# 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-axiomatically 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*, thus requiring 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, and is also practically relevant, as it democratizes fair ML by providing concrete training algorithms and rigorous generalization guarantees.

## 1 Introduction

It is now well-understood that contemporary ML systems exhibit differential accuracy across gender and race 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 based metrics require a notion of (positive) utility, and we argue that this is not natural to many ML tasks, where 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* rather than *utility*. We then cast fair ML as a direct *malfare minimization* problem.

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

Perhaps surprisingly, minimizing a *malfare function* is *not equivalent* to maximizing some *welfare function*, while taking utility to be negative loss (except in the trivial cases of egalitarian and utilitarian malfare). 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 notion of generic fair machine learning, termed *fair-PAC* (FPAC) *learning*, where the goal is to learn models for which finite training samples may guarantee high-probability near-optimal malfare, for *any* (axiomatically justified) malfare concept. This definition extends Valiant's ([28]) 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 *realizable 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*, then training $\varepsilon$-$\delta$ malfare-optimal models, like risk-optimal models, requires polynomial time. We briefly summarize our contributions below.

1. We derive in § 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 demographic-parity (e.g., equality of *opportunity* or *outcome*) thus correcting for some forms of data or algorithmic bias. While noble in intent and intuitive by design, fairness by demographic-parity constraints has several prominent flaws: most notably, several popular parity constraints are mutually unsatisfiable [14], 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, 13] have shown that *welfare* and even *disadvantaged group utility* can be harmed by such fairness constraints, calling into question whether demographic parity constraints even benefit those 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, 11] 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 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 a particular axiomatically justified objective, but remains simple enough to study as a mathematical object, leading to informative reductions between various PAC and FPAC learnable classes.

## 2 Quantifying Population-Level Sentiment

A generic *aggregator function* $\mathrm{M}(\mathcal{S}; \boldsymbol{w})$ summarizes some *sentiment vector* $\mathcal{S} \in \mathbb{R}^g_{0+}$, across a population of $g$ groups, weighted by *probability vector* $\boldsymbol{w} \in \mathbb{R}^g_+$ s.t. $\|\boldsymbol{w}\|_1 = 1$. When $\mathcal{S}$ measures a *desirable quantity*, generally termed *utility*, the aggregator function is a measure of *cardinal welfare* [17], and thus quantifies overall *wellbeing*. We also consider the inverse-notion, that of overall *illbeing*, termed *malfare*, in terms of an *undesirable* $\mathcal{S}$, generally *loss* or *risk*, which naturally extends the concept. We show an equivalent *axiomatic justification* for malfare, and argue that its use is more natural in many situations, particularly when considering or optimizing *loss functions*.

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

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

### 2.1 Axioms of Cardinal Welfare and Malfare

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

1. (Strict) Monotonicity: $\mathcal{S}' \neq \mathbf{0} \implies \mathrm{M}(\mathcal{S}; \boldsymbol{w}) < \mathrm{M}(\mathcal{S} + \mathcal{S}'; \boldsymbol{w})$.
2. Symmetry: $\forall$ permutations $\pi$ over $\{1, \ldots, g\}$: $\mathrm{M}(\mathcal{S}; \boldsymbol{w}) = \mathrm{M}(\pi(\mathcal{S}); \pi(\boldsymbol{w}))$.
3. Continuity: $\mathrm{M}(\mathcal{S}, \boldsymbol{w})$ is a continuous function in both $\mathcal{S}$ and $\boldsymbol{w}$.
4. Independence of unconcerned agents:
$$\mathrm{M}(\langle \mathcal{S}_{1:g-1}, \alpha \rangle; \boldsymbol{w}) \leq \mathrm{M}(\langle \mathcal{S}'_{1:g-1}, \alpha \rangle; \boldsymbol{w}) \implies \mathrm{M}(\langle \mathcal{S}_{1:g-1}, \beta \rangle; \boldsymbol{w}) \leq \mathrm{M}(\langle \mathcal{S}'_{1:g-1}, \beta \rangle; \boldsymbol{w}) \ .$$
5. Independence of common scale: $\mathrm{M}(\mathcal{S}; \boldsymbol{w}) \leq \mathrm{M}(\mathcal{S}'; \boldsymbol{w}) \implies \mathrm{M}(\alpha\mathcal{S}; \boldsymbol{w}) \leq \mathrm{M}(\alpha\mathcal{S}'; \boldsymbol{w})$.
6. Multiplicative linearity: $\mathrm{M}(\alpha\mathcal{S}; \boldsymbol{w}) = \alpha\mathrm{M}(\mathcal{S}; \boldsymbol{w})$.
7. Unit scale: $\mathrm{M}(\mathbf{1}; \boldsymbol{w}) = \mathrm{M}(1, \ldots, 1; \boldsymbol{w}) = 1$.
8. Pigou-Dalton transfer principle: Suppose $\mu = \boldsymbol{w} \cdot \mathcal{S} = \boldsymbol{w} \cdot \mathcal{S}'$, and $\forall i \in \{1, \ldots, g\} : \left|\mu - \mathcal{S}'_i\right| \leq |\mu - \mathcal{S}_i|$. Then $\mathrm{W}(\mathcal{S}'; \boldsymbol{w}) \geq \mathrm{W}(\mathcal{S}; \boldsymbol{w})$.
9. Anti-Pigou-Dalton transfer principle: Suppose as in 8, but conclude $\Mu(\mathcal{S}'; \boldsymbol{w}) \leq \Mu(\mathcal{S}; \boldsymbol{w})$.

