365. DOI: 10.1609/aies.v8i3.36738.

[61] Martin J. Wainwright and Michael I. Jordan. "Graphical Models, Exponential Families, and Variational Inference". In: *Found. Trends Mach. Learn.* 1.1–2 (Jan. 1, 2008), pp. 1–305. ISSN: 1935-8237. DOI: 10.1561/2200000001.

[62] Linda F. Wightman. *LSAC National Longitudinal Bar Passage Study*. LSAC Research Report Series. Law School Admission Council, 1998, pp. 1–112.

[63] Frank Wilcoxon. "Individual Comparisons by Ranking Methods". In: *Biometrics Bulletin* 1.6 (1945), pp. 80–83. DOI: 10.2307/3001968.

[64] Jiaqi Xue et al. "TrojFair: Trojan Fairness Attacks". In: *Proceedings of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis*. LAMPS '24. Association for Computing Machinery, Nov. 19, 2024, pp. 47–56. ISBN: 979-8-4007-1209-8. DOI: 10.1145/3689217.3690620.

[65] S. Yadav and S. Shukla. "Analysis of k-fold Cross-validation over Hold-out Validation on Colossal Datasets for Quality Classification". In: *Proceedings of the IEEE 6th International Conference on Advanced Computing (IACC '16)*. IEEE, 2016, pp. 78–83.

[66] Muhammad Bilal Zafar et al. "Fairness Constraints: Mechanisms for Fair Classification". In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. Ed. by Aarti Singh and Jerry Zhu. Vol. 54. Proceedings of Machine Learning Research. PMLR, Apr. 20, 2017, pp. 962–970. URL: https://proceedings.mlr.press/v54/zafar17a.html.

[67] Mengchen Zhao et al. "Efficient Label Contamination Attacks against Black-Box Learning Models". In: *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. IJCAI'17. AAAI Press, Aug. 19, 2017, pp. 3945–3951. ISBN: 978-0-9992411-0-3.

[68] Pinlong Zhao et al. *Data Poisoning in Deep Learning: A Survey*. Mar. 1, 2025. DOI: 10.48550/arXiv.2503.22759.

Appendix

.1 Additional Tables

Table 8: Class imbalance statistics and trivial classifier behaviour for all five datasets. Trivial classifier always predicts the majority class. Datasets below the resampling threshold of 0.15 are marked with (†).

Dataset	N	$Y=1$	$Y=0$	$\%Y=1$	Y**-ratio**	**Trivial acc.**
Bank marketing†	45,211	5,289	39,922	11.70%	0.132	0.8830
Law school†	20,798	19,758	1,040	95.00%	0.053	0.9500
Adult	45,222	7,508	37,714	16.60%	0.199	0.8340
COMPAS	6,150	3,283	2,867	53.38%	0.873	0.5338
German credit	1,000	700	300	70.00%	0.429	0.7000

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
École polytechnique de Louvain
Rue Archimède, 1 bte L6.11.01, 1348 Louvain-la-Neuve, Belgique | **www.uclouvain.be/epl**