An Axiomatic Theory of Provably-Fair Welfare-Centric Machine Learning

Cyrus Cousins
Department of Computer Science
Brown University
cyrus_cousins@brown.edu

Abstract
We address an inherent difficulty in welfare-theoretic fair ML, by proposing an equivalently-axiometrically justified alternative setting, and studying the resulting computational and statistical learning questions. Welfare metrics quantify overall wellbeing across a population of groups, and welfare-based objectives and constraints have recently been proposed to incentivize fair ML methods to satisfy their diverse needs. However, many ML problems are cast as loss minimization tasks, rather than utility maximization, and thus require nontrivial modeling to construct utility functions. We define a complementary metric, termed malfare, measuring overall societal harm, with axiomatic justification via the standard axioms of cardinal welfare, and cast fair ML as malfare minimization over the risk values (expected losses) of each group. Surprisingly, the axioms of cardinal welfare (malfare) dictate that this is not equivalent to simply defining utility as negative loss and maximizing welfare. Building upon these concepts, we define fair-PAC learning, where a fair-PAC learner is an algorithm that learns an \( \varepsilon-\delta \) malfare-optimal model, with bounded sample complexity for any data distribution and (axiomatically justified) malfare concept. We show conditions under which many standard PAC-learners may be converted to fair-PAC learners. This places fair-PAC learning on firm theoretical ground, as it yields statistical, and in some cases computational, efficiency guarantees for many well-studied machine-learning models. Fair-PAC learning is also practically relevant, as it democratizes fair ML by providing concrete training algorithms with rigorous generalization guarantees.

1 Introduction
It is now well-understood that contemporary ML systems exhibit differential accuracy across gender, race, and other protected groups, for tasks like facial recognition [4,6,5], in medical settings [15,2], and many others. This exacerbates existing inequality, as facial recognition in policing yields disproportionate false-arrest rates, and medical ML yields disproportionate health outcomes. Welfare-centric ML methods encode both accuracy and fairness in a welfare function defined on a set of groups, and optimize or constrain welfare to learn fairly. This addresses differential accuracy and bias issues across groups by ensuring that (1) each group is seen and considered during training, and (2) an outcome is incentivized that is desirable overall, ideally according to some mutually-agreed-upon welfare function. Unfortunately, welfare metrics require a notion of (positive) utility, and we argue that this is not natural to many ML tasks, wherein we instead minimize some negatively connoted loss value. We thus define a complementary measure to welfare, which we term malfare, measuring societal harm (rather than wellbeing). In particular, malfare arises naturally when one applies the standard axioms of cardinal welfare to loss values rather than utility values. We then cast fair ML as a direct malfare minimization problem, and study its computational and statistical aspects.
Perhaps surprisingly, minimizing a malfare function is not equivalent to maximizing some welfare function while taking utility to be negative loss, except in the egalitarian and utilitarian edge cases. This is because nearly every function satisfying the standard axioms of cardinal welfare requires nonnegative inputs, and it is in general impossible to contort a loss function into a utility function while satisfying this requirement. For example, while minimizing the 0-1 loss, which simply counts the number of mistakes a classifier makes, is isomorphic to maximizing the 1-0 gain, which counts number of correct classifications, minimizing some malfare function defined on 0-1 loss over groups is not in general equivalent to maximizing any welfare function defined on 1-0 gain.

Building upon these concepts, we develop a generic notion of fair machine learning, termed fair-PAC (FPAC) learning, where the goal is to learn models for which finite training samples may guarantee near-optimal malfare (with high probability), for any (axiomaticaly justified) malfare concept. This definition extends Valiant’s classic PAC-learning formalization, and we show that, with appropriate modifications, many (standard) PAC-learners may be converted to FPAC learners. In particular, we show via a constructive polynomial reduction that realizables FPAC-learning reduces to realizable PAC-learning. Furthermore, we show, non-constructively, that for learning problems where PAC-learnability implies uniform convergence, it is equivalent to FPAC-learnability. We also show that when training is possible via convex optimization under standard assumptions, then training ε-δ malfare-optimal models, like risk-optimal models, requires polynomial time. We briefly summarize our contributions below.

1. We derive in section 2 the malfare concept, extending welfare to measure negatively-connoted attributes, and show that malfare-minimization naturally generalizes risk-minimization to produce fairness-sensitive machine-learning objectives that consider multiple protected groups.
2. Section 3 extends PAC-learning to FPAC-learning, where we consider minimization of malfare rather than risk (expected loss) objectives. Both PAC and FPAC learning are parameterized by a learning task (model space and loss function), and we explore the resulting learnability hierarchy.
3. We show that for many loss functions, PAC and FPAC learning are statistically equivalent, and convexity conditions for computationally efficient PAC-learning are also sufficient for FPAC-learning.

1.1 Related Work

Constraint-based notions of algorithmic fairness have risen to prominence in fair ML, with the potential to ensure fairness (i.e., via parity constraints between the various groups, such as equalized odds, demographic-parity, equality of opportunity, equality of outcome, etc.), thus correcting for some forms of data or algorithmic bias. While noble in intent and intuitive by design, fairness via such statistical constraints has several prominent flaws: most notably, several popular parity constraints are mutually unsatisfiable [4], and their constraint-based formulation inherently puts accuracy and fairness at odds, where additional tolerance parameters are required to strike a balance between the two. Furthermore, recent works [12] have shown that welfare and even disadvantaged group utility can be harmed by such fairness constraints, calling into question whether these constraints actually benefit those that they purport to aid.

Perhaps in response to these issues, some recent work has trended toward welfare-based fairness-concepts, wherein both accuracy and fairness are encoded in a welfare function defined on a group of subpopulations. Welfare is then directly optimized [12][21][25] or constrained [26][31] to promote fair learning across all groups. Perhaps the most similar to our work is a method of [12], wherein they directly maximize empirical welfare over linear classifiers; however as with other previous works, an appropriate utility function must be selected, which we avoid by instead using malfare. We argue that empirical welfare maximization is an effective strategy when an appropriate and natural measure of utility is available, but in machine learning contexts, there is no “correct” or clearly neutral way to convert loss to utility. We avoid this issue by working directly with malfare and loss.

