Survey on PAC Learnability and Fairness: A Unified Framework from Cardinal Welfare and Metric Fairness

Arnav Gattani
University of Pennsylvania, CIS
agattani@seas.upenn.edu

Akash Anickode
University of Pennsylvania, CIS
aanickod@sas.upenn.edu

Vadim Popov
University of Pennsylvania, CIS
vadmipop@seas.upenn.edu

Abstract

This paper attempts to incorporate fairness into the Probably Approximately Correct (PAC) learning framework. It integrates perspectives from three critical areas: individual fairness through probably approximately correct and fair (PACF) learning, group fairness based on cardinal welfare axioms, and the influence of adversarial data corruption on fairness assurances. Using a micro-lending dataset, the group-fairness framework is demonstrated in a practical context, through satisfying cardinal axioms and evaluating fairness metrics. The proposed Metric-Weighted Fairness Framework (MWFF) harmonizes individual and group fairness by combining metric-based fairness constraints with aggregate utility measures, utilizing weighted power-mean functions to enable flexible trade-offs between egalitarian and utilitarian fairness. The study underscores the MWFF's resilience and versatility while suggesting directions for future research, such as improving adversarial robustness and addressing domain-specific fairness needs. This work lays the groundwork for developing more equitable and scalable machine learning systems.

1. Introduction

Machine learning is quickly becoming the foundation for modern decision-making, leading to strides in numerous fields including healthcare, finance, and the justice system. However, this increased prevalence makes the concerns about the fairness of machine learning all the more pressing. Without explicitly addressing fairness during training, machine learning models risk perpetuating or even exacerbating biases found in the data, leading to discriminatory results during prediction [1]. This heavy risk warrants an assessment of the performance of existing frameworks in this regard, as fairness is crucial not only for ethical reasons, but also to build trust in the community around machine learning driven systems.

In order to address these concerns, researchers have constructed various methodologies for measuring and optimizing fairness of maching learning algorithms. Most of this work is centered on group fairness, wherein decisions made by an algorithm must not discriminate on the basis of protected attributes such as race, gender, or socioeconomic class. Such strategies include preprocessing data, modifying training protocols, and processing models after learning to fit certain fairness metrics [7].

However, these methodologies have limitations, as real-world datasets often suffer from biases, inaccuracies, and corruption. Furthermore, with the potential for adversarial manipulation of data, where malicious actors deliberately introduce biases to undermine fairness objectives, these issues can be aggravated [2]. While researchers have attempted to develop methods to mitigate the effects of various forms of corruption, the field lacks a comprehensive framework for guaranteeing fairness under arbitrary, adversarial data manipulations.

In this paper, we specifically explore the Probably Approximately Correct (PAC) learning framework from the lens of fairness described above. Introduced by Valiant in 1984 [10], PAC learning formalizes the conditions under which an algorithm can achieve a hypothesis that is both accurate and generalizable with high probability. Because of this overall objective, the PAC learning framework tends to perform well with balancing multiple objectives during learning. We begin by surveying a paper on the limits of

PAC learning, then move onto two papers approaching fairness under the PAC framework. The second paper goes into the notion of individual fairness while the third paper explains flaws of fairness constraints and how they can be resolved through implementing group fairness. We then show an implementation of the group fairness framework in the third paper through evaluating various ”cardinal axioms” that correspond to fairness metrics. We improve on this through proposing our own fairness framework by incorporating our earlier findings.

2. Analysis of existing results

2.1. Fairness-Aware PAC Learning from Corrupted Data [6]

This paper discusses the problem of fairness-aware PAC learning in the context of a malicious adversary. In the practical setting, datasets are often unreliable, whether due to noise, human bias, or a malicious agent. To model these situations, the authors consider an adversary capable of arbitrarily modifying a certain fraction of the dataset. Allowing this adversary to perform any type of data corruption creates a strong adversarial model that can provide a certificate of fairness. That is, if a system can perform well under this adversarial model, we can conclude that it will perform well under any situation of data corruption.

The paper uses the following standard group fairness classification framework: The product space is of the form $X \times A \times Y$, where X is an input space, Y is a binary classification label, and A is a binary protected attribute (i.e. being part of the majority or minority group). The clean data is assumed to be sampled from a true distribution P. Finally, denote the hypothesis space of classifiers as $\mathcal{H}$.

A fairness-aware learner $\mathcal{L}$ is a function that takes as input a dataset from the product space and outputs a hypothesis from the hypothesis space. The learner's performance is measured by the expected 0/1 loss with respect to the distribution P. Specifically,

$$\mathcal{R}(h, P) = \mathbb{P}_{(X,Y)\sim P}(h(X) \neq Y)$$

The paper also proposes to use the following methods for measuring the fairness of the classifier. First is demographic parity, which checks that the classifier's decisions are independent of the protected attribute. More formally:

$$\mathbb{P}(h(X) = 1 \mid A = 0) = \mathbb{P}(h(X) = 1 \mid A = 1)$$

The second proposed measure is an equal opportunity, which requires true positive rates to be consistent across protected groups. More formally:

$$\mathbb{P}(h(X) = 1 \mid A = 0, Y = 1) = \mathbb{P}(h(X) = 1 \mid A = 1, Y = 1)$$

Note that this definition assumes that $Y = 1$ is a positive result. The paper outlines that due to the rarity of achieving perfect fairness, the goal will instead be to control the amount of unfairness present in the hypothesis. Specifically, the mean difference score measure is used.

For demographic parity:

$$D^{par}(h, P) = |\mathbb{P}(h(X) = 1 \mid A = 0) - \mathbb{P}(h(X) = 1 \mid A = 1)|$$

For equal opportunity:

$$\begin{aligned} D^{opp}(h, P) = |&\mathbb{P}(h(X) = 1 \mid A = 0, Y = 1) - \\ &\mathbb{P}(h(X) = 1 \mid A = 1, Y = 1)| \end{aligned}$$

Also, define $P_a = \mathbb{P}(A = a) > 0$ and $P_{1a} = \mathbb{P}(Y = 1, A = a) > 0$ for $a \in \{0, 1\}$

Finally, the paper formally defines the malicious adversary model of data generation as follows:

- Sample an i.i.d clean dataset $S^c = \{(x_i^c, a_i^c, y_i^c)\}$ of points from P, indexed from 1 to n
- Independently mark each index with some probability $\alpha \in [0, 0.5)$, adding the marked indices to a set M.
- For all indices in M, the adversary is now allowed to replace the corresponding marked data points $\{(x_i^c, a_i^c, y_i^c)\}$ in an arbitrary manner with some $\{(x_i^p, a_i^p, y_i^p)\}$. Note that the adversary can do so with no assumptions whatsoever on the corrupted data points $\{(x_i^p, a_i^p, y_i^p)\}$.
- The corrupted dataset S^p is then given to the learner, which then computes $\mathcal{L}(S^p)$

It is important to note that since the adversary is acting under no assumptions on the corrupted data points, they are allowed to depend on the learner $\mathcal{L}$, the distribution P, or any other parameter of the problem. Specifically, this assumes that the adversary is working under full knowledge of the learning setup and no computational constraints.

The authors continue on to discuss how to measure the quality of a hypothesis in terms of both the expected loss $\mathcal{R}(\mathcal{L}(S^p), P)$ and its level of unfairness $D(\mathcal{L}(S^p), P)$. They propose two possible ways of doing so.