We take a moment to comment on each axiom, to preview their purpose and assure the reader of their necessity. Axioms 1-5 are the standard *axioms of cardinal welfarism* [8, 10, 22, 20]. Together, they imply (via the Debreu-Gorman theorem) that any aggregator function can be decomposed as a *monotonic function* of a *sum* (over groups) of *log* or *power* functions. Axiom 6 is a natural and useful property, and one which enables *dimensional analysis* on mean functions; in particular, the *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. 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 $\mathcal{S}$, but *scale* does as well (making comparisons like "$\mathcal{S}_i$ is *above* the population-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*.

**The Power Mean** We now define the *p-power mean* $\mathrm{M}_p(\cdot;\cdot)$, for any $p \in \mathbb{R} \cup \pm\infty$, which we shall use to quantify both malfare and welfare. Power means arise when analyzing aggregator functions obeying the various axioms of def. 2.2, and are thus a particularly important class of aggregator function.

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

$$\mathrm{M}_p(\mathcal{S};\boldsymbol{w}) \doteq \sqrt[p]{\sum_{i=1}^{g} \boldsymbol{w}_i \mathcal{S}_i^p} \ , \ \mathrm{M}_0(\mathcal{S};\boldsymbol{w}) \doteq \exp\left(\sum_{i=1}^{g} \boldsymbol{w}_i \ln \mathcal{S}_i\right) \ , \ \mathrm{M}_{\pm\infty}(\mathcal{S};\boldsymbol{w}) \doteq \pm \max_{i\in 1,\ldots,g} \pm\mathcal{S}_i \ ,$$

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 $\mathcal{S}, \mathcal{S}' \in \mathbb{R}^g_{0+}$ and probability vector $\boldsymbol{w} \in \mathbb{R}^g_+$. The following then hold.

1. Monotonicity: $\mathrm{M}_p(\mathcal{S};\boldsymbol{w})$ is weakly-monotonically-increasing in $p$.
2. Subadditivity: $\forall p \geq 1 : \ \mathrm{M}_p(\mathcal{S}+\mathcal{S}';\boldsymbol{w}) \leq \mathrm{M}_p(\mathcal{S};\boldsymbol{w}) + \mathrm{M}_p(\mathcal{S}';\boldsymbol{w})$.
3. Contraction: $\forall p \geq 1 : \left|\mathrm{M}_p(\mathcal{S};\boldsymbol{w}) - \mathrm{M}_p(\mathcal{S}';\boldsymbol{w})\right| \leq \mathrm{M}_p(\left|\mathcal{S}-\mathcal{S}'\right|;\boldsymbol{w}) \leq \left\|\mathcal{S}-\mathcal{S}'\right\|_\infty$.
4. Curvature: $\mathrm{M}_p(\mathcal{S};\boldsymbol{w})$ is concave in $\mathcal{S}$ for $p \leq 1$ and convex for $p \geq 1$.

### 2.2 Properties of Welfare and Malfare Functions

We now derive properties of welfare and malfare from the axioms of def. 2.2.

**Theorem 2.5** (Aggregator Function Properties)**.** Suppose aggregator function $\mathrm{M}(\mathcal{S};\boldsymbol{w})$, and assume arbitrary sentiment vector $\mathcal{S} \in \mathbb{R}^g_{0+}$ and probability vector $\boldsymbol{w} \in \mathbb{R}^g_+$. If $\mathrm{M}(\cdot;\cdot)$ satisfies (subsets of) the aggregator-function axioms (see def. 2.2), then $\mathrm{M}(\cdot;\cdot)$ exhibits the following properties.

1. *Identity*: Axioms 6-7 imply that $\forall \alpha \in \mathbb{R}_{0+}$: $\mathrm{M}(\alpha\mathbf{1};\boldsymbol{w}) = \alpha$.
2. Axioms 1-5 imply $\exists p \in \mathbb{R}$, strictly-monotonically-increasing continuous $F : \mathbb{R} \to \mathbb{R}_{0+}$ s.t.
$$\mathrm{M}(\mathcal{S};\boldsymbol{w}) = F\left(\sum_{i=1}^{g} f_p(\mathcal{S}_i)\right) \ , \ \text{with} \begin{cases} p = 0 & f_0(x) \doteq \ln(x) \\ p \neq 0 & f_p(x) \doteq \mathrm{sgn}(p)x^p \end{cases} .$$
3. Axioms 1-7 imply $F(x) = f_p^{-1}(x)$, thus $\mathrm{M}(\mathcal{S};\boldsymbol{w}) = \mathrm{M}_p(\mathcal{S};\boldsymbol{w})$.
4. Axioms 1-5 and 8 imply $p \in (-\infty, 1]$.
5. Axioms 1-5 and 9 imply $p \in [1, \infty)$.