Our most poignant contrast to existing theoretical work is the Seldonian learner [27] framework, which essentially extends PAC-learning to learning problems with both constraints and arbitrary nonlinear objectives. We argue that this generality is harmful to the utility of the concept as a mathematical or practical object, as nearly any ML problem can be posed as a constrained nonlinear optimization task. The utility in FPAC learning is that it is sophisticated enough to handle fairness issues with an axiomaticaly-justified objective, but remains simple enough to study as a mathematical object, leading to informative reductions between various PAC and FPAC learnable classes.
2 Aggregating Sentiment within Populations

A generic aggregator function \( M(S; w) \) summarizes some sentiment vector \( S \in \mathbb{R}^d_+ \), across a population of \( g \) groups, weighted by a probability vector \( w \in \mathbb{R}^d_+ \) s.t. \( \|w\|_1 = 1 \). When \( S \) measures a desirable quantity, generally termed utility, the aggregator function is a measure of \textit{cardinal welfare} \cite{17}, and thus quantifies overall \textit{wellbeing}. We also consider the inverse-notion, that of overall \textit{illbeing}, termed \textit{malfare}, in terms of an undesirable \( S \), generally \textit{loss} or \textit{risk}, which naturally extends the welfare concept. We show an equivalent \textit{axiomatic justification} for malfare, and argue that its use is more natural in many situations, particularly when considering or optimizing \textit{loss} functions.

\textbf{Definition 2.1 (Aggregator Functions: Welfare and Malfare).} An aggregator function \( M(S; w) \) measures the overall sentiment of a population, given sentiment vector \( S \) and probability vector \( w \). If \( S \) denotes a desirable quantity (e.g., utility), we call \( M(S; w) \) a \textit{welfare function}, written \( W(S; w) \), and inversely, if it is undesirable (e.g., disutility, loss, or risk), we call \( M(S; w) \) a \textit{malfare function}, written \( \Lambda(S; w) \).

For now, think of the term \textit{aggregator function} as signifying that an entire population, with diverse and subjective desiderata, is considered and summarized into an \textit{aggregate value} \( M(S; w) \), rather than a single group’s perspective (i.e., some \( S_g \)). As we introduce axioms and show consequent properties, the appropriateness of the term shall become more apparent. We use the term \textit{sentiment} to refer to \( S \) with neutral connotation, but when discussing welfare or malfare, we often refer to \( S \) as \textit{utility} or \textit{risk}, respectively, as in these cases, \( S \) describes a well-understood \textit{pre}existing concept. We shall see that often aggregator functions on utility values and risk values are mathematically interchangeable, however, in order to promote fairness, the desirable axioms of malfare and welfare functions differ slightly. The notation reflects the distinction; \( M(S; w) \) is \( M \) for \textit{mean}, whereas \( W(S; w) \) is \( W \) for \textit{welfare}, and \( \Lambda(S; w) \) is \( \Lambda \) (inverted \( W \)), to emphasize its inverted nature.

2.1 Axioms of Cardinal Welfare and Malfare

\textbf{Definition 2.2 (Axioms of Cardinal Welfare and Malfare).} We define the \textit{aggregator-function axioms} for aggregator function \( M(S; w) \) below. For each item, assume (if necessary) that the axiom applies \( \forall S, S' \in \mathbb{R}^d_+ \), scalars \( \alpha, \beta \in \mathbb{R}_{0+} \), and probability vector \( w \in \mathbb{R}^d_+ \).

1. \textit{Strict Monotonicity}: \( S' \neq 0 \implies M(S; w) < M(S + S'; w) \).
2. \textit{Weighted Symmetry}: Suppose \( S' \in \mathbb{R}^d_+ \) and probability vector \( w' \in \mathbb{R}^d_+ \), such that for all \( u \in \mathbb{R}_{0+} \) it holds \( \sum_{i:S_i = u} w_i = \sum_{i:S_i = w'} w'_{i} \). Then \( M(S; w) = M(S'; w') \).
3. \textit{Continuity}: \( M(S; w) \) is a continuous function in both \( S \) and \( w \).
4. \textit{Independence of unconcerned agents}:
   \[ M(S_{1:g-1}, \alpha; w) \leq M(S_{1:g-1}, \alpha; w) \implies M((S_{1:g-1}), \beta; w) \leq M((S_{1:g-1}), \beta; w) \cdot \]
5. \textit{Independence of common scale}: \( M(S; w) \leq M(S'; w) \implies M(\alpha S; w) \leq M(\alpha S'; w) \).
6. \textit{Multiplicative linearity}: \( M(\alpha S; w) = \alpha M(S; w) \).
7. \textit{Unit scale}: \( M(1; w) = M((1, \ldots, 1); w) = 1 \).
8. \textit{Pigou-Dalton transfer principle}: Suppose \( \mu = w \cdot S = w \cdot S' \), and for all \( i \in \{1, \ldots, g\} \):
   \[ |\mu - S'_i| \leq |\mu - S_i| \]. Then \( W(S'; w) = W(S; w) \).
9. \textit{Anti-Pigou-Dalton transfer principle}: Suppose as in \( 8 \) but conclude \( M(S'; w) \leq M(S; w) \).

We take a moment to comment on each axiom, to preview their purpose and assure the reader of their necessity. Axioms\cite{15} are essentially the standard \textit{axioms of cardinal welfarism} \cite{22, 20}. Together, they imply (via the Debreu-Gorman theorem \cite{8, 10}) that any aggregator function can be decomposed as a \textit{monotonic function of a sum} (over groups) of \textit{logarithm} or \textit{power} functions.

Axiom \( 6 \) is a natural and useful property, and one which enables \textit{dimensional analysis} on mean functions; in particular, the \textit{units} of aggregator function match those of sentiment values. Note that axiom\( 6 \) implies axiom\( 5 \) and is thus a simple strengthening of a traditional cardinal welfare axiom.

\footnote{In the unweighted case, it is standard to define symmetry as simply \( M(S) = M(\pi(S)) \) for all permutations \( \pi \) over \( \{1, \ldots, g\} \), but with weightings, the simple extension \( M(S; w) = M(\pi(S); \pi(w)) \) is not quite sufficient. In particular, in this weaker form, nowhere is \textit{additive decomposability}, wherein a group may be decomposed into multiple groups of the same sentiment value and total weight without impacting the aggregate, codified.}
We will also see that it is essential to show convenient statistical and learnability properties. Axiom\(^7\) furthers this theme, as it ensures that not only do units of means match those of \(S\), but scale does as well (making comparisons like “\(S\) is above the population’s welfare” meaningful), and also enabling comparison across populations (i.e., comparing averages rather than sums).