First is to predetermine a weight parameter λ and have the learner minimize:

$$Q_\lambda(h) = \mathcal{R}(h) + \lambda D(h)$$

This weight parameter λ can be assigned variably depending on the use case. Thus, the quality of the hypothesis $\mathcal{L}(S^p)$ can be measured by $Q_\lambda(\mathcal{L}(S^p)) - min_{h \in H} Q_\lambda(h)$

The second is to make comparisons between various hypotheses by defining $\beta(h) = (R(h), D(h))$. This induces the relation between two hypotheses h_1 and h_2 as $\beta(h_1) \leq \beta(h_2)$ if $R(h_1) \leq R(h_2)$ and $D(h_1) \leq D(h_2)$. Further, the authors assume that there exists a classifier h^* that is optimal under this relation. In other words, there exists $h^* \in \mathcal{H}$ such that $h^* \in argmin_{h \in \mathcal{H}} R(h)$ and $h^* \in argmin_{h \in \mathcal{H}} D(h)$ so that $\beta(h^*) \leq \beta(h)$ for all $h \in \mathcal{H}$. Thus, the quality of the hypothesis $\mathcal{L}(S^p)$ can be measured by $\beta(\mathcal{L}(S^p)) - \beta(h^*)$

The paper now moves to show lower and upper bounds as a series of hardness results to show that fair learning in the malicious adversary setting is provably impossible in a PAC-learning sense.

The following theorems are the results found in the paper. Here we will focus on identifying and discussing the results. For formal proofs of these theorems, refer to the original paper.

Theorem 1 *Let* $0 \leq \alpha < 0.5$, $0 < P_0 \leq 0.5$. *For any input set* $\mathcal{X}$ *with at least four distinct points, there exists a finite hypothesis space* $\mathcal{H}$*, such that for any learning algorithm* $\mathcal{L} : \bigcup_{n \in \mathbb{N}} (\mathcal{X} \times \mathcal{A} \times \mathcal{Y})^n \to \mathcal{H}$*, there exists a distribution* P *for which* $\mathbb{P}(A = 0) = P_0$*, a malicious adversary* $\mathcal{A}$ *of power* α *and a hypothesis* $h^* \in \mathcal{H}$*, such that with probability at least* 0.5

$$\mathcal{R}(\mathcal{L}(S^p), P) - \mathcal{R}(h^*, P) \geq \min\left\{\frac{\alpha}{1-\alpha}, 2P_0P_1\right\}$$

and

$$\mathcal{D}^{par}(\mathcal{L}(S^p), P) - \mathcal{D}^{par}(h^*, P) \geq \min\left\{\frac{\alpha}{2P_0P_1(1-\alpha)}, 1\right\}.$$

This result implies that there is no learner that can guarantee to reach the Pareto front of the accuracy-fairness optimization problem in the demographic parity case. In detail, this conclusion is due to the fact that the adversary can, with non vanishing probability, force the learner to return a hypothesis that is a constant factor away from optimality for both of the metrics. A similar conclusion can be made for the equal opportunity case as seen in Theorem 2 of the paper.

Theorem 3 *Let* $0 \leq \alpha < 0.5$, $0 < P_0 \leq 0.5$. *For any input set* $\mathcal{X}$ *with at least four distinct points, there exists a finite hypothesis space* $\mathcal{H}$*, such that for any learning algorithm* $\mathcal{L} : \bigcup_{n \in \mathbb{N}} (\mathcal{X} \times \mathcal{A} \times \mathcal{Y})^n \to \mathcal{H}$*, there exists a distribution* P *for which* $\mathbb{P}(A = 0) = P_0$*, a malicious adversary* $\mathcal{A}$ *of power* α *and a hypothesis* $h^* \in \mathcal{H}$*, such that with probability at least* 0.5

$$\mathcal{R}(\mathcal{L}(S^p), P) = \mathcal{R}(h^*, P) = \min_{h \in \mathcal{H}} \mathcal{R}(h, P)$$

and

$$\mathcal{D}^{par}(\mathcal{L}(S^p), P) - \mathcal{D}^{par}(h^*, P) \geq \min\left\{\frac{\alpha}{2P_0}, 1\right\}.$$

This theorem reveals a case in which the model has good accuracy but is still problematic. Specifically, it is shown that the adversary can force the learner to produce a model that is optimal in accuracy but contains a high amount of unfairness in the demographic parity case. A similar conclusion can be made for the equal opportunity case as seen in Theorem 4 of the paper.

Next, the paper discusses upper bounds to show that the lower bounds above are tight to constant factors. In particular, two types of fairness-aware algorithms are analyzed, specifically the two hypothesis quality measures discussed previously (i.e., the λ-weighted objective and the β hypothesis comparison objective).

Theorem 5 *Let* $\mathcal{H}$ *be any hypothesis space with* $d = VC(\mathcal{H}) < \infty$. *Let* $P \in \mathcal{P}(\mathcal{X} \times \mathcal{A} \times \mathcal{Y})$ *be a fixed distribution and let* $\mathcal{A}$ *be any malicious adversary of power* $\alpha < 0.5$. *Denote by* P^A *the probability distribution of the corrupted data* S^p*, under the random sampling of the clean data, the marked points, and the randomness of the adversary. Then for any* $\delta \in (0, 1)$ *and*

$$n \geq \max\left\{\frac{8\log(16/\delta)}{(1-\alpha)P_0}, \frac{12\log(12/\delta)}{\alpha}, \frac{d}{2}\right\}, \textit{ we have:}$$

$$P^A\left(L_\lambda^{par}(\hat{h}) \leq \min_{h \in \mathcal{H}} L_\lambda^{par}(h) + \Delta_\lambda^{par}\right) > 1 - \delta,$$

where $\hat{h} := \mathcal{L}_\lambda^{par}(S^p)$ *is the hypothesis returned by the learner,* $L_\lambda^{par}(h) = \mathcal{R}(h) + \lambda\mathcal{D}^{par}(h)$ *is the* λ*-weighted objective and*

$$\Delta_\lambda^{par} = 3\alpha + \lambda(2\Delta^{par}) + \tilde{\mathcal{O}}\left(\sqrt{\frac{d}{n}} + \lambda\sqrt{\frac{d}{P_0 n}}\right)$$

and

$$\Delta^{par} = \frac{2\alpha}{\frac{P_0}{3} + \alpha} = \mathcal{O}\left(\frac{\alpha}{P_0}\right).$$

This theorem reveals that, in the demographic parity case, for any hypothesis class $\mathcal{H}$ of finite VC-dimension, any distribution P, any adversary $\mathcal{A}$ of power α, and sufficiently large sample size, the learner can get the λ-weighted objective ($Q_\lambda(h) = \mathcal{R}(h) + \lambda D(h)$) below a certain threshold with high probability. In particular,

$$L_\lambda^{par}(\hat{h}) - \min_{h \in \mathcal{H}} L_\lambda^{par}(h) \leq \mathcal{O}\left(\alpha + \lambda\frac{\alpha}{P_0}\right)$$

A similar result is made for the equal opportunity case in Theorem 6.

The paper continues on to discuss some more upper bounds in the component-wise learning case, in which the authors define a set of hypothesis $\mathcal{H}_1$ that is not far from optimal in terms of accuracy and another set of hypothesis $\mathcal{H}_2$ that is not far from optimal in terms of fairness. Here, the learner aims to return a hypothesis in the intersection of $\mathcal{H}_1$ and $\mathcal{H}_2$. Theorems 7 and 8 provide more detail into this scenario.

Finally, Theorem 9 provides an interesting result with the added assumption that a perfectly accurate classifier exists. In this case, Theorem 9 shows that the model can converge to an order-optimal error in both accuracy and fairness at fast statistical rates.

From the lower bounds shown in the results, we can conclude that fair learning in the malicious adversarial model is impossible in the PAC learning setting. We are unable to obtain arbitrarily small errors due to these bounds. To develop accurate yet equitable models, the importance of clean data cannot be overstated. However, the authors suggest that strict data collection practices may make it possible to construct provably fair models in the PAC setting.

2.2. Probably Approximately Metric-Fair Learning [9]

This paper addresses limitations in existing fairness models in ML, specifically focusing on *individual fairness*, which means that every two individuals with similar features from the standpoint of the classification algorithm should receive similar predicted labels. This is a more granular version of the group-level version of fairness, which aims to eliminate statistical biases that could arise due to a particular group being underrepresented in the training dataset [4]. The paper distinguishes between the notions of *perfect metric-fairness*, which is computationally infeasible, and *approximate metric-fairness*, a more relaxed version of the former. The approach developed by [9] introduces approximate metric fairness, which generalizes fairness guarantees to the underlying population distribution rather than just the training dataset. The authors develop polynomial-time Probably Approximately Correct and Fair (PACF) learning algorithms for specific model classes, such as linear and logistic predictors, leveraging generalization bounds based on Rademacher complexity.