Thm. 2.5 tells us that the mild conditions of axioms 1-5 (generally assumed for welfare), along with 6 (multiplicative linearity), imply 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 multiplicative linearity, 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 malfare" meaningful. Finally, we have that $\mathrm{W}_p(\cdot;\boldsymbol{w})$ promotes equitable redistribution of utility for $p \leq 1$, and $\mathrm{M}_p(\cdot;\boldsymbol{w})$ 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 [28], 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 $\mathcal{X} \to \mathcal{Y}$. 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 $\mathcal{H}_d \doteq \left\{\vec{x} \mapsto \operatorname{sgn}\left(\vec{x} \cdot \langle \vec{w}_{1:d}, \vec{0} \rangle\right) \;\middle|\; \vec{w} \in \mathbb{R}^d\right\}$. 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 $h : \mathcal{X} \to \mathcal{Y}$ w.r.t. loss $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ on distribution $\mathcal{D}$ over $(\mathcal{X} \times \mathcal{Y})$, and the *empirical risk* on sample $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^m$, as

$$\mathrm{R}(h; \ell, \mathcal{D}) \doteq \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\ell(y, h(x))\right] \quad \& \quad \hat{\mathrm{R}}(h; \ell, \boldsymbol{z}) \doteq \mathop{\hat{\mathbb{E}}}_{(x,y)\in\boldsymbol{z}}\left[\ell(y, h(x))\right] \;,$$

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

$$\hat{h} \doteq \operatorname*{argmin}_{h\in\mathcal{H}} \Lambda\Big(i \mapsto \hat{\mathrm{R}}(h; \ell, \boldsymbol{z}_i); \boldsymbol{w}\Big) \quad \& \quad h^* \doteq \operatorname*{argmin}_{h\in\mathcal{H}} \Lambda\big(i \mapsto \mathrm{R}(h; \ell, \mathcal{D}_i); \boldsymbol{w}\big) \;.$$

Although empirical risk is an *unbiased estimate* of risk, with $\Lambda_p(\cdot; \boldsymbol{w})$ 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*

$$\forall h \in \mathcal{H} : \; (\ell \circ h)(x, y) \doteq \ell(y, h(x)) \quad \& \quad \ell \circ \mathcal{H} \doteq \{\ell \circ h \,|\, h \in \mathcal{H}\} \;.$$

The *empirical Rademacher average* is a well-studied [23, 16] data-dependent measurement of the *capacity to overfit* of a model class $\mathcal{H}$ w.r.t. loss function $\ell$, defined as

$$\hat{\mathfrak{R}}_m(\ell \circ \mathcal{H}, \boldsymbol{z}) \doteq \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} \left| \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{\sigma}_i \ell\big(\boldsymbol{y}_i, h(\boldsymbol{x}_i)\big) \right|\right] \;,$$

where $\boldsymbol{\sigma}_{1:m}$ 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 $\mathcal{H} \subseteq \mathcal{X} \to \mathcal{Y}$, bounded *loss function* $\ell : (\mathcal{Y} \times \mathcal{Y}) \to [0, r]$, and per-group samples $\boldsymbol{z}_i \sim \mathcal{D}_i^m$. Then, with probability at least $1 - \delta$ over choice of $\boldsymbol{z}$, it holds, *simultaneously for all* fair *power-mean malfare* functions $\Lambda_p(\cdot; \cdot)$ (i.e., $p \geq 1$) and *probability vectors* $\boldsymbol{w} \in \mathbb{R}_+^g$, that

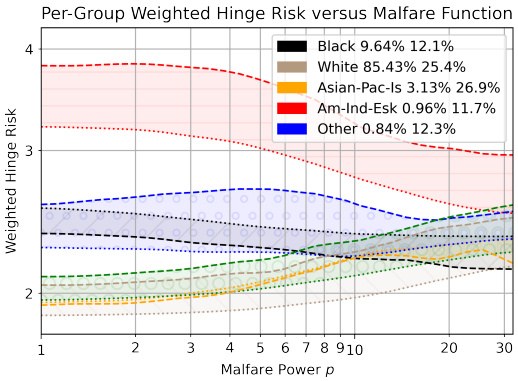
$$\sup_{h\in\mathcal{H}} \left| \Lambda_p\big(i \mapsto \mathrm{R}(h; \ell, \mathcal{D}_i); \boldsymbol{w}\big) - \Lambda_p\big(i \mapsto \hat{\mathrm{R}}(h; \ell, \boldsymbol{z}_i); \boldsymbol{w}\big) \right| \leq \Lambda_p\left(i \mapsto 2\hat{\mathfrak{R}}_m(\ell \circ \mathcal{H}, \boldsymbol{z}_i) + 3r\sqrt{\tfrac{\ln \frac{g}{\delta}}{2m}}; \boldsymbol{w}\right) \;.$$