Finally, axiom\(^8\) the Pigou-Dalton transfer principle \(^[19, 7]\), is also standard in cardinal welfare theory, as it ensures fairness in the sense that welfare is higher when utility values are more uniform (i.e., incentivizing equitable redistribution of “wealth” in welfare). Its antithesis, axiom\(^9\) encourages the opposite; in the context of welfare, this perversely incentivizes inequality, but for malfare, which we generally wish to minimize, the opposite occurs, thus this axiom characterizes fairness in the context of malfare. We thus state axioms\(^8, 9\) specifically for welfare or malfare, respectively, whereas axioms\(^1, 7\) apply equally well to both welfare and malfare.

Axioms\(^1, 5\) are modified from their standard presentation to include the weighting \(w\), wherein continuity \(^3\) must also hold over \(w\), and weighted symmetry \(^2\) specifies additionally that weight may be transferred between groups with equal sentiment value. Axioms\(^6, 7\) are novel to this work, and are key in strengthening the Debreu-Gorman theorem to ensure that all welfare and malfare functions are power means in the sequel. Axiom\(^9\) is also novel, as it is necessary to flip the inequality of axiom\(^8\) when the sense of the aggregator function is inverted from welfare to malfare; in particular, the semantic meaning shifts from requiring that “redistribution of utility is desirable” to “redistribution of disutility is not undesirable.”

**The Power-Mean** We now define the \(p\)-power-mean \(M_p(\cdot; \cdot)\), for any \(p \in \mathbb{R} \cup \{\pm \infty\}\), which we shall use to quantify both welfare and malfare. Power means arise when analyzing aggregator functions obeying the various axioms of definition \(^2, 2\) and are thus particularly important to this work.

**Definition 2.3** (Power-Mean Welfare and Malfare). Suppose \(p \in \mathbb{R} \cup \{\pm \infty\}\), sentiment vector \(S \in \mathbb{R}_{0+}^g\), and probability vector \(w \in \mathbb{R}_0^g\). We then define the \(p\)-weighted-power-mean as

\[
M_p(S; w) = \sum_{i=1}^{g} w_i S_i^p , \quad M_0(S; w) = \exp \left( \sum_{i=1}^{g} w_i \ln S_i \right) , \quad M_{\pm \infty}(S; w) = \pm \max_{i=1,\ldots,g} \pm S_i ,
\]

where \(p \in \{-\infty, 0, \infty\}\) attain their limits, the minimum, geometric mean, and maximum, respectively.

**Theorem 2.4** (Properties of the Power-Mean). Suppose sentiment vectors \(S, S' \in \mathbb{R}_{0+}^g\), and probability vector \(w \in \mathbb{R}_0^g\). The following then hold.

1. **Monotonicity:** \(M_p(S; w)\) is weakly-monotonically-increasing in \(p\).
2. **Subadditivity:** \(\forall p \geq 1 : M_p(S + S'; w) \leq M_p(S; w) + M_p(S'; w)\).
3. **Contraction:** \(\forall p \geq 1 : |M_p(S; w) - M_p(S'; w)| \leq M_p(|S - S'| ; w) \leq \|S - S'\|_\infty\).
4. **Curvature:** \(M_p(S; w)\) is concave in \(S\) for \(p \leq 1\) and convex in \(S\) for \(p \geq 1\).

### 2.2 Properties of Welfare and Malfare Functions

We now derive properties of welfare and malfare from the axioms of definition \(^2, 2\)

**Theorem 2.5** (Aggregator Function Properties). Suppose aggregator function \(M(S; w)\), and assume arbitrary sentiment vector \(S \in \mathbb{R}_{0+}^g\) and probability vector \(w \in \mathbb{R}_0^g\). If \(M(\cdot; \cdot)\) satisfies (subsets of) the aggregator-function axioms (see definition \(^2, 2\)), then \(M(\cdot; \cdot)\) exhibits the following properties.

1. **Identity:** Axioms\(^6, 7\) imply that \(\forall \alpha \in \mathbb{R}_{0+} : M(\alpha; w) = \alpha\).
2. **Debreu-Gorman Factorization:** Axioms\(^1, 5\) imply \(\exists \alpha \in \mathbb{R},\) strictly-monotonically-increasing continuous \(F: \mathbb{R} \to \mathbb{R}_{0+}\) s.t.

\[
M(S; w) = F \left( \sum_{i=1}^{g} w_i f_p(S_i) \right) , \quad \text{with} \quad \begin{cases} p = 0 \quad f_0(x) \doteq \ln(x) \\ p \neq 0 \quad f_p(x) \doteq \text{sgn}(p)x^p \end{cases}.
\]

3. **Power-Mean Factorization:** Axioms\(^1, 7\) imply \(F(x) = f_p^{-1}(x)\), thus \(M(S; w) = M_p(S; w)\).
4. **Fair Welfare:** Axioms\(^1, 5\) and \(^8\) imply \(p \in (-\infty, 1]\).
5. **Fair Malfare:** Axioms\(^1, 5\) and \(^9\) imply \(p \in [1, \infty)\).
We define the unbiased estimate of risk, with 6 (multiplicative linearity), that welfare and utility, or malfare and loss, are measured in the same units (e.g., nats of cross-entropy loss, or dollars of income utility), and the power-mean is effectively the only reasonable family of welfare or malfare function; even without axiom 6, axioms 1-5 imply aggregator function functions are still monotonic transformations of power-means. Furthermore, the entirely milquetoast unit scale axiom (7) implies that sentiment values and aggregator functions have the same scale, making comparisons like “the risk of group i is above (or below) the population’s malfare” meaningful. Finally, we have that $W_p(\cdot; w)$ promotes equitable redistribution of utility for $p \leq 1$, and $\Lambda_p(\cdot; w)$ promotes equitable redistribution of loss for $p \geq 1$.