2.2.1. Key definitions and theorems

The paper is building off of [4], except they move away from using perfect metric fairness, which ensures that for any two similar individuals from the dataset, the predicted classification probabilities should be similar. While this objective ensures individual fairness, Rothblum et al. show it to be computationally intractable in certain settings and also point out the fact that this notion of fairness doesn't generalize well from the sampled dataset to the underlying distribution. A solution proposed by Rothblum et al. is to allow for a small probability δ of fairness failure (similarly to the PAC model, this represents the probability of an unrepresentative dataset draw), as well as a small additive slack γ in fairness when evaluating how close the output distributions for similar individuals must be.

Definition 2.1 (Approximate Metric-Fairness)**.** *A classifier h is (α, γ)-approximately metric-fair with respect to a similarity metric d and a data distribution $\mathcal{D}$ if:*

$$\mathcal{L}_\gamma^F \triangleq \mathbb{P}_{x,x' \sim \mathcal{D}} \left[|h(x) - h(x')| > d(x,x') + \gamma \right] \leq \alpha$$

This ensures fairness violations are bounded for most individual pairs, where α is the fraction of violations and γ is the slack in fairness. Importantly, approximate metric-fairness (AMF) guarantees that *every* group G of fractional size significantly larger than α is protected from discrimination, in the sense that members of G are treated similarly to similar individuals outside of G, on average. Equipped with this notion of AMF, the paper then defines the framework of PACF learning, which encapsulates the idea that the algorithm should attempt to maximize accuracy only after ensuring that it has met the fairness constraints, since we do not want to sacrifice fairness for marginal performance improvements.

Definition 2.2 (PACF Learning)**.** *A learning algorithm $\mathcal{A}$ PACF-learns a hypothesis class $\mathcal{H}$ if for every metric d and population distribution $\mathcal{D}$, every required fairness parameters $\alpha, \gamma \in [0, 1)$, every failure probability $\delta \in (0, 1)$, and every error parameters $\epsilon, \epsilon_\alpha, \epsilon_\gamma \in (0, 1)$, the following holds:*

There exists a sample complexity

$$m = poly\left(\frac{\log|\mathcal{X}| \cdot \log(1/\delta)}{\alpha \cdot \gamma \cdot \epsilon \cdot \epsilon_\alpha \cdot \epsilon_\gamma}\right)$$

and constants $\alpha', \gamma' \in [0, 1)$ (specified below), such that with all but δ probability over an i.i.d. sample of size m and $\mathcal{A}$'s coin tosses, the output predictor h satisfies the following two conditions:

*1. **Fairness:** h is (α, γ)-approximately metric-fair w.r.t. the metric d and the distribution $\mathcal{D}$.*

*2. **Accuracy:** Let $\mathcal{H}'_F$ denote the subclass of hypotheses in $\mathcal{H}$ that are $(\alpha' - \epsilon_\alpha, \gamma' - \epsilon_\gamma)$-approximately metric-fair, then:*

$$err_D(h) \leq \min_{h' \in \mathcal{H}'_F} err_D(h') + \epsilon$$

We say that $\mathcal{A}$ is efficient if it runs in time poly(m). If accuracy holds for $\alpha' = \alpha$ and $\gamma' = \gamma$, then we say that $\mathcal{A}$ is a strong PACF learning algorithm. Otherwise, we say that $\mathcal{A}$ is a relaxed PACF learning algorithm.

Now, it is important to note here that the main reason why the paper develops the AMF learning theory as opposed to just working with perfect MF learning is due to out-of-sample generalizability. The PACF learning definition asks that the learning algorithm performs almost as well as the best AMF algorithm, which, in lieu of the following result, means that the algorithm achieves good performance out-of-distribution while adhering to the MF standard.

Theorem 2.1 (Generalization of AMF)**.** *Let $\mathcal{H}$ be a hypothesis class with Rademacher complexity $R_m(\mathcal{H}) = r/\sqrt{m}$. For every $\delta \in (0,1)$ and every $\epsilon_\alpha, \epsilon_\gamma \in (0,1)$, there exists a sample complexity*

$$m = O\left(\frac{r^2 \cdot \ln(1/\delta)}{\epsilon_\alpha^2 \cdot \epsilon_\gamma^2}\right),$$

Such that the following holds:

With probability at least $1 - \delta$ over an i.i.d. sample $S \sim \mathcal{D}^m$, simultaneously for every $h \in \mathcal{H}$: if h is (α, γ)-approximately metric-fair on the sample S, then h is also $(\alpha + \epsilon_\alpha, \gamma + \epsilon_\gamma)$-approximately metric-fair on the underlying distribution $\mathcal{D}$.

In addition to generalization guarantees, the paper offers two more important results. First, after introducing the PACF learning model, they give algorithms for relaxed PACF learning in the setting of linear and logistic regression.

For linear regression, the predictor $h(x) = \frac{1+\langle w,x\rangle}{2}$ is optimized by relaxing the fairness constraint with an ℓ_1-based convex approximation, enabling efficient optimization using standard convex tools. Logistic regression, with $h(x) = \frac{1}{1+\exp(-4L\langle w,x\rangle)}$, handles the additional non-convexity of the sigmoid function by embedding the problem into a Reproducing Kernel Hilbert Space. While both approaches achieve relaxed PACF guarantees, logistic regression's complexity grows exponentially with the slope parameter L.

The final contribution of the paper is proof that under certain cryptographic assumptions, achieving perfect metric fairness is computationally intractable for some tasks, even when a perfectly fair and accurate classifier exists. Specifically, it shows that any polynomial-time learning algorithm constrained to perfect metric fairness will likely output a trivial, nearly random classifier with an error rate close to $\frac{1}{2}$.

To summarize, this work bridges individual and group fairness, offering a practical alternative to strict metric-based fairness by ensuring broad fairness protections without sacrificing computational feasibility.

2.3. Revisiting Fair-PAC Learning and the Axioms of Cardinal Welfare [3]

The previous papers focus primarily on individual fairness and adversarial robustness. It would be interesting to consider such fairness effects in a group setting. While an attempt to correct algorithmic bias by ensuring fairness or parity across demographics is important with ML's increased usage in modern society, Cousins' paper goes into several technical challenges that it labels as "flaws" of fairness constraints. Some of these are obvious, such as the trade-off between fairness and accuracy, however, some are surprising with studies demonstrating that welfare and utility can decrease if fairness constraints are enforced. Before diving into why this is the case, it is important to define what these terms mean and why they are worth analyzing:

2.3.1. Motivations for Cardinal Fairness Axioms

Utility, which must be $\geq$ 0, simply measures how much satisfaction it provides a user, similar to an economic setting. In contrast, malfare is a measure of general disunity. Therefore, we seek to maximize the utility and use malfare for fair loss minimization.

We formalize aggregating unity and disunity across different groups through "Cardinal Fairness", which is a set of axioms that seek to fairly compromise between multiple groups that send their data and feedback (their wants and needs). In the status quo, research is focused on minimizing malfare, such as minimizing work case risk, but does not deeply consider welfare optimization. We will motivate cardinal objectives and how they summarize unity and disunity across g groups using standard axioms.