The *contraction property* (thm. 2.4 item 3) of malfare is key to this result, and this property does not hold for welfare. Strikingly, fair welfare functions $\mathrm{W}_p(\mathcal{S}; \boldsymbol{w})$ for $p \in [0, 1)$ are Lipschitz-discontinuous; e.g., the *Nash social welfare* (geometric welfare) $\mathrm{W}_0(\mathcal{S}; \frac{1}{g}) = \sqrt[g]{\prod_{i=1}^{g} \mathcal{S}_i}$ is unstable to perturbations of each $\mathcal{S}_i$ around 0, which causes great difficulty in sampling-based welfare estimation. For example, if utility values are BERNOULLI$(q)$-distributed for some group, the sample complexity of $\varepsilon$-$\delta$ estimating $\mathrm{W}_0(\mathcal{S}; \boldsymbol{w})$ grows unboundedly as $q \to 0$.

**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 $\Lambda_p(\cdot; \boldsymbol{w})$-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[1]), *white* and *Asian* training risks increase, as the remaining risks decrease. At first, most improvement

[1]Note that for continuous loss functions and $g = 2$ groups, group training risks are monotonic in $p$, as seen in the supplementary *gender-group* experiments.

Figure 1: We minimize malfare on a *weighted hinge-loss* SVM, with $g = 5$ ethnoracial protected groups, listed in the legend with *group weight* $\boldsymbol{w}_i$ (population frequency) and *class bias* $\boldsymbol{b}_i$ (proportion with income $\geq$ \$50k/yr.). Due to existing societal inequity, *class imbalance* varies widely by group, so we weight risk as $\frac{1}{\boldsymbol{b}_i}\hat{\mathrm{R}}(h; \ell_{\text{hinge}}, \boldsymbol{z}_i)$. We report per-group training (dotted) and test (dashed) hinge risk, along with the malfare objective (green) of the EMM solution $\hat{h} \doteq \underset{h \in \mathcal{H}}{\operatorname{argmin}}\, \Lambda_p\left(i \mapsto \frac{1}{\boldsymbol{b}_i}\hat{\mathrm{R}}(h; \ell_{\text{hinge}}, \boldsymbol{z}_i); \boldsymbol{w}\right)$ , as a function of $p \in [1, 32]$. Detailed description of the experimental setup is provided in appx. A.1.




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, as although SVM generalization error is well-understoond [see 23, Chapter 26], bounds are generally vaccuous for tiny subpopulations of $\approx 400$ individuals. In general, overfitting increases with $p$, due to higher relative importance of small high-risk groups on $\hat{h}$. 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 (appx. 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 $\mathcal{X} \to \mathcal{Y}$, and *loss function* $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{0+}$. We say $\mathcal{H}$ is *PAC-learnable* w.r.t. $\ell$ if $\exists$ a (randomized) algorithm $\mathcal{A}$, such that $\forall$:

1. sequence indices $d$;
2. instance distribution $\mathcal{D}$ 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., $\hat{h} \leftarrow \mathcal{A}(\mathcal{D}, \varepsilon, \delta, d)$, such that

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

$$\mathrm{R}(\hat{h}; \ell, \mathcal{D}) \leq \inf_{h^* \in \mathcal{H}} \mathrm{R}(h^*; \ell, \mathcal{D}) + \varepsilon \ .$$

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

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

We now generalize this concept to *fair-PAC learnability*. In particular, we replace the *univariate risk-minimization* task with a *multivariate malfare-minimization* task. Following the theory of § 2.2, we do not commit to any particular objective, but instead require that a FPAC-learner is able to minimize *any* fair malfare function satisfying the standard axioms. Furthermore, here problem instances grow not just in problem complexity $d$, but also in the number of groups $g$.

**Definition 3.4** (Fair-PAC (FPAC) Learnability)**.** Suppose *hypothesis class sequence* $\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \cdots \subseteq \mathcal{X} \to \mathcal{Y}$, and *loss function* $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{0+}$. We say $\mathcal{H}$ is *fair PAC-learnable* w.r.t. and *loss function* $\ell$ if $\exists$ a (randomized) algorithm $\mathcal{A}$, such that $\forall$:

1. sequence indices $d$;
2. $g$ instance distributions $\mathcal{D}_{1:g}$ over $(\mathcal{X} \times \mathcal{Y})^g$;
3. probability vectors $\boldsymbol{w} \in \mathbb{R}_+^g$;
4. malfare concepts $\Lambda$ satisfying axioms 1-7+9;
5. additive approximation errors $\varepsilon > 0$; and
6. failure probabilities $\delta \in (0, 1)$;

$\mathcal{A}$ can identify a hypothesis $\hat{h} \in \mathcal{H}$, i.e., $\hat{h} \leftarrow \mathcal{A}(\mathcal{D}_{1:g}, \boldsymbol{w}, \Lambda, \varepsilon, \delta, d)$, such that