3 Statistical and Computational Learning-Efficiency Guarantees

In this section, we define a formal notion of fair-learnability, termed $FPAC$-learning, where a loss function and hypothesis class are $FPAC$-learnable essentially if any distribution can be learned to approximate malfare-optimality from a finite sample (w.h.p.). We then construct various $FPAC$ learners, and relate the concept to standard PAC learning [23], with the understanding that this allows the vast breadth of research on PAC-learning algorithms to be applied to $FPAC$ learning. In particular, we show a hierarchy of fair-learnability via generic statistical and computational reductions.

Definition 3.1 (Hypothesis Classes and Class Sequences). A hypothesis class is a family of functions mapping domain $\mathcal{X}$ to codomain $\mathcal{Y}$, and a hypothesis class sequence $\mathcal{H} = \mathcal{H}_1, \mathcal{H}_2, \ldots$ is a nested sequence of hypothesis classes, each mapping $\mathcal{X} \rightarrow \mathcal{Y}$. In other words, $\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \cdots$.

For example, linear classifiers naturally form a sequence of families using their dimension as $\mathcal{H}_d = \{ \tilde{x} \mapsto \text{sgn}(\tilde{x} \cdot \langle w_1, d \rangle) \mid \tilde{w} \in \mathbb{R}^d \}$. The hypothesis class sequence concept allows us to distinguish statistically-easy problems, like learning hyperplanes in finite-dimensional $\mathbb{R}^d$, from statistically-challenging problems, like learning hyperplanes in $\mathbb{R}^\infty$. It is also used to analyze the computational complexity of learners as $d$ increases.

3.1 Empirical Malfare Minimization and the Rademacher Average

We define the risk of hypothesis $h : \mathcal{X} \rightarrow \mathcal{Y}$ w.r.t. loss $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_+$ on distribution $\mathcal{D}$ over $(\mathcal{X} \times \mathcal{Y})$, and the empirical risk on sample $z \in (\mathcal{X} \times \mathcal{Y})^n$, as

$$R(h; \ell, D) = \frac{1}{n} \sum_{(x, y) \in D} \ell(y, h(x))$$

and

$$\hat{R}(h; \ell, z) = \frac{1}{n} \sum_{(x, y) \in z} \ell(y, h(x))$$

respectively. The goal in ML is generally to recover the hypothesis that minimizes (true) risk, and empirical risk minimization (ERM) computes $\hat{h} = \arg\min_{h \in \mathcal{H}} \hat{R}(h; \ell, z)$ as a proxy for $h^* = \arg\min_{h \in \mathcal{H}} R(h; \ell, D)$. Analogously, for fair machine learning, we define empirical malfare minimization (EMM), given $\Lambda(\cdot; w)$, $\mathcal{D}_{1:g}$, and $z_{1:g}$, with empirical and optimal models

$$\hat{h} = \arg\min_{h \in \mathcal{H}} \Lambda(i \mapsto \hat{R}(h; \ell, z_i); w)$$

and

$$h^* = \arg\min_{h \in \mathcal{H}} \Lambda(i \mapsto R(h; \ell, D_i); w)$$

Although empirical risk is an unbiased estimate of risk, with $\Lambda_p(\cdot; w)$ for $p > 1$, empirical malfare is a biased estimator of malfare. Fortunately, empirical malfare is a consistent estimator of malfare, and we show that standard tools for bounding the error of ERM can be applied to EMM. Here it is necessary to consider not the loss function or hypothesis class in isolation, but their composition

$$\forall h \in \mathcal{H} : (\ell \circ h)(x, y) = \ell(y, h(x)) \quad \& \quad \ell \circ \mathcal{H} = \{ \ell \circ h \mid h \in \mathcal{H} \}$$

The empirical Rademacher average is a well-studied [23] data-dependent measurement of the capacity to overfit of a model class $\mathcal{H}$ w.r.t. loss function $\ell$, defined as

$$\hat{R}_m(\ell \circ \mathcal{H}, z) = \mathbb{E}_{\sigma} \left[ \sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^m \sigma_i \ell(y_i, h(x_i)) \right]$$

where $\sigma_{1:m}$ are drawn i.i.d. Rademacher (uniform on $\pm 1$), which essentially measures the ability of the hypothesis class to spuriously correlate loss values with noise.

Theorem 3.2 (Generalization Guarantees for Malfare Estimation). Suppose hypothesis class $\mathcal{H} \subseteq \mathcal{X} \rightarrow \mathcal{Y}$, bounded loss function $\ell : (\mathcal{Y} \times \mathcal{Y}) \rightarrow [0, r]$, and per-group samples $z_i \sim \mathcal{D}_i^n$. Then, with
probability at least 1 − δ over choice of z, it holds simultaneously for all fair malfare functions Λ(·; ·) (i.e., Λp(·; ·) for p ≥ 1) and probability vectors w ∈ R^n, that

\[ \sup_{h \in H} \left| \Lambda\left( i \mapsto R(h; \ell, D_i); w \right) - \Lambda\left( i \mapsto \hat{R}(h; \ell, z_i); w \right) \right| \leq \Lambda\left( i \mapsto 2\hat{R}_m(\ell \circ H, z_i) + 3r \sqrt{\frac{\ln \frac{1}{\delta}}{2m}}; w \right) . \]

The contraction property (theorem 2.4 item 3) of malfare is key to this result, and this property does not hold for welfare. Strikingly, fair welfare functions W_p(S; w) for p ∈ (0, 1) are Lipschitz-discontinuous; e.g., the Nash social welfare (geometric welfare) W_0(S; z) = \sqrt{\prod_{i = 1}^m S_i} is unstable to perturbations of each S_i around 0, which causes great difficulty in sampling-based welfare estimation. For example, if utility samples are BERNOLLI(q)-distributed for S_i, the sample complexity of ε-δ estimating W_0(S; w) grows unboundedly as w_1 → 0, q → 0 (jointly).