2.3.2. Weighted Aggregator Functions Setup and Use Cases

We can express the overall sentiment (positive welfare or negative malfare) of a singular group using aggregator functions. The sentiment vector represents the sentiment/utility of a group, $u \in \mathbb{R}_+^g$, a vector with positive values (since utility is $\geq$ 0 and for g distinct groups. Each group is then weighted by the importance of their aggregated sentiment through weights $w \in \Delta_g$, ensuring weights are positive and that the sum of the weights of all g groups equals 1. For example, the utility vector $u =$ (0.3, 0.5, 0.9) represents utility for $g = 3$ groups with increasing satisfaction. The weight vector $w =$ (0.5, 0.25, 0.25) means group 1's satisfaction is weighted as much as group 2's and group 3's combined, and group 1 should be prioritized in the aggregation. Cousins defines generic aggregator functions, $M(u; w)$, to combine group utilities u and group weights w into a single measure of overall sentiment. He also adjacently defines $W(u; w)$ and $M(u; w)$

for welfare and malfare functions (positive and negative sentiments respectively). This is a useful representation since sentiment vectors directly correspond to outcomes of real-life scenarios, such as health outcomes for each group in the vector. The aggregator functions are used to set a cardinal preference ordering to the groups over the sentiment vectors. The function, therefore, assigns a calculated ranking to real-life (grounded) situations based on the utilities of the groups. This ordering describes which grounded situations are favorable based on the aggregate function values.

Therefore, any general algorithm can optimize the preference order by selecting the grounded situation that maximizes welfare or minimizes malfare. Furthermore, the algorithm does not have to explicitly look at the grounded situation and can objectively infer solely from each group's sentiments. Thus, the notion of fairness is not set in stone but is circumstantial based on how the sentiments are balanced among the groups. We can approach fairness with mathematical rigor while still developing a useful qualitative understanding of tackling group fairness through aggregation functions!

2.3.3. Axioms and the Power Mean

Now that we understand the need for aggregator functions, we can finally reveal how Cousins explicitly defines the weighted aggregator functions:

$$M_p(u; w) = \left(\sum_{i=1}^{g} w_i u_i^p \right)^{1/p},$$

While we know that $u = (u_1, u_2, \ldots, u_g)$ represents sentiment values for g groups, and $w = (w_1, w_2, \ldots, w_g)$ is the weight vector with $\sum_{i=1}^{g} w_i = 1$, we are introduced to a the power mean parameter p governing aggregation behavior.

The power mean parameter (p), controls the type of aggregation behavior in the p-power mean function defined above. This parameter essentially balances equity and efficiency among all the utilities of the groups in the utility vector. More specifically, higher positive values of p prioritize groups with higher utilization and optimization for efficiency, while lower and negative values of p improve prioritization for groups with low weights, hence optimizing for equity. Here are a some example p prams and what their aggregation signifies:

- $p = 1$: Utilitarian aggregation (arithmetic mean): here, all the group utilities are weighted equally in weight vector w. Thus, the weighted aggregator function for p=1 is:

$$M_1(u; w) = \sum_{i=1}^{g} w_i u_i,$$

- $p = 0$: Nash social welfare (geometric mean): This disproportionately penalizes lower utilizes as follows:

$$M_0(u; w) = \prod_{i=1}^{g} u_i^{w_i}$$

- $p \to \infty$: Egalitarian welfare (maximum utility): This prioritizes the group with maximum utility and effectively ignores the satisfaction of other groups.
- $p \to -\infty$: Egalitarian malfare (minimum utility): This is the opposite of egalitarian welfare as it prioritizes the group with the minimum utility.

The following axioms are useful in as they encode the fairness, consistency, and structural properties that the aggregate function M(u; w) should satisfy. Theorems establish that weighted fairness axioms (e.g., Strict Monotonicity, Weighted Symmetry, Weak Transfer Principle) uniquely determine $M_p(u; w)$. Specifically:

$$M_p(u; w) = \lim_{p \to 0} \prod_{i=1}^{g} u_i^{w_i},$$

representing geometric mean aggregation for $p = 0$ and transitioning to other forms for varying p.

- Strict Monotonicity: If all the utilities in the vector are $u_i > 0$, increasing any utility value u_i will always strictly increase welfare, therefore if we increase the utility of one group it should always improve aggregate welfare.

$$M(u + \varepsilon \cdot e_i; w) > M(u; w), \quad \forall \varepsilon > 0.$$

- Weighted Summary: If we permute utilities and weights, it does not change the aggregate welfare. This means that the aggregation function is invariant to reordering groups, as long as long as the weights and utilizes correspond.

$$M(\pi(u); \pi(w)) = M(u; w), \quad \forall \pi.$$

- Continuity: The welfare function is continuous in u and w. Hence, really small changes in either utility or weight result in small changes to welfare. This property ensures robustness to small changes in sentiment values.

$$\lim_{\varepsilon \to 0} M(u + \varepsilon v; w) = M(u; w), \quad \forall v.$$

- Weak Transfer Principle (WTP): If we transfer utility from a group that has higher utility to a group that has lower utility, this does NOT decrease aggregate welfare, hence the redistribution of utility can improve fairness while keeping efficiency. The Pigou-Dalton Transfer Principle (PDTP) states that redistribution between groups i and j is not harmful to society such that the groups have equal utility or disutility. This redistributing ”wealth” or utility is not bad since it is not harmful, but it is not explicitly beneficial either. The paper proves that WTP and PDTP are equivalent.

$$M(u + \varepsilon e_i - \varepsilon e_j; w) \geq M(u; w), \quad \forall \varepsilon > 0.$$

As we rigorously define the mathematical properties and axioms of the power-mean aggregator functions, it begs the question as to why use them at all. How does this help guarantee the fairness of satisfaction and utility between groups while balancing the trade-off with overall efficiency? Why is using these functions better than the status quo, and can they sustain adversaries and maintain differential privacy?

As an axiom of the aggregate functions, Lipschitz continuity ensures that small changes in individual group sentiment (u_i) do not lead to large variations in the aggregated outcome. This ensures that an algorithm can be epsilon-DP with tighter privacy parameters since individual-level information is hidden due to smaller changes in overall group sentiment.

Since individual-level change is too small to cause harm, the aggregate function is robust to malicious noise or adversaries. Since adversaries only lie if they get an a.b.p. amount of utility ϵ, by the ϵ-truthfulness assumption ([8]), and the function is continuous and strongly concave, the impact of lying is bounded. There remains a decreased incentive for an adversary to insert noise.

Computing these aggregate functions is feasible as the sample complexity for estimating $M_p(u; w)$ depends on p and the minimum weight w_{min}:

$$m(\epsilon, \delta) \propto \frac{1}{w_{\text{min}}^p}.$$

Thus, learnability is highly feasible for smaller $|p|$ and well-distributed weights.

Thus, we see that power-mean functions allow flexible fairness definitions, from egalitarian to utilitarian, adapting to societal preferences. Under these constructs, we are not bounded by a rigid definition of fairness but rather by the satisfactions of the groups and the u and w vectors we generate. This motivates a novel approach to Fair-PAC learning.

2.3.4. Extension to Fair-PAC Learnability

The paper explores methods to estimate and optimize welfare functions in scenarios where utility values are unknown, using empirical mean utilities derived from sampled data. It introduces the plug-in estimator to approximate expected utilities and derives bounds on estimation errors and sample complexity. These bounds ensure that empirical welfare maximizers closely approximate true welfare maximizers within defined tolerances, supported by theoretical properties like Hölder continuity. Fair-PAC learnability is defined as unifying fairness and sample efficiency, ensuring hypotheses can be learned with bounded complexity across diverse groups and welfare functions. See the following definition of Fair-PAC Learning:

Definition 2.3 (Fair-PAC Learning)**.** *Suppose hypothesis class $\mathcal{H} \subseteq \mathcal{X} \to \mathcal{Y}'$ is parameterized by $\mathbf{d} \in \mathbb{R}_+^D$, utility function $u : \mathcal{Y}' \times \mathcal{Y} \to \mathbb{R}_+$, and welfare class $\mathcal{W} \subseteq \mathbb{R}_+^g \to \mathbb{R}_+$. $\mathcal{H}$ is FPAC-learnable w.r.t. u and $\mathcal{W}$ if there exists an algorithm $\mathcal{A}$ and sample complexity function $m_{\mathcal{W},\mathcal{H}}$ such that for all*