1. there exists some *sample complexity* function $\mathrm{m}(\varepsilon, \delta, d, g) : \big(\mathbb{R}_+ \times (0, 1) \times \mathbb{N} \times \mathbb{N}\big) \to \mathbb{N}$ s.t. $\mathcal{A}(\mathcal{D}_{1:g}, \boldsymbol{w}, \Lambda, \varepsilon, \delta, d)$ consumes no more than $\mathrm{m}(\varepsilon, \delta, d, g)$ samples (finite sample complexity); and
2. with probability at least $1 - \delta$ (over randomness of $\mathcal{A}$), $\hat{h}$ obeys

$$\Lambda\left(i \mapsto \mathrm{R}(\hat{h}; \ell, \mathcal{D}_i); \boldsymbol{w}\right) \leq \inf_{h^* \in \mathcal{H}} \Lambda\left(i \mapsto \mathrm{R}(h^*; \ell, \mathcal{D}_i); \boldsymbol{w}\right) + \varepsilon \ .$$

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

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

**Observation 3.5** (Malfare Functions and Special Cases)**.** By assumption, $\Lambda(\cdot; \cdot)$ must be $\Lambda_p(\cdot; \cdot)$ for some $p \in [1, \infty)$. Taking $g = 1$ implies $\boldsymbol{w} = \langle 1 \rangle$, and $\Lambda_p(\mathcal{S}; \boldsymbol{w}) = \mathcal{S}_1$, thus reducing the problem to standard PAC-learning. Similarly, taking $p = 1$ converts the problem to *weighted loss minimization* (weights determined by $\boldsymbol{w}$), and $p = \infty$ yields a *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 *statistical* but also the *computational* performance of learning, generally requiring that $\mathcal{A}$ have *polynomial* time complexity (and thus implicitly sample complexity). In other words, they require that $\mathcal{A}(\mathcal{D}, \varepsilon, \delta, d)$ terminates in $\mathrm{m}(\varepsilon, \delta, d) \in \mathrm{Poly}(\frac{1}{\varepsilon}, \frac{1}{\delta}, d)$ steps. A similar concept of *polynomial-time* FPAC-learnability is equally interesting, where here we assume $\mathcal{A}(\mathcal{D}_{1:g}, \boldsymbol{w}, \Lambda, \varepsilon, \delta, d)$ may be computed by a Turing machine (with access to *sampling* and *entropy* oracles) in $\mathrm{m}(\varepsilon, \delta, d, g) \in \mathrm{Poly}(\frac{1}{\varepsilon}, \frac{1}{\delta}, d, g)$ steps. We denote these concepts $\mathrm{PAC}_{\mathrm{Poly}}$, $\mathrm{PAC}^0_{\mathrm{Poly}}$, $\mathrm{FPAC}_{\mathrm{Poly}}$, and $\mathrm{FPAC}^0_{\mathrm{Poly}}$.

**Observation 3.6** (On FPAC-Learnability and Weighted Loss Functions)**.** Per-group weighted-loss functions are a well-studied object in fairness, representing the idea that different types of incorrect or undesirable answer can impact different groups in different ways.[2] It is not immediately obvious that the FPAC framework covers this case, however observe that the framework is compatible with weighted loss functions, where the weighting is associated with (the label of) *each individual sample*, as 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 defs. 3.3 and 3.4) that PAC-learning is a special case of FPAC-learning; in particular, taking $g = 1$ implies $\mathrm{M}_p(\mathcal{S}) = \mathrm{M}_1(\mathcal{S}) = \mathcal{S}_1$, thus *malfare-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. $(\mathcal{H}, \ell) \in \mathrm{PAC}^0 \implies (\mathcal{H}, \ell) \in \mathrm{FPAC}^0$; &
2. $(\mathcal{H}, \ell) \in \mathrm{PAC}^0_{\mathrm{Poly}} \implies (\mathcal{H}, \ell) \in \mathrm{FPAC}^0_{\mathrm{Poly}}$.

We construct a(n) (efficient) FPAC-learner for $(\mathcal{H}, \ell)$ by noting that there exists some $\mathcal{A}'$ with sample-complexity $\mathrm{m}_{\mathcal{A}'}(\varepsilon, \delta, d)$ and time complexity $\mathrm{t}_{\mathcal{A}'}(\varepsilon, \delta, d)$ to PAC-learn $(\mathcal{H}, \ell)$, and taking $\mathcal{A}(\mathcal{D}_{1:g}, \boldsymbol{w}, \Lambda, \varepsilon, \delta, d) \doteq \mathcal{A}'(\mathrm{mix}(\mathcal{D}_{1:g}), \frac{\varepsilon}{g}, \delta, d)$. Then $\mathcal{A}$ FPAC-learns $(\mathcal{H}, \ell)$, with sample-complexity $\mathrm{m}_{\mathcal{A}}(\varepsilon, \delta, d, g) = \mathrm{m}_{\mathcal{A}'}(\frac{\varepsilon}{g}, \delta, d)$, and time-complexity $\mathrm{t}_{\mathcal{A}}(\varepsilon, \delta, d, g) = \mathrm{t}_{\mathcal{A}'}(\frac{\varepsilon}{g}, \delta, d)$.

[2] The experiments of § 3.1 represent a special case of this, wherein incorrect classifications disproportionately impact marginalized groups.