Experimental Validation Figure 1 presents a brief experiment on the lauded adult dataset, where the task is to predict whether income is above or below $50k/yr. We train \( \Lambda_p(·; ·) \)-minimizing SVM, and find significant risk-variation between groups; generally low risk for the white and Asian groups, and high risk for the native American and other groups. The p = 1 model is a standard weighted SVM, with poor performance for small and traditionally marginalized groups, as expected in an 85.43% majority-white population. As p increases (towards egalitarianism), we observe interesting fairness tradeoffs; training malfare increases monotonically, and in general (but not monotonically, white and Asian training risks increase, as the remaining risks decrease. At first, most improvement is in the relatively-large (9.64%), high-risk Black group, but for larger p, the much smaller (0.96%), but higher-risk, native American group sharply improves.

Both training and test performance generally improve for high-risk groups, but significant overfitting occurs in small groups and malfare. This is unsurprising, since although SVM generalization error is well-understood [see [23], chapter 26], bounds are generally vacuous for tiny subpopulations of \( \approx 400 \) individuals. In general, overfitting increases with p, due to the higher relative-importance of small but high-risk groups on h. This experiment validates EMM as a fair-learning technique, with the capacity to specify tradeoffs between majority and marginalized groups, while demonstrating overfitting to fairness, which we formally treat in the sequel. We observe similar fairness tradeoffs in our supplementary experiments (appendix A.2), on weighted and unweighted SVM and logistic regressors with race and gender groups.

### 3.2 Fair Probably Approximately Correct Learning

For context, we first present a generalized notion of PAC-learnability, which we then generalize to FPAC-learnability. Standard presentations consider only classification under 0-1 loss, but we follow the generalized learning setting of [29], which treats generic learning problems.

**Definition 3.3 (PAC-Learnability).** Suppose hypothesis class sequence \( \mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \ldots \) all over \( X \rightarrow \mathcal{Y} \), and loss function \( \ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_{0+} \). We say \( \mathcal{H} \) is PAC-learnable w.r.t. \( \ell \) if there exists a (randomized) algorithm \( \mathcal{A} \), such that for all

\[ \mathbb{P}(\mathcal{A}(\ell, \mathcal{H}) \notin \mathcal{H}^* \cup \mathcal{H}^\ell) \leq \delta \]

for all \( \delta, \varepsilon > 0 \), where \( \mathcal{H}^* \) is the class of all optimal hypotheses for \( \ell \).
1. sequence indices d;
2. instance distribution D over \( \mathcal{X} \times \mathcal{Y} \);
3. additive approximation error \( \varepsilon > 0 \); and
4. failure probability \( \delta \in (0, 1) \);

\( \mathcal{A} \) can identify a hypothesis \( \hat{h} \in \mathcal{H} \), i.e., \( \hat{h} \leftarrow \mathcal{A}(D, \varepsilon, \delta, d) \), such that

1. there exists some sample complexity function \( m(\varepsilon, \delta, d) : (\mathbb{R}_+ \times (0, 1) \times \mathbb{N}) \to \mathbb{N} \) s.t. \( \mathcal{A}(D, \varepsilon, \delta, d) \)
consumes no more than \( m(\varepsilon, \delta, d) \) samples from \( D \) (i.e., finite sample complexity); and
2. with probability at least \( 1 - \delta \) (over randomness of \( A \)), \( \hat{h} \) obeys
\[
R(\hat{h}; \ell, D) \leq \inf_{h^* \in \mathcal{H}} R(h^*; \ell, D) + \varepsilon.
\]

The class of such learning problems is denoted \( \text{PAC} \), with membership denoted \((\mathcal{H}, \ell) \in \text{PAC}\).

Furthermore, if for all \( d \), the space of \( D \) is restricted such that \( \exists h \in \mathcal{H}_d \) s.t. \( R(h; \ell, D) = 0 \), then \((\mathcal{H}, \ell) \) is \( \text{realizable-PAC-learnable} \), written \((\mathcal{H}, \ell) \in \text{PAC}^0\).

We now generalize this concept to \textit{fair-PAC learnability}. In particular, we replace the \textit{univariate risk-minimization} task with a \textit{multivariate malfare-minimization} task. Following the theory of section 2.2, we do not commit to any particular objective, but instead require that a \textit{FPAC-learner} is able to...

### Definition 3.4 (Fair-PAC (FPAC) Learnability)

Suppose hypothesis class sequence \( \mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \cdots \subseteq \mathcal{X} \rightarrow \mathcal{Y} \), and loss function \( \ell : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}_{\geq 0} \). We say \( \mathcal{H} \) is \textit{fair PAC-learnable} w.r.t. and loss function \( \ell \) if there exists a (randomized) algorithm \( A \), such that for all

1. sequence indices \( d \);
2. loss instance distributions \( D_{1,g} \) over \( (\mathcal{X} \times \mathcal{Y})^d \);
3. probability vectors \( w \in \mathbb{R}^d_+ \);
4. failure probabilities \( \delta \in (0, 1) \); and
5. instance distribution \( D \) over \( \mathcal{X} \times \mathcal{Y} \);
6. failure probabilities \( \delta \in (0, 1) \);

\( \mathcal{A} \) can identify a hypothesis \( \hat{h} \in \mathcal{H} \), i.e., \( \hat{h} \leftarrow \mathcal{A}(D_{1,g}, w, \mathcal{H}_1, \varepsilon, \delta, d) \), such that

1. there exists some sample complexity function \( m(\varepsilon, \delta, d, g) : (\mathbb{R}_+ \times (0, 1) \times \mathbb{N} \times \mathbb{N}) \to \mathbb{N} \) s.t. \( \mathcal{A}(D_{1,g}, w, \mathcal{H}_1, \varepsilon, \delta, d) \)
consumes no more than \( m(\varepsilon, \delta, d, g) \) samples (finite sample complexity); and
2. with probability at least \( 1 - \delta \) (over randomness of \( A \)), \( \hat{h} \) obeys
\[
\mathcal{L}(i \mapsto R(\hat{h}; \ell, D_{i,g}); w) \leq \inf_{h^* \in \mathcal{H}} \mathcal{L}(i \mapsto R(h^*; \ell, D_{i,g}); w) + \varepsilon.
\]

The class of such fair-learning problems is \textit{FPAC}, with membership denoted \((\mathcal{H}, \ell) \in \text{FPAC}\).

Finally, if for all \( d \), the space of \( D \) is restricted such that \( \exists h \in \mathcal{H}_d \) s.t. \( \max_{i \in 1,\ldots,g} R(h; \ell, D_i) = 0 \), then \((\mathcal{H}, \ell) \) is \textit{realizable-FPAC-learnable}, written \((\mathcal{H}, \ell) \in \text{FPAC}^0\).