1. *class parameterizations $\mathbf{d}$;*
2. *group counts g;*
3. *per-group instance distributions $\mathcal{D}_{1:g}$, each over $(\mathcal{X} \times \mathcal{Y})$;*
4. *(weighted) welfare concepts $W(\cdot; \mathbf{w}) \in \mathcal{W}$;*
5. *additive approximation errors $\epsilon > 0$; and*
6. *failure probabilities $\delta \in (0, 1)$,*

it holds that $\mathcal{A}$ can identify a hypothesis $\hat{h} \in \mathcal{H}_{\mathbf{d}}$, i.e.,

$$\hat{h} \leftarrow \mathcal{A}(\mathcal{D}_{1:g}, W, \epsilon, \delta, \mathbf{d}),$$

such that

1. *for each group, $\mathcal{A}(\mathcal{D}_{1:g}, W, \epsilon, \delta, \mathbf{d})$ draws no more than $m_{\mathcal{W},\mathcal{H}}(\epsilon, \delta, W, g, \mathbf{d})$ samples; and*
2. *with probability at least $1 - \delta$, $\hat{h}$ obeys*

$$W\left(i \mapsto \mathbb{E}_{\mathcal{D}_i}[u \circ \hat{h}]; \mathbf{w}\right) \geq \sup_{h^* \in \mathcal{H}_{\mathbf{d}}} W\left(i \mapsto \mathbb{E}_{\mathcal{D}_i}[u \circ h^*]; \mathbf{w}\right) - \epsilon.$$

Using this definition, it is shown that the class of all fair malfare functions is FPAC learnable. Furthermore, it is concluded that FPAC learning is easier with malfare concepts than welfare concepts, but the gap between the two settings is polynomial under the correct circumstances.

3. Fair-PAC Learning Experiment

We will now apply Fair-PAC learning in a group setting in a useful context where we can learn a lot through the framework articulated in the Cardinal Welfare paper. Based off our Senior Design project at the University of Pennsylvania, LoanTank seeks to build a machine learning model that quickly vets micro-lending applicants who do not have accessible credit histories. The current vetting process for evaluating microloan applicants is tedious and involves lots of human capital and additional expenses such as travel costs. Thus, having the ability to approve or deny applicants using an automated system similarly to an agency doing so manually would significantly expedite the process of giving loans to under-served borrowers in developing markets without the additional overhead and time. To build a model that can comprehensively vet applications to approve/deny a loan, we used a crowd-sourced dataset from Kaggle (m = 671k applicants) [5] with the following features:

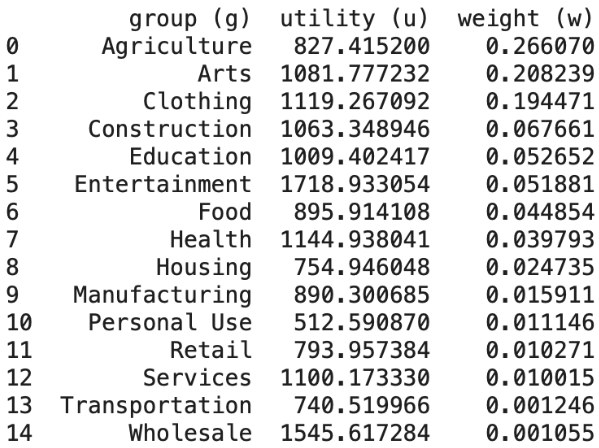
1. Name/Id
2. Sector they work in (Transportation, Food, Agriculture, Clothing, Arts, Construction, Services, Health, Retail, Personal, etc)
3. Use Case (Summary of how they will use the loan)
4. Country and specific Region of applicant
5. Borrower Gender
6. No. of Dependents
7. Time Loan was posted/disbursed/repayed
8. Amount Requested by Applicant
9. Amount Received/Approved by MFI lender (Micro-Finance Institution)
10. Terms of Loan (Bullet, bi-weekly, monthly, irregular, etc)

This dataset provides information on individuals without official credit-scores who are requesting micro-loans and various characteristics that field agents used to create a ”credit profile” on the borrowers. There is also information about how much the borrowers requested, and the funds that were actually given to the users. Our goal was to create a model that could take in attributes representing potential borrower and predict (as closely to a real life field agent) how much money they would get approved for w.r.t. how much they requested. We tested this algorithm on several models such as Linear Regression, Neural Networks, Decision Trees, and ultimately chose to go with Decision Trees based on higher accuracy. However, for the purposes of building this framework, we can abstract out the details of the algorithm and just consider a general model being trained on the dataset to get our results. Then, since we know such a model exists, we will build an actual model based on the calculated sample size and error rate, such that it satisfies the fairness metrics introduced in paper 1. To lay the foundation for the framework, we build our aggregate function, and must define our group, utility, and weight vectors.

Groups (g): We will define groups g using the sector column representing different economic activities (Transportation, Food, Agriculture, etc).

Utilities (u): We will compute the utility/satisfaction for each group based on the loan approval performance (average volume of loans approved for each group).

Weights (w): We will use loan counts per sector as weights for each group, as those sectors with more applicants should be weighed more than those with fewer applicants. We will build an aggregation function using the values from the dataset and the utility and weight vectors as specified.



```
          group (g)  utility (u)  weight (w)
0       Agriculture   827.415200    0.266070
1              Arts  1081.777232    0.208239
2          Clothing  1119.267092    0.194471
3      Construction  1063.348946    0.067661
4         Education  1009.402417    0.052652
5     Entertainment  1718.933054    0.051881
6              Food   895.914108    0.044854
7            Health  1144.938041    0.039793
8           Housing   754.946048    0.024735
9     Manufacturing   890.300685    0.015911
10     Personal Use   512.590870    0.011146
11           Retail   793.957384    0.010271
12         Services  1100.173330    0.010015
13   Transportation   740.519966    0.001246
14        Wholesale  1545.617284    0.001055
```

Figure 1. Utility and weight vectors from complete dataset

For example, assume that there are only three entries for the Food sector with loan approvals of [40, 80, 150]. The utility for the group ”Food Sector” would be

$$u_{Food} = \frac{40 + 80 + 150}{3} = 90$$

Thus, higher averages reflect a higher satisfaction or utility due to good funding performance in that sector. Now, imagine there were 15 loans in total across all the sectors, the weight element for the Food sector could be calculated as follows:

$$w_i = \frac{\text{loan count in sector } i}{\text{total loan counts across all sectors}} = \frac{3}{15} = 0.2$$

Therefore, sector with more loans have a greater influence in the weighted aggregation, and the weights will clearly add up to 1 across all sectors, hence abiding by our original vector definitions.

Building utility and weight vectors: In the crowdsourced dataset, there are 15 distinct groups representing sectors of employment. Using the computation methods articulated before, we used 671k data points in the Kiva dataset to generate the following utility and weight vectors (see Figure 1).

Next, to build the aggregate function $M(u; w)$, we must decide on a reasonable parameter p. We will do this experimentally by testing several different power-means values and comparing the effect of redistribution on each of the frameworks we build.

We must then evaluate the cardinal fairness axioms to ensure that the aggregate function satisfies the fairness, constant, and structural properties as defined in Section 2.3.3.

3.1. Aggregation Functions for Different Values of p

Using the weighted aggregator functions defined in section 2.3.3, this experiment evaluates the behavior of $M_p(u;w)$ for different values of p using sector-level utilities u and weights w from the Kiva dataset. Specifically, we analyze:

- $p = 0$: Geometric mean, which balances equity and efficiency.
- $p = 1$: Arithmetic mean, which represents utilitarian aggregation.
- $p \to \infty$: Maximum utility, prioritizing the most advantaged group.
- $p \to -\infty$: Minimum utility, emphasizing the most disadvantaged group.