Unfortunately, the argument here strongly depends on the properties of realizablility, 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-learning algorithms. In particular, we show 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 $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, r]$ and hypothesis class $\mathcal{H} \subseteq \mathcal{X} \to \mathcal{Y}$. We say $(\mathcal{H}, \ell) \in \mathrm{UC}$ if

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

We stress that this definition is both uniform over $\ell$ composed with the *hypothesis class* $\mathcal{H}$ and uniform over *all possible distributions* $\mathcal{D}$. 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

$$\mathrm{m}_{\mathrm{UC}}(\ell \circ \mathcal{H}, \varepsilon, \delta) \doteq \operatorname{argmin} \left\{ m \ \middle| \ \sup_{\mathcal{D} \text{ over } \mathcal{X}\times\mathcal{Y}} \mathbb{P}\left( \sup_{h\in\mathcal{H}} \left| \mathop{\mathbb{E}}_{\mathcal{D}}[\ell \circ h] - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{z}\sim\mathcal{D}^m}[\ell \circ h] \right| > \varepsilon \right) \leq \delta \right\} \ ,$$

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.

**Theorem 3.9** (Fundamental Theorem of Statistical Learning (Classification) [23, thm. 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 fig. 2).

**Theorem 3.10** (Fundamental Theorem of Fair Statistical Learning)**.** Suppose $\ell$ such that $\forall \mathcal{H}$ : $(\mathcal{H}, \ell) \in \mathrm{PAC}^0 \implies (\mathcal{H}, \ell) \in \mathrm{UC}$. 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 $\mathrm{PAC} = \mathrm{FPAC}$, here we focus on the subset of models with polynomial time training-efficiency

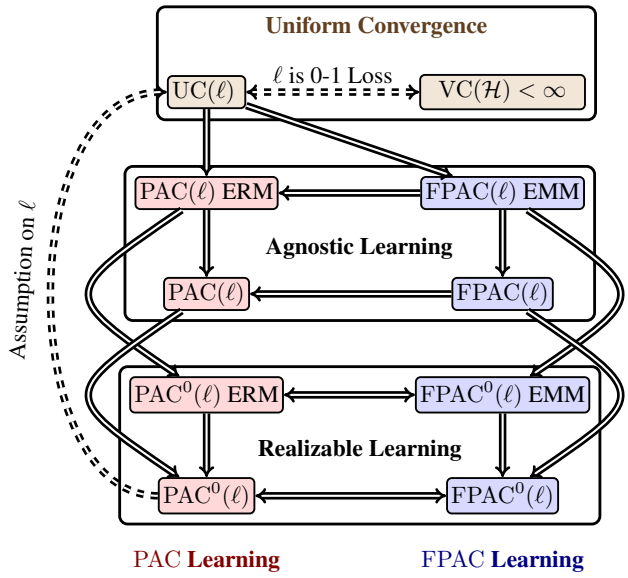



Figure 2: Implications between membership in PAC and FPAC classes. In particular, for arbitrary fixed $\ell$, implication denotes *implication of membership* of some $\mathcal{H}$ (i.e., containment); see thms. 3.7 and 3.10. Dashed implication arrows hold conditionally on $\ell$. Note that when the assumption on $\ell$ (see thm. 3.10) holds, the hierarchy collapses, and in general, under realizability, some classes are known to coincide.

guarantees, i.e., $\mathrm{PAC}_{\mathrm{Poly}} = \mathrm{FPAC}_{\mathrm{Poly}}$. Thm. 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 $\mathrm{PAC}_{\mathrm{Poly}} = \mathrm{FPAC}_{\mathrm{Poly}}$, 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 *subgradient method* [24]. This result is immediately practical, and can be applied verbatim to problems like *generalized linear models* [18] and many *kernel methods*.