### Observation 3.5 (Malfare Functions and Special Cases)

By assumption, \( \mathcal{L}_i(\cdot; \cdot) \) must be \( \mathcal{L}_p(\cdot; \cdot) \) for some \( p \in [1, \infty) \). Taking \( g = 1 \) implies \( w = (1) \), and \( \mathcal{L}_p(S; w) = S_1 \), thus reducing the problem to standard PAC-learning. Similarly, taking \( p = 1 \) converts the problem to \textit{weighted loss minimization} (weights determined by \( w \)), and \( p = \infty \) yields a \textit{minimax optimization problem}, where the max is over groups, as commonly encountered in adversarial and robust learning settings.

### On Computational Efficiency

Some authors consider not just the \textit{statistical} but also the \textit{computational} performance of learning, generally requiring that \( A \) have \textit{polynomial} time complexity (and thus implicitly sample complexity). In other words, they require that \( \mathcal{A}(D, \varepsilon, \delta, d) \) terminates in \( m(\varepsilon, \delta, d) \) steps. A similar concept of \textit{polynomial-time} FPAC-learnability is equally interesting, where here we assume \( \mathcal{A}(D_{1,g}, w, \mathcal{H}_1, \varepsilon, \delta, d) \) may be computed by a Turing machine (with access to \textit{sampling} and \textit{entropy oracles}) in \( m(\varepsilon, \delta, d, g) \in \text{Poly}(\frac{1}{\varepsilon}, \frac{1}{\delta}, d, g) \) steps. We denote these concepts \( \text{PAC}_{\text{Poly}}, \text{PAC}^0_{\text{Poly}}, \text{FPAC}_{\text{Poly}}, \text{FPAC}^0_{\text{Poly}} \).

### Observation 3.6 (FPAC-Learnability and Weighted Loss Functions)

Per-group weighted-loss functions are a well-studied object in the fairness literature, representing the idea that different types of incorrect or undesirable outcome can impact different groups in different ways.\[3\] It is not immediately obvious that the FPAC framework covers this case, however observe that it is compatible with weighted loss functions, where the weighting is associated with \textit{each individual sample}, as

---

\[3\]The experiments of section 3.1 represent a special case of this, wherein incorrect classifications disproportionately impact marginalized groups, who already have disproportionately many low-income members.
this is just a particular choice of loss function. This is actually a much more general case, but since FPAC-learnability requires uniform learnability over all distributions, this also includes distributions where per-group weightings are constant.

We also observe (immediately from definitions 3.3 and 3.4) that PAC-learning is a special case of FPAC-learning; in particular, taking \( g = 1 \) implies \( M_g(S) = M_1(S) = S_1 \), thus welfare-minimization coincides with risk minimization. The more interesting question, which we seek to answer in the remainder of this document, is when and whether the converse holds. Furthermore, when possible, we would like to show practical constructive reductions.

We now show that in the realizable case, PAC-learnability implies FPAC-learnability via a constructive polynomial-time reduction. Our reduction simply takes a sufficiently number of samples from the uniform mixture distribution over all \( g \) groups, and PAC-learns on this distribution. As the reduction is constructive (and efficiency-preserving), this gives us generic algorithms for (efficient) realizable FPAC-learning in terms of algorithms for realizable PAC-learning.

**Theorem 3.7 (Realizable Reductions).** Suppose loss function \( \ell \) and hypothesis class \( \mathcal{H} \). Then

1. \( (\mathcal{H}, \ell) \in \text{PAC}^0 \implies (\mathcal{H}, \ell) \in \text{FPAC}^0 \), and
2. \( (\mathcal{H}, \ell) \in \text{PAC}^\text{poly} \implies (\mathcal{H}, \ell) \in \text{FPAC}^\text{poly} \).

We construct a (n) (efficient) FPAC-learner for \( (\mathcal{H}, \ell) \) by noting that there exists some \( \mathcal{A}' \) with sample-complexity \( m_{\mathcal{A}'}(\varepsilon, \delta, d) \) and time complexity \( t_{\mathcal{A}'}(\varepsilon, \delta, d) \) to PAC-learn \( (\mathcal{H}, \ell), \) and taking \( \mathcal{A}(D_{1:g}, w, \mathcal{H}, \varepsilon, \delta, d) = \mathcal{A}'(\text{mix}(D_{1:g}), \frac{\varepsilon}{g}, \delta, d) \), where \( \text{mix}(D_{1:g}) \) denotes the uniformly-weighted mixture of distributions \( D_{1:g} \). Then \( \mathcal{A} \)-FPAC-learns \( (\mathcal{H}, \ell) \), with sample-complexity \( m_{\mathcal{A}'}(\varepsilon, \delta, g) = m_{\mathcal{A}'}(\frac{\varepsilon}{g}, \delta, d) \), and time-complexity \( t_{\mathcal{A}'}(\varepsilon, \delta, g) = t_{\mathcal{A}'}(\frac{\varepsilon}{g}, \delta, d) \).

Unfortunately, the argument here strongly depends on the properties of realizability, and does not extend to the agnostic case. Furthermore, we note that, philosophically speaking, realizable FPAC-learning is rather uninteresting, essentially because in a world where all parties may be satisfied completely, the obvious solution is to do so. Thus nontrivial unfairness and bias issues logically only arise in a world of conflict, e.g., in zero-sum or resource-constrained settings, which foster competition between groups. We henceforth focus our efforts on the more interesting agnostic-learning setting.

### 3.3 Characterizing Fair Statistical Learnability with FPAC-Learners

We first consider only questions of statistical learning, i.e., we ignore computation and show that there exist FPAC-learners, which is a generalization of the fundamental theorem of statistical learning to fair learning problems. The aforementioned result relates uniform convergence and PAC-learnability, and is generally stated for binary classification only. We define a natural generalization of uniform convergence to arbitrary learning problems within our framework, and then show conditions under which a generalized fundamental theorem of (fair) statistical learning holds. In particular, we show that, neglecting computational concerns, PAC-learnability and FPAC-learnability are equivalent for learning problems with a particular no-free-lunch guarantee.