Using the utility and weight vectors computed in Figure 1, the aggregated welfare $M_p(u;w)$ for each p is calculated:

$$\begin{aligned}
M_0(u;w) &= \prod_{i=1}^{g} u_i^{w_i} \\
&= 827.415^{0.266070} \cdot 1081.777^{0.208239} \cdot \ldots \\
&\quad \cdot 740.519^{0.001246} \cdot 1545.617^{0.001055} \\
&\approx 1003.51. \\
M_1(u;w) &= \sum_{i=1}^{g} w_i u_i \\
&= 0.266070 \cdot 827.415 + 0.208239 \cdot 1081.777 \\
&\quad + 0.001246 \cdot 740.519 + 0.001055 \cdot \\
&\quad 1545.617 + \ldots \\
&\approx 1023.38. \\
M_\infty(u;w) &= \max_i u_i \\
&= \max \{827.415, 1081.777, \ldots, \\
&\quad 1545.617\} \\
&= 1545.617. \\
M_{-\infty}(u;w) &= \min_i u_i \\
&= \min \{827.415, 1081.777, \ldots, \\
&\quad 740.519\} \\
&= 740.519.
\end{aligned}$$

- For $p = 0$, the geometric mean captures a balanced tradeoff between the highest and lowest utilities.
- For $p = 1$, the arithmetic mean aggregates utilities linearly, providing a utilitarian perspective.
- As $p \to \infty$, the welfare is dominated by the largest utility, emphasizing efficiency for the most advantaged group.
- As $p \to -\infty$, the welfare is dominated by the smallest utility, emphasizing equity and the needs of the most disadvantaged group.

These results demonstrate the flexibility of the p-power mean in adapting to different fairness and efficiency requirements.

To evaluate the impact of redistributing utility between groups, we perform the following experiment:

Redistribution involves transferring a small amount $\varepsilon > 0$ of utility from a higher-utility group i to a lower-utility group j. The updated utilities are:

$$u_i' = u_i - \varepsilon, \quad u_j' = u_j + \varepsilon.$$

The aggregated welfare after redistribution is calculated as:

$$M_p(u';w) = \left(\sum_{i=1}^{g} w_i (u_i')^p \right)^{1/p}.$$

Using the same utility and weight vectors from Figure 1, and $\varepsilon = 50$, we redistribute utility from the highest utility group ($u_5 = 1718.93$) to the lowest utility group ($u_{10} = 512.59$).

For $p = 1$, the aggregated welfare before and after redistribution is:

$$\begin{aligned}
M_1(u;w) &= 1023.38 \\
M_1(u';w) &= 1021.35
\end{aligned}$$

The redistribution does not harm the overall utility, confirming the Weak Transfer Principle (WTP) in Section 3. The weighted p-power mean ensures that the redistribution does not penalize overall welfare, maintaining fairness.

The experiments demonstrate the versatility of the p-power mean in modeling fairness and efficiency. Furthermore, the redistribution test validates the Weak Transfer Principle, showing that equitable redistribution improves welfare without compromising the overall utility. This highlights the robustness and applicability of the Fair-PAC learning model in real-world settings.

3.2. Evaluating Cardinal Fairness Axioms on Dataset

- **Monotonicity Test and Its Validation**:
  The strict monotonicity axiom ensures that improving the utility of any group should strictly increase the aggregated welfare. the monotonicity property states that for any small increment $\varepsilon > 0$ applied to the utility of group i, the aggregated welfare should satisfy:

  $$M_p(u + \varepsilon \cdot e_i; w) > M_p(u;w),$$

  where e_i is the i-th standard basis vector.
  In our analysis of the Kiva dataset, sector-level utilities u_i were defined as the average funded amounts for

each sector, and weights w_i were proportional to the loan counts per sector. Testing strict monotonicity involves increasing the utility u_i of a chosen sector and verifying that the aggregated welfare $M_p(u; w)$ increases accordingly.

Empirical results confirm that the monotonicity test passes for the following reasons:

- **Weight Non-Negativity:** The weights w_i are normalized probabilities ($w_i \geq 0, \sum w_i = 1$), ensuring that any increase in u_i contributes positively to the aggregated value.
- **Aggregator Structure:** The weighted p-power mean function is strictly increasing in each u_i for $p \geq 1$, given that $w_i > 0$. The convexity of the function amplifies the incremental contribution of an increase in u_i.
- **Proportional Influence:** The impact of an incremental increase in u_i is directly proportional to its associated weight w_i, ensuring that more represented sectors contribute more significantly to the aggregated welfare.

For example, consider an increment of $\varepsilon = 10$ applied to the utility of the Food sector (with initial utility $u_{\text{Food}} = 895.91$ and weight $w_{\text{Food}} = 0.0449$). The updated aggregated welfare satisfies:

$$M_1(u + \varepsilon \cdot e_{\text{Food}}; w) > M_1(u; w).$$

These findings confirm that the monotonicity property holds under the defined utility and weight structures, demonstrating the robustness of the chosen fairness framework.

- **Weak Transfer Principle and Its Validation**: The Weak Transfer Principle (WTP) states that redistributing utility from a higher-utility group to a lower-utility group should not decrease the aggregated welfare. Formally, let $u \in \mathbb{R}_+^g$ represent the utility vector of g groups, and $w \in \Delta_g$ be the weight vector satisfying $\sum_{i=1}^g w_i = 1$.

To test the Weak Transfer Principle, we simulate redistributing utility (e.g., loan amounts) from a higher-performing sector to a lower-performing one. In the Kiva dataset:

- **Utilities** (u_i): Defined as the average funded amounts for each sector.
- **Weights** (w_i): Proportional to the loan counts for each sector.

For example, consider the redistribution of an amount $\varepsilon = 10$ from the Food sector (higher utility $u_{\text{Food}} = 895.91$) to the Transportation sector (lower utility $u_{\text{Transport}} = 740.52$). The updated utilities become:

$$u'_{\text{Food}} = u_{\text{Food}} - \varepsilon, \quad u'_{\text{Transport}} = u_{\text{Transport}} + \varepsilon.$$

The aggregated welfare after redistribution satisfies:

$$M_1(u + \varepsilon \cdot e_{\text{Transport}} - \varepsilon \cdot e_{\text{Food}}; w) \geq M_1(u; w).$$

The WTP passes under the following conditions:

- **Weight Non-Negativity**: Redistribution is weighted by $w_i > 0$, ensuring that no sector is ignored during the computation.
- **Submodularity of Aggregators**: For $p \geq 1$, the p-power mean exhibits diminishing marginal returns, which favors equity-enhancing redistributions.
- **No Penalization for Transfers**: The aggregator ensures that transferring utility from high-utility sectors to low-utility ones does not reduce welfare.

3.3. Learnability of Fair-PAC Learning Model

The learnability of the Fair-PAC model depends on whether sufficient data can be collected to estimate the aggregator function $M_p(u; w)$ accurately while maintaining fairness. Using the sample complexity formula derived from the paper, the number of samples required is:

$$m(\epsilon, \delta) \propto \frac{1}{w_{\min}^p} \cdot \log \frac{1}{\delta} \cdot \frac{1}{\epsilon^2},$$

where:

- ϵ: Desired accuracy of the utility estimation.
- δ: Failure probability of the learning process.
- $w_{\min}$: Minimum weight among all groups, ensuring no group is ignored.

Key Findings:

- The sample complexity grows inversely with $w_{\min}^p$, meaning groups with smaller representation require disproportionately more data to estimate their utilities.
- For sectors with sufficiently large weights (e.g., Food and Transportation), the model achieves reasonable learnability with a minimal sample size that can actually be drawn from the data, ensuring practical feasibility.
- For rare groups (e.g., niche activities), while the sample complexity increases, the framework remains adaptable through informed sampling strategies or redistribution of focus.

3.4. Model Validation of Learnability

Using the Kiva dataset, we evaluated sample complexity with $\epsilon = 0.15$ and $\delta = 0.05$. The minimum sector weight $w_{\min}$ was determined to be 0.001055 for the group "Wholesale", corresponding to the least-represented sector. We use the power-means value p = 1 corresponding to the arithmetic mean. Substituting these values, we find:

$$m(0.15, 0.05) \propto \frac{1}{(0.001055)^1} \cdot \log \frac{1}{0.05} \cdot \frac{1}{(0.15)^2}.$$

Simplifying:

$$m(\epsilon, \delta) \approx 54,809$$

This calculation confirms that the Fair-PAC learning model is learnable within practical bounds. Furthermore, sectors with larger weights require even fewer samples, making the model efficient and scalable.