**Algorithm 1** Approximate Empirical Malfare Minimization via the Subgradient Method

**procedure** $\mathcal{A}_{\mathrm{PSG}}(\ell, \mathcal{H}, \theta_0, \mathrm{m}_{\mathrm{UC}}(\cdot,\cdot), \mathcal{D}_{1:g}, \boldsymbol{w}, \mathrm{M}(\cdot;\cdot), \varepsilon, \delta)$
  **Input**: $\lambda_\ell$-Lipschitz loss function $\ell$, $\lambda_{\mathcal{H}}$-Lipschitz hypothesis class $\mathcal{H}$ with parameter space $\Theta$ s.t. $\ell \circ \mathcal{H}$ is convex, initial guess $\theta_0 \in \Theta$, uniform-convergence sample-complexity bound $\mathrm{m}_{\mathrm{UC}}(\cdot,\cdot)$, group distributions $\mathcal{D}_{1:g}$, group weights $\boldsymbol{w} \in \mathbb{R}_+^g$, malfare function $\mathrm{M}(\cdot;\cdot)$, and additive error guarantee $\varepsilon$-$\delta$
  **Output**: $\varepsilon$-$\delta$-$\mathrm{M}(\cdot;\cdot)$-optimal $\hat{h} \in \mathcal{H}$
  $\mathrm{m}_{\mathcal{A}} \leftarrow \mathrm{m}_{\mathrm{UC}}(\frac{\varepsilon}{3}, \frac{\delta}{g})$; $\boldsymbol{z}_{1:g,1:\mathrm{m}_{\mathcal{A}}} \sim \mathcal{D}_1^{\mathrm{m}_{\mathcal{A}}} \times \cdots \times \mathcal{D}_g^{\mathrm{m}_{\mathcal{A}}}$ ▷ Draw sufficient sample for each group
  $n \leftarrow 1 + \left(\frac{3\,\mathrm{Diam}(\Theta)\lambda_\ell\lambda_{\mathcal{H}}}{\varepsilon}\right)^2$; $\alpha \leftarrow \frac{\mathrm{Diam}(\Theta)}{\lambda_\ell\lambda_{\mathcal{H}}\sqrt{n}}$ ▷ Set iteration count $n$ and learning rate $\alpha$
  $f(\theta) : \Theta \mapsto \mathbb{R}_{0+} \doteq \mathrm{M}\big(i \mapsto \hat{\mathrm{R}}(h(\cdot;\theta); \ell, \boldsymbol{z}_i); \boldsymbol{w}\big)$ ▷ Define empirical malfare objective
  $\hat{\theta} \leftarrow \textsc{ProjectedSubgradient}(f, \Theta, \theta_0, n, \alpha)$ ▷ Run PSG algorithm on empirical malfare
  **return** $h(\cdot;\hat{\theta})$ ▷ Return $\varepsilon$-$\delta$ optimal model
**end procedure**

**Theorem 3.11** (Efficient FPAC Learning via Convex Optimization)**.** Suppose each hypothesis space $\mathcal{H}_d \in \mathcal{H}$ is indexed by $\Theta_d \subseteq \mathbb{R}^{\mathrm{Poly}(d)}$, i.e., $\mathcal{H}_d = \{h(\cdot;\theta) \mid \theta \in \Theta_d\}$, s.t. (Euclidean) $\mathrm{Diam}(\Theta_d) \in \mathrm{Poly}(d)$, and $\forall x \in \mathcal{X}, \theta \in \Theta_d$, $h(x;\theta)$ can be evaluated in $\mathrm{Poly}(d)$ time, and $\tilde{\theta} \in \mathbb{R}^{\mathrm{Poly}(d)}$ can be Euclidean-projected onto $\Theta_d$ in $\mathrm{Poly}(d)$ time. Suppose also $\ell$ such that $\forall x \in \mathcal{X}, y \in \mathcal{Y} : \theta \mapsto \ell(y, h(x;\theta))$ is a *convex function*, and suppose Lipschitz constants $\lambda_\ell, \lambda_{\mathcal{H}} \in \mathrm{Poly}(d)$ and some norm $\|\cdot\|_{\mathcal{Y}}$ over $\mathcal{Y}$ s.t. $\ell$ is $\lambda_\ell \|\cdot\|_{\mathcal{Y}} \|\cdot\|$-Lipschitz in $\hat{y}$, i.e.,

$$\forall y, \hat{y}, \hat{y}' \in \mathcal{Y} : \left|\ell(y,\hat{y}) - \ell(y,\hat{y}')\right| \leq \lambda_\ell \left\|\hat{y} - \hat{y}'\right\|_{\mathcal{Y}} \quad ,$$

and also that each $\mathcal{H}_d$ is $\lambda_{\mathcal{H}} \|\cdot\|_2 \|\cdot\|_{\mathcal{Y}}$-Lipschitz in $\theta$, i.e.,

$$\forall x \in \mathcal{X}, \theta, \theta' \in \Theta_d : \left\|h(x;\theta) - h(x;\theta')\right\|_{\mathcal{Y}} \leq \lambda_{\mathcal{H}} \left\|\theta - \theta'\right\|_2 \quad .$$

Finally, assume $\ell \circ \mathcal{H}_d$ exhibits $\varepsilon$-$\delta$ *uniform convergence* with sample complexity $\mathrm{m}_{\mathrm{UC}}(\varepsilon, \delta, d) \in \mathrm{Poly}(\frac{1}{\varepsilon}, \frac{1}{\delta}, d)$. It then holds that, for arbitrary initial guess $\theta_0 \in \Theta_d$, for any group distributions $\mathcal{D}_{1:g}$, group weights $\boldsymbol{w}$, and fair malfare function $\Lambda(\cdot;\cdot)$, the algorithm (see alg. 1)

$$\mathcal{A}(\mathcal{D}_{1:g}, \boldsymbol{w}, \Lambda(\cdot;\cdot), \varepsilon, \delta, d) \doteq \mathcal{A}_{\mathrm{PSG}}\big(\ell, \mathcal{H}_d, \theta_0, \mathrm{m}_{\mathrm{UC}}(\cdot,\cdot,d), \mathcal{D}_{1:g}, \boldsymbol{w}, \Lambda(\cdot;\cdot), \varepsilon, \delta\big)$$

FPAC-learns $(\mathcal{H}, \ell)$ with sample complexity $\mathrm{m}(\varepsilon, \delta, d, g) = g \cdot \mathrm{m}_{\mathrm{UC}}(\frac{\varepsilon}{3}, \frac{\delta}{g}, d)$, and (training) time-complexity $\in \mathrm{Poly}(\frac{1}{\varepsilon}, \frac{1}{\delta}, d, g)$, thus $(\mathcal{H}, \ell) \in \mathrm{FPAC}_{\mathrm{Poly}}$.