We first define a generalized notion of uniform convergence. In particular, our definition applies to any bounded loss function, thus greatly generalizes the standard notion for binary classification.

**Definition 3.8 (Uniform Convergence).** Suppose \( \ell : \mathcal{Y} \times \mathcal{Y} \rightarrow [0, r] \) and hypothesis class \( \mathcal{H} \subseteq \mathcal{X} \rightarrow \mathcal{Y} \). We say \( (\mathcal{H}, \ell) \in \text{UC} \) if

\[
\lim_{m \to \infty} \sup_{D \text{ over } \mathcal{X} \times \mathcal{Y}} \sup_{z \sim D^m} \mathbb{E} \left[ \sup_{h \in \mathcal{H}} \left| \hat{R}(h; \ell, z) - R(h; \ell, D) \right| \right] = 0.
\]

We stress that this definition is both uniform over \( \ell \) composed with the hypothesis class \( \mathcal{H} \) and uniform over all possible distributions \( D \). Standard uniform convergence definitions consider only the convergence of empirical frequencies of events to their true frequencies, whereas we generalize to consider uniform convergence of the empirical means of functions to their expected values. It is also helpful to consider the sample complexity of \( \varepsilon, \delta \) uniform-convergence, where we take

\[
m_{\text{UC}}(\ell \circ \mathcal{H}, \varepsilon, \delta) = \arg\min \left\{ m \mid \sup_{D \text{ over } \mathcal{X} \times \mathcal{Y}} \mathbb{E} \left[ \sup_{h \in \mathcal{H}} \left| \hat{R}(h; \ell, D) - \hat{R}(h; \ell, \mathcal{H}) \right| > \varepsilon \right] \leq \delta \right\},
\]

i.e., the minimum sufficient sample size to ensure \( \varepsilon, \delta \) uniform-convergence over the loss family \( \ell \circ \mathcal{H} \). For classification, uniform convergence bijectively implies PAC-learnability as follows.
Figure 2: Implications between membership in PAC and FPAC classes. For arbitrary fixed \( \ell \), \( \implies \) denotes implication of membership of some \( \mathcal{H} \) (i.e., containment); see theorems 3.7 and 3.10 and =\( \implies \) holds conditionally on \( \ell \). Note that when the assumption on \( \ell \) (see theorem 3.10) holds, the hierarchy collapses, and in general, under realizability, some classes are known to coincide.

**Theorem 3.9** (Fundamental Theorem of Statistical Learning (Classification) [23, theorem 6.2]). Suppose \( \ell \) is the 0-1 loss. Then the following are equivalent:
1. \( \forall d \in \mathbb{N}: \mathcal{H}_d \) has finite VC dimension.
2. \( \forall d \in \mathbb{N}: (\ell, \mathcal{H}_d) \) has the uniform convergence property.
3. Any ERM rule is a successful agnostic-PAC learner for \( \mathcal{H} \), thus \( \mathcal{H} \) is agnostic-PAC learnable.
4. Any ERM rule is a successful realizable-PAC learner for \( \mathcal{H} \), thus \( \mathcal{H} \) is realizable-PAC learnable.

It is rather subtle to generalize this result beyond classification, as there are PAC-learnable problems for which uniform convergence does not hold. However, [1] show similar results for various regression problems, and we now consider FPAC learning (summarized in figure 2).

**Theorem 3.10** (Fundamental Theorem of Fair Statistical Learning). Suppose \( \ell \) such that \( \forall \mathcal{H}: (\mathcal{H}, \ell) \in \text{PAC}^0 \implies (\mathcal{H}, \ell) \in \text{UC} \). Then \( \forall \mathcal{H} \), the following are equivalent:
1. \( \forall d \in \mathbb{N}: (\ell, \mathcal{H}_d) \) has the generalized uniform convergence property.
2. Any EMM rule is an agnostic-FPAC learner for \( (\ell, \mathcal{H}) \), thus \( (\ell, \mathcal{H}) \) is agnostic-FPAC learnable.
3. Any EMM rule is a realizable-FPAC learner for \( (\ell, \mathcal{H}) \), thus \( (\ell, \mathcal{H}) \) is realizable-FPAC learnable.

### 3.4 Characterizing Computational Learnability with Efficient FPAC Learners

In this section, we consider the more granular question of whether FPAC learning is computationally harder than PAC learning. In other words, where previously we showed conditions under which \( \text{PAC} = \text{FPAC} \), here we focus on the subset of models with polynomial time training-efficiency guarantees, i.e., \( \text{PAC}_{\text{Poly}} = \text{FPAC}_{\text{Poly}} \). Theorem 3.7 has already characterized the computational complexity of realizable FPAC-learning, so we now focus on the agnostic case. Here we show neither a generic reduction, nor a non-constructive proof that \( \text{PAC}_{\text{Poly}} = \text{FPAC}_{\text{Poly}} \), nor do we show a counterexample, but rather we show that under standard convexity conditions often leveraged as sufficient for efficient PAC-learning, so too is efficient FPAC-learning possible.

Here we show concretely and constructively the existence of FPAC-learners under standard convex optimization assumptions via the projected subgradient method [24]. This result is immediately practical, and can be applied verbatim to problems like generalized linear models [18] and many kernel methods.

**Theorem 3.11** (Efficient FPAC Learning via Convex Optimization). Suppose each hypothesis space \( \mathcal{H}_d \in C \) is indexed by \( \Theta_d \subseteq \mathbb{R}^{\text{Poly}(d)} \), i.e., \( \mathcal{H}_d = \{ h(\cdot; \theta) \mid \theta \in \Theta_d \} \), s.t. (Euclidean) \( \text{Diam}(\Theta_d) \in \text{Poly}(d) \), and \( \forall x \in \mathcal{X}, \theta \in \Theta_d, h(x; \theta) \) can be evaluated in \( \text{Poly}(d) \) time, and \( \tilde{\theta} \in \mathbb{R}^{\text{Poly}(d)} \) can be Euclidean-projected onto \( \Theta_d \) in \( \text{Poly}(d) \) time. Suppose also \( \ell \) such that
Approximate Empirical Malfare Minimization via the Projected Subgradient Method