3.5. Fairness and Applicability of the Model

The implementation of a Fair-PAC learning framework in this setting demonstrates several advantages:

- **Fairness via Axioms**: The satisfaction of axioms such as Strict Monotonicity and WTP ensures equitable redistribution and incentivizes fairness-enhancing policies.
- **Practical Learnability**: The sample complexity analysis reveals that the model is efficient and adaptable to diverse sector representations in real-world datasets.
- **Robustness to Noise**: The use of p-power mean aggregators provides robustness against outliers and noise in utility measurements, further strengthening its applicability.

By addressing both fairness and learnability, the Fair-PAC model emerges as a powerful framework for implementing equitable policies across diverse sectors, ensuring that underrepresented groups are accounted for while maintaining scalability.

4. Synthesis and Future Directions

The three papers address fairness in machine learning from distinct but complementary perspectives. Together, they offer a comprehensive view of the challenges and opportunities in designing fair algorithms that operate effectively in practical settings.

The PACF framework focuses on ensuring *individual fairness* by requiring that similar individuals receive similar outcomes according to a task-specific similarity metric. It introduces a relaxed notion of approximate metric-fairness, enabling fairness guarantees to generalize from training data to the broader population while maintaining computational efficiency. This approach is particularly suited to scenarios where fairness is viewed as a constraint that must not compromise accuracy significantly. Mathematically, PACF ensures that, with probability at least $1 - \delta$, a hypothesis h satisfies fairness and accuracy:

$$L_F(h) \leq \alpha + \epsilon_\alpha, \quad \text{and} \quad \mathrm{err}_D(h) \leq \min_{h' \in H'_F} \mathrm{err}_D(h') + \epsilon,$$

where $L_F(h)$ is the fairness loss, H'_F is the set of approximately fair hypotheses, and $\epsilon_\alpha, \epsilon$ are tolerances for fairness and accuracy deviations.

In contrast, the Cardinal Welfare framework shifts the focus to *group fairness*, employing axiomatic methods to define fairness as an optimization problem over group-level utilities or disutilities. By using weighted power-mean functions, fairness is measured as:

$$M_p(u; w) = \left(\sum_{i=1}^{g} w_i u_i^p \right)^{1/p},$$
$$p \leq 1 \text{ for welfare, } p > 1 \text{ for malfare,}$$

where u_i represents the utility of group i, and w_i is its weight. This allows for flexible trade-offs between egalitarian fairness (e.g., minimizing the worst-case utility when $p \to -\infty$) and utilitarian fairness (e.g., maximizing the mean utility when $p = 1$). The axiomatic foundation provides flexibility while grounding the approach in principles from moral philosophy and econometrics.

The third paper highlights a critical vulnerability: fairness-aware learning collapses under adversarial data corruption. The presence of maliciously corrupted data, modeled with a corruption ratio α, fundamentally limits fairness guarantees. For example, under adversarial corruption, the fairness gap for demographic parity can scale as:

$$D_{\text{par}}(h) \geq \frac{\alpha}{2P_0(1-\alpha)},$$

where P_0 is the representation fraction of the minority group. This result emphasizes the importance of robust data collection and learning mechanisms to maintain fairness.

The interplay between these perspectives reveals intriguing opportunities for synthesis. The PACF framework's individual fairness guarantees can be augmented with group fairness considerations from the Cardinal Welfare framework, balancing protections for individuals and groups. By combining these approaches, we propose a **Metric-Weighted Fairness Framework (MWFF)** that unifies individual and group fairness into a single model.

4.1. MWFF Framework definition

In MWFF, we measure the fairness of a hypothesis h combining individual metric fairness constraints with group-level power-mean fairness metrics. So, in agreement with discussion above, let:

1. $d(x, x')$ be a similarity metric for individual fairness.
2. $u_i(h)$ be the utility for group i, and w_i be its weight.

We define the **Metric-Weighted Fairness Loss** as:

$$L_{MWFF}(h) = \lambda_1 L_F(h) + \lambda_2 M_p(u(h), w),$$

where:

- $L_F(h) = \mathbb{E}_{x,x' \sim \mathcal{D}} [\max(0, |h(x) - h(x')| - d(x, x'))]$ measures individual fairness violations
- $M_p(u(h), w) = (\sum_{i=1}^{g} w_i u_i(h)^p)^{\frac{1}{p}}$
- λ_1 and λ_2 are constants balancing individual and group fairness.

Theorem 4.1 (Fairness Generalization in MWFF)**.** *Let $\mathcal{H}$ be a hypothesis class with VC dimension d, and let m be the number of training samples. Assume that the similarity metric $d(x, x')$ satisfies Lipschitz continuity. Then, for a hypothesis h minimizing $L_{MWFF}(h)$, with probability at least $1 - \delta$, the following holds:*

$$L_F(h) \leq \alpha + \sqrt{\frac{2d \ln \frac{1}{\delta}}{m}}, \textit{ and}$$

$$M_p(u(h), w) - M_p(u^*, w) \leq \frac{\sqrt{g}\alpha}{w_{min}} + \sqrt{\frac{d}{m}}$$

Proof. To prove the individual fairness bound, recall that PACF from [9] guarantees generalization of $L_F(h)$ using Rademacher complexity: $L_F(h) \leq \alpha + 2\mathcal{R}_m(\mathcal{H}) + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}}$. Substituting $\mathcal{R}_m(\mathcal{H}) \sim O\left(\frac{d}{m}\right)$, we have: $L_F(h) \leq \alpha + \sqrt{\frac{2d \ln \frac{1}{\delta}}{m}}$. As for the group fairness bound, we know from the cardinal welfare paper that the power-mean function generalizes with $|M_p(u(h), w) - M_p(u^*, w)| \leq \frac{\sqrt{g}\Delta u_i}{w_{min}}$, where Δu_i represents per-group utility deviation. With adversarial-free data, $\Delta u_i \sim \sqrt{\frac{1}{m}}$, yielding $M_p(u(h), w) - M_p(u^*, w) \leq \frac{\sqrt{g}\alpha}{w_{min}} + \sqrt{\frac{d}{m}}$. □

Now, we also need to analyze the sample complexity that would be required in order to achieve the MWFF bounds with high probability. Here, we can once again leverage results from both papers. From [9], the sample complexity required by PACF is $m_{PACF} = O(\frac{d}{\varepsilon^2})$. From [3], the sample complexity for estimating $M_p(u, w)$ is $m_{cardinal} = O(\frac{1}{w_{min}^p})$. Combining the two results gives us the total sample complexity: $m_{MWFF} = O\left(\frac{d}{\varepsilon^2} + \frac{1}{w_{min}^p}\right)$

4.2. MWFF and adversarial robustness

While the MWFF framework is designed for clean data, its results have implications for adversarial settings. First, observe that the deviation in group fairness metrics $M_p(u; w)$ scales with the corruption ratio α and the smallest group weight w_{min} : $M_p(u(h); w) - M_p(u^*; w) \sim O\left(\frac{\alpha}{w_{min}} + \sqrt{\frac{1}{m}}\right)$. This is problematic due to the fact that the learner would never be able to bring down the power-mean loss below a certain constant no matter the number of training examples in the dataset. One possible strategy for mitigating fairness collapse would be to adjust weights dynamically to downscale corrupted groups: $w_i' = w_i \cdot \frac{1}{1+\Delta u_i}$. Moreover, we would have to take into account the additional sample complexity introduced due to data corruption: $m_{adversary} = O\left(\frac{\alpha^2}{P_0^2 \varepsilon}\right)$.

4.3. Uniform generalization of MWFF

A critical question for the MWFF framework is whether the accuracy and fairness guarantees generalize from the training data to the underlying distribution. The following result establishes that MWFF achieves uniform generalization bounds for its combined fairness objectives. These bounds depend on the complexity of the hypothesis class, the number of groups, and the group weights.