# 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*, stands on an equal axiomatic footing, and it stands to reason that the right tool (malfare) should be used for the right job (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 malfare properties do not hold for welfare. In particular, non-Lipschitz welfare functions (i.e., $p \in [0, 1)$), create great difficulty in both the *statistical* and *computational* aspects of learning. For this reason, straightforward translation of our FPAC framework into a welfare setting is rather vacuous. In contrast, the *contraction property* (thm. 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 generalizes PAC, known hardness results (e.g., lower-bounds on learning complexity) apply, thus the interesting question is whether *malfare minimization* is harder than *risk minimization*. Thm. 3.7 answers this question in the negative *under realizability*, as does thm. 3.10 for *sample complexity* in the general (agnostic) case, under appropriate conditions on the loss function. The remaining cases are left open, though § 3.4 shows that convexity conditions sufficient for PAC-learnability are also sufficient for 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-learning algorithms.

# 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

[1] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM (JACM)*, 44(4):615–631, 1997.

[2] A. Ashraf, S. Khan, N. Bhagwat, M. Chakravarty, and B. Taati. Learning to unlearn: Building immunity to dataset bias in medical imaging studies. *arXiv preprint arXiv:1812.01716*, 2018.

[3] P. S. Bullen. *Handbook of means and their inequalities*, volume 560. Springer Science & Business Media, 2013.

[4] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[5] J. G. Cavazos, P. J. Phillips, C. D. Castillo, and A. J. O'Toole. Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2020.

[6] C. M. Cook, J. J. Howard, Y. B. Sirotin, J. L. Tipton, and A. R. Vemury. Demographic effects in facial recognition and their dependence on image acquisition: An evaluation of eleven commercial systems. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 1(1):32–41, 2019.

[7] H. Dalton. The measurement of the inequality of incomes. *The Economic Journal*, 30(119):348–361, 1920.
[8] G. Debreu. Topological methods in cardinal utility theory. Technical report, Cowles Foundation for Research in Economics, Yale University, 1959.
[9] D. Dua and C. Graff. UCI machine learning repository, 2021.
[10] W. M. Gorman. The structure of utility functions. *The Review of Economic Studies*, 35(4):367–390, 1968.
[11] H. Heidari, C. Ferrari, K. Gummadi, and A. Krause. Fairness behind a veil of ignorance: A welfare analysis for automated decision making. In *Advances in Neural Information Processing Systems*, pages 1265–1276, 2018.
[12] L. Hu and Y. Chen. Fair classification and social welfare. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 535–545, 2020.
[13] M. Kasy and R. Abebe. Fairness, equality, and power in algorithmic decision-making. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 576–586, 2021.
[14] J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67, page 43. Schloß Dagstuhl–Leibniz-Zentrum für Informatik, 2017.
[15] B. Mac Namee, P. Cunningham, S. Byrne, and O. I. Corrigan. The problem of bias in training data in regression problems in medical decision support. *Artificial intelligence in medicine*, 24(1):51–70, 2002.
[16] M. Mitzenmacher and E. Upfal. *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis*. Cambridge university press, second edition, 2017.
[17] H. Moulin. *Fair division and collective welfare*. MIT Press, 2004.
[18] J. A. Nelder and R. W. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972.
[19] A. C. Pigou. *Wealth and welfare*. Macmillan and Company, limited, 1912.
[20] K. W. Roberts. Interpersonal comparability and social choice theory. *The Review of Economic Studies*, pages 421–439, 1980.
[21] E. Rolf, M. Simchowitz, S. Dean, L. T. Liu, D. Björkegren, M. Hardt, and J. Blumenstock. Balancing competing objectives with noisy data: Score-based classifiers for welfare-aware machine learning. *arXiv preprint arXiv:2003.06740*, 2020.
[22] A. Sen. On weights and measures: Informational constraints in social welfare analysis. *Econometrica: Journal of the Econometric Society*, pages 1539–1572, 1977.
[23] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.
[24] N. Z. Shor. *Minimization methods for non-differentiable functions*, volume 3. Springer Science & Business Media, 2012.
[25] U. Siddique, P. Weng, and M. Zimmer. Learning fair policies in multi-objective (deep) reinforcement learning with average and discounted rewards. In *International Conference on Machine Learning*, pages 8905–8915. PMLR, 2020.
[26] T. Speicher, H. Heidari, N. Grgic-Hlaca, K. P. Gummadi, A. Singla, A. Weller, and M. B. Zafar. A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2239–2248, 2018.
[27] P. S. Thomas, B. C. da Silva, A. G. Barto, S. Giguere, Y. Brun, and E. Brunskill. Preventing undesirable behavior of intelligent machines. *Science*, 366(6468):999–1004, 2019.
[28] L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.
[29] V. Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.