Algorithm 1

1: procedure $\mathcal{A}_{PSG}(\ell, H, \theta_0, m_{UC}(\cdot), D_{1:g}, w, \Lambda(\cdot), \varepsilon, \delta)$
2: Input: $\ell_r$-Lipschitz loss function $\ell$, $\lambda_H$-Lipschitz hypothesis class $H$ with parameter space $\Theta$ s.t. $\ell \circ \Theta$ is convex, initial guess $\theta_0 \in \Theta$, uniform-convergence sample-complexity bound $m_{UC}(\cdot)$, group distributions $D_{1:g}$, group weights $w \in \mathbb{R}_+^g$, welfare function $\Lambda(\cdot)$, and additive error guarantee $\varepsilon, \delta$
3: Output: $\varepsilon, \delta$-$\Lambda(\cdot)$-optimal $\hat{h} \in H$
4: $m_A \leftarrow m_{UC}(\frac{\varepsilon}{\delta}, \frac{\delta}{g}); z_{1:g,1:m_A} \sim D_{1:g} \times \cdots \times D_{m_A}$ \triangleright Draw sufficient sample for each group
5: $n \leftarrow 1 + \left( \frac{3\text{Diam}(\Theta)\lambda_H^2}{\varepsilon} \right)^2; \alpha \leftarrow \frac{3\text{Diam}(\Theta)}{\lambda_H \sqrt{n}}$ \triangleright Set iteration count $n$ and learning rate $\alpha$
6: $f(\theta) : \Theta \rightarrow \mathbb{R}_+ \triangleright$ Define empirical welfare objective
7: $\hat{\theta} \leftarrow \text{ProjectedSubgradient}(f, \theta, \theta_0, n, \alpha)$ \triangleright Run PSG algorithm on empirical welfare
8: return $\hat{h}(\hat{\theta})$
9: end procedure

$\forall x \in X, y \in Y : \theta \rightarrow \ell(y, h(x; \theta))$ is a convex function, and suppose Lipschitz constants $\lambda_\ell, \lambda_H$ are $\text{Poly}(d)$ and some norm $\|\cdot\|_Y$ over $Y$ s.t. $\ell$ is $\lambda_\ell \|\cdot\|_Y$-Lipschitz in $\hat{y}$, i.e.,

$$\forall y, y', \hat{y} \in Y : |\ell(y, \hat{y}) - \ell(y, y')| \leq \lambda_\ell \|\hat{y} - y'\|_Y,$$

and also that each $H_d$ is $\lambda_H \|\cdot\|_2 \|\cdot\|_Y$-Lipschitz in $\theta$, i.e.,

$$\forall x \in X, \theta, \theta' \in \Theta_d : \|h(x; \theta) - h(x; \theta')\|_Y \leq \lambda_H \|\theta - \theta'\|_2.$$

Finally, assume $\ell \circ \Theta_d$ exhibits $\varepsilon, \delta$ uniform convergence with sample complexity $m_{UC}(\varepsilon, \delta, d) \in \text{Poly}(\frac{1}{\varepsilon}, \frac{1}{\delta}, d)$. It then holds that, for arbitrary initial guess $\theta_0 \in \Theta_d$, for any group distributions $D_{1:g}$, group weights $w$, and fair welfare function $\Lambda(\cdot)$, the algorithm (see algorithm $\mathcal{A}$)

$$A(D_{1:g}, w, \Lambda(\cdot), \varepsilon, \delta, d) \triangleq A_{PSG}(\ell, H, \theta_0, m_{UC}(\cdot), \Theta, d, D_{1:g}, w, \Lambda(\cdot), \varepsilon, \delta)$$

FPAC-learns $(H, \ell)$ with sample complexity $m(\varepsilon, \delta, d, g) = g \cdot m_{UC}(\frac{\varepsilon}{g}, \frac{\delta}{g}, d)$, and (training) time-complexity $\in \text{Poly}(\frac{1}{\varepsilon}, \frac{1}{\delta}, d, g)$, thus $(H, \ell) \in \text{FPAC}_{\text{Poly}}$.

4 Conclusion

This paper introduces malfare minimization as a fair learning task, which we argue is better aligned to address ML tasks cast as loss minimization than is welfare. We then show statistical and computational relationships between malfare and risk minimization. We do not claim that malfare is a better or more useful concept than welfare; rather only that it is significantly different, enjoys equivalent axiomatic footing, and it stands to reason that the right tool (malfare) should be used for the task (fair risk minimization). We acknowledge that some learning tasks, e.g., bandit problems and reinforcement learning tasks, are more naturally phrased as maximizing utility or (discounted) reward, however, most supervised learning problems are naturally cast as minimizing nonnegative loss functions.

We are highly interested in exploring a parallel theory of fair welfare maximization, however some key welfare properties do not hold for welfare. In particular, non-Lipschitz welfare functions, i.e., $W_p(\cdot)$ for $p \in [0, 1)$, create great statistical and computational challenges in learning. For this reason, straightforward translation of our FPAC framework into a welfare setting is rather vacuous. In contrast, the contraction property (theorem 2.4 item 3) of malfare leads to uniform sample complexity bounds, which are crucial to FPAC-learning. It thus seems that similar bounds for welfare would need either to either impose additional assumptions to avoid non-Lipschitz behavior (e.g., artificially limit the permitted range of $p$), or otherwise provide weaker (non-uniform) learning guarantees.

As FPAC learning generalizes PAC learning, known hardness reductions and lower-bounds apply, thus the interesting computation-theoretic question is whether malfare minimization is harder than risk minimization. Theorem 3.7 answers this question in the negative under realizability, as does theorem 3.10 for sample complexity in the general (agnostic) case, under appropriate conditions on the loss function. The remaining cases are left open, though section 3.4 shows that convexity conditions sufficient for efficient PAC-learnability are also sufficient for efficient FPAC-learnability. We hope that deeper inquiry into these questions will lead to both a better understanding of what is and is not FPAC-learnable, as well as more practical and efficient reductions and FPAC-learners.
Acknowledgments

The author graciously acknowledges NSF grant RI-1813444 and DARPA/AFRL grant FA8750 for funding this work in part. The author also wishes to thank Cristina Menghini for her assistance with the experimental code, and Eli Upfal for insightful discussion regarding the intricacies of a generalized PAC-learnability concept.

References