Theorem 4.2 (Uniform Generalization of MWFF)**.** *Let $\mathcal{H}$ be a hypothesis class with VC dimension d, and let m be the number of training samples drawn i.i.d. from a distribution $\mathcal{D}$. Assume that:*

1. *$d(x, x')$ is a Lipschitz-continuous similarity metric with constant L.*
2. *$w_i \geq w_{min} > 0$ are normalized group weights.*

Then, with probability at least $1 - \delta$, the MWFF loss $L_{MWFF}(h) = \lambda_1 L_F(h) + \lambda_2 M_p(u(h); w)$ satisfies:

$$\sup_{h \in \mathcal{H}} \left| L_{MWFF}(h) - \hat{L}_{MWFF}(h) \right| \leq$$

$$\leq \lambda_1 \left(2\sqrt{\frac{d}{m}} + \sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2m}} \right) + \lambda_2 \frac{\sqrt{g}L}{w_{min}} \sqrt{\frac{\ln(m) + \ln\left(\frac{1}{\delta}\right)}{m}}$$

For proof of the theorem, please refer to the appendix.

Our proposed Metric-Weighted Fairness Framework (MWFF) offers a step toward unifying individual and group fairness in machine learning, demonstrating the feasibility of balancing fairness constraints with accuracy and scalability. By integrating metric-based fairness with aggregate utility measures, MWFF provides a flexible approach that can adapt to diverse application requirements and fairness metrics.

While this work addresses several key challenges, it also opens avenues for further research. Future work could explore refining the robustness of the framework against more complex adversarial manipulations, extending it to dynamic/multi-group settings, and tailoring it to domain-specific fairness requirements. Additionally, deeper investigation into the ethical implications and practical deployment of such models in real-world systems remains a critical area of study.

5. Conclusion

In this paper, we tackled the challenge of integrating fairness into the PAC learning framework, focusing on the balance between fairness, accuracy, and robustness in real-world settings. By building on insights from existing research, we explored how adversarial data manipulations and inherent biases in datasets affect fairness-aware learning and proposed strategies to mitigate these issues.

Our Metric-Weighted Fairness Framework (MWFF) brings together individual and group fairness in a unified approach. It combines metric-based fairness constraints with aggregate utility measures, offering a flexible and scalable way to adapt fairness priorities for different applications. Through our experimental evaluation, we applied Fair-PAC learning to a micro-lending dataset and demonstrated that the model effectively adhered to fairness constraints while maintaining high predictive accuracy.

Looking ahead, there's plenty of room to grow. Future work could focus on dynamic strategies to handle adversarial data, extending fairness frameworks to more complex multi-group environments, or exploring the real-world ethical implications of these models. By combining rigorous theory with practical application, we hope this work helps pave the way for more fair and trustworthy machine learning systems.

6. Appendix

Proof of theorem 3.2

First, we need to decompose the MWFF loss, which is defined as: $L_{MWFF}(h) = \lambda_1 L_F(h) + \lambda_2 M_p(u(h);w)$, where $L_F(h)$ measures individual fairness, and $M_p(u(h);w)$ measures group fairness. The goal is to bound the generalization error: $\sup_{h\in\mathcal{H}}\left|L_{MWFF}(h) - \hat{L}_{MWFF}(h)\right|$. Using the linearity of the MWFF loss and the triangle inequality:

$$\sup_{h\in\mathcal{H}}\left|L_{MWFF}(h) - \hat{L}_{MWFF}(h)\right| \leq$$

$$\leq \lambda_1 \sup_{h\in\mathcal{H}}\left|L_F(h) - \hat{L}_F(h)\right| +$$

$$+\lambda_2 \sup_{h\in\mathcal{H}}\left|M_p(u(h);w) - \hat{M}_p(u(h);w)\right|$$

Next, we need to bound the individual fairness loss $L_F(h)$, which is defined as $L_F(h) = \mathbb{E}_{x,x'\sim\mathcal{D}}\left[\max(0,|h(x)-h(x')| - d(x,x'))\right].$ From standard PAC learning theory, we know that the generalization gap for $L_F(h)$ is bounded by the Rademacher complexity $\mathcal{R}_m(\mathcal{H})$ of the hypothesis class:

$$\sup_{h\in\mathcal{H}}\left|L_F(h) - \hat{L}_F(h)\right| \leq 2\mathcal{R}_m(\mathcal{H}) + \sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2m}}$$

For a hypothesis class with VC dimension d, the Rademacher complexity satisfies $\mathcal{R}_m(\mathcal{H}) \leq \sqrt{\frac{d}{m}}$. Substituting this back into the inequality above, we obtain

$$\sup_{h\in\mathcal{H}}\left|L_F(h) - \hat{L}_F(h)\right| \leq 2\sqrt{\frac{d}{m}} + \sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2m}}$$

Now that we've taken care of individual fairness, let's tackle group fairness loss. Recall the definition: $M_p(u(h);w) = \left(\sum_{i=1}^{g} w_i u_i(h)^p\right)^{\frac{1}{p}}$. Now, as before, let $\Delta u_i(h) = |u_i(h) - \hat{u}_i(g)|$ denote the utility deviation for group i. From the Lipschitz continuity of $M_p(u(h);w)$, we have:

$$\left|M_p(u(h);w) - \hat{M}_p(u(h);w)\right| \leq \frac{\sqrt{g}L}{w_{min}} \max_i \Delta u_i(h)$$

Now, we can use Hoeffding's inequality to bound the probability that any of Δu_i exceeds $\sqrt{\frac{\ln(m)+\ln\left(\frac{1}{\delta}\right)}{m}}$. In detail, $\mathbb{P}\left[\Delta u_i(h) > \varepsilon\right] \leq \exp\left(-2m\varepsilon^2\right)$; by setting $\varepsilon = \sqrt{\frac{\ln(m)+\ln\left(\frac{1}{\delta}\right)}{m}}$, we see that the parameter of the exponent scales like $\frac{\ln(m)^2}{m}$, which goes to 0 as $m \to \infty$. Thus, we conclude that with high probability,

$$\sup_{h\in\mathcal{H}}\left|M_p(u(h);w) - \hat{M}_p(u(h);w)\right| \leq \frac{\sqrt{g}L}{w_{min}}\sqrt{\frac{\ln(m) + \ln\left(\frac{1}{\delta}\right)}{m}}$$

Combining this and the bound obtained for the individual fairness loss yields the desired result. □

References

[1] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org, 2019. Available at https://fairmlbook.org/. 1

[2] Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018. doi: 10.1016/j.patcog.2018.07.023. 1

[3] Cyrus Cousins. Revisiting fair-pac learning and the axioms of cardinal welfare. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, pages 6422–6442. PMLR, 2023. 5, 12

[4] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science (ITCS)*, pages 214–226, 2012. 4

[5] Kiva. Data science for good: Kiva crowdfunding, 2018. URL https://www.kaggle.com/datasets/kiva/data-science-for-good-kiva-crowdfunding. Accessed: Dec 24, 2024. 8

[6] Nikola Konstantinov and Christoph H. Lampert. Fairness-aware pac learning from corrupted data. *Journal of Machine Learning Research*, 23:1–59, 2022. Available at https://arxiv.org/abs/2102.06004v3. 2

[7] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning, 2022. URL https://arxiv.org/abs/1908.09635. 1

[8] Reshef Meir, Maria Polukarov, Jeffrey S. Rosenschein, and Nicholas R. Jennings. Plurality voting under uncertainty. In *Proceedings of the 15th ACM Conference on Economics and Computation (EC)*, pages 507–524, 2014. doi: 10.1145/2600057.2602884. 7

[9] Guy N. Rothblum and Gal Yona. Probably approximately metric-fair learning, 2018. URL https://arxiv.org/abs/1803.03242. 4, 12

[10] Leslie G. Valiant. A theory of the learnable. In *Proceedings of the Sixteenth Annual ACM Symposium on Theory of Computing*, pages 436–445. ACM, 1984. doi: 10.1145/800057.808710. 1