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 machine learning (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*, 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 ε-δ malfare-optimal model with bounded sample complexity, for *any data distribution* and (axiomatically justified) malfare concept. Finally, we show conditions under which many standard PAC-learners may be converted to fair-PAC learners, which places fair-PAC learning on firm theoretical ground, as it yields statistical — and in some cases computational — efficiency guarantees for many well-studied ML 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 [5–7], in medical settings [2, 17], 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.

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

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 ML, termed *fair-PAC* (FPAC) *learning*, where the goal is to learn models for which finite training samples may guarantee (additively) ε-optimal malfare, with probability at least $1 - \delta$, for *any* (axiomatically justified) malfare concept. This definition extends Valiant's [31] 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* under standard assumptions, then training ε-δ malfare-optimal models, like training 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* ML 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 [16], 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 [14, 15] 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 [14, 23, 28] or constrained [13, 29] to promote *fair learning* across *all groups*. Perhaps the most similar to our work is a method of [14], 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 ML contexts, there is no "correct" or clearly neutral way to generically 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* [30] 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 axiomatically-justified objective, but remains simple enough to study as a mathematical object, leading to informative reductions between various PAC and FPAC learnable classes. A similar framework, termed "multi-group agnostic PAC learning" [3, 24], also considers per-group risk values, but there *regret* (maximum relative dissatisfaction), rather than malfare, is minimized.

2 Aggregating Sentiment within Populations

A generic *aggregator function* $\mathrm{M}(\mathcal{S};\boldsymbol{w})$ summarizes some *sentiment vector* $\mathcal{S} \in \mathbb{R}_{0+}^{g}$, across a population of g groups, weighted by a *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* [19], 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 welfare 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 $\rotatebox[origin=c]{180}{W}(\mathcal{S};\boldsymbol{w})$.

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 value* $\mathrm{M}(\mathcal{S};\boldsymbol{w})$, rather than a single group's 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 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; $\mathrm{M}(\mathcal{S};\boldsymbol{w})$ is M for *mean*, whereas $\mathrm{W}(\mathcal{S};\boldsymbol{w})$ is W for *welfare*, and $\rotatebox[origin=c]{180}{W}(\mathcal{S};\boldsymbol{w})$ is $\rotatebox[origin=c]{180}{W}$ (*inverted* 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 $\forall \mathcal{S}, \mathcal{S}' \in \mathbb{R}_{0+}^{g}$, scalars $\alpha, \beta \in \mathbb{R}_{0+}$, and probability vectors $\boldsymbol{w} \in \mathbb{R}_{+}^{g}$.

1. *Strict Monotonicity*: $\mathcal{S}' \neq \mathbf{0} \implies \mathrm{M}(\mathcal{S};\boldsymbol{w}) < \mathrm{M}(\mathcal{S}+\mathcal{S}';\boldsymbol{w})$.
2. *Weighted Symmetry*:[1] Suppose $g' \in \mathbb{Z}_{+}$, $\mathcal{S}' \in \mathbb{R}_{0+}^{g'}$, and probability vector $\boldsymbol{w}' \in \mathbb{R}_{0+}^{g'}$, such that for all $u \in \mathbb{R}_{0+}$, it holds that $\sum_{i \text{ s.t. } \mathcal{S}_i = u} \boldsymbol{w}_i = \sum_{i \text{ s.t. } \mathcal{S}'_i = u} \boldsymbol{w}'_i$. Then $\mathrm{M}(\mathcal{S};\boldsymbol{w}) = \mathrm{M}(\mathcal{S}';\boldsymbol{w}')$.
3. *Continuity*: $\mathrm{M}(\mathcal{S};\boldsymbol{w})$ is a continuous function in both $\mathcal{S}$ and $\boldsymbol{w}$.
4. *Independence of Unconcerned Agents*:
$$\mathrm{M}(\langle \mathcal{S}_{1:g-1}, \alpha\rangle;\boldsymbol{w}) \leq \mathrm{M}(\langle \mathcal{S}'_{1:g-1}, \alpha\rangle;\boldsymbol{w}) \implies \mathrm{M}(\langle \mathcal{S}_{1:g-1}, \beta\rangle;\boldsymbol{w}) \leq \mathrm{M}(\langle \mathcal{S}'_{1:g-1}, \beta\rangle;\boldsymbol{w}) \enspace .$$
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*: $\mathrm{M}(\mathbf{1};\boldsymbol{w}) = \mathrm{M}(\langle 1,\dots,1\rangle;\boldsymbol{w}) = 1$.
8. *Pigou-Dalton Transfer Principle*: Suppose $\mu = \boldsymbol{w}\cdot\mathcal{S} = \boldsymbol{w}\cdot\mathcal{S}'$, and for all $i \in \{1,\dots,g\}$: $|\mu - \mathcal{S}'_i| \leq |\mu - \mathcal{S}_i|$. Then $\mathrm{W}(\mathcal{S}';\boldsymbol{w}) \geq \mathrm{W}(\mathcal{S};\boldsymbol{w})$.
9. *Negated Pigou-Dalton Transfer Principle*: Suppose as in 8, but conclude $\rotatebox[origin=c]{180}{W}(\mathcal{S}';\boldsymbol{w}) \leq \rotatebox[origin=c]{180}{W}(\mathcal{S};\boldsymbol{w})$.

We take a moment to comment on each axiom, to preview their purpose and assure the reader of their necessity. Axioms 1-5 are essentially the standard *axioms of cardinal welfarism* [22, 25]. Together, they imply (via the Debreu-Gorman theorem [10, 12]) that any aggregator function can be decomposed as a *monotonic function* of a *sum* (over groups) of *logarithm* 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

[1] In the unweighted case, it is standard to define symmetry as simply $\mathrm{M}(\mathcal{S}) = \mathrm{M}\big(\pi(\mathcal{S})\big)$ for all permutations π over $\{1,\dots,g\}$, but with weightings, the simple extension $\mathrm{M}(\mathcal{S};\boldsymbol{w}) = \mathrm{M}\big(\pi(\mathcal{S});\pi(\boldsymbol{w})\big)$ is not quite sufficient. In particular, in this weaker form, nowhere is *additive decomposability*, wherein a group may be decomposed into multiple groups of the same sentiment value and total weight without impacting the aggregate, codified.

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's welfare" meaningful), and also enabling comparison *across populations* (i.e., comparing *averages* rather than *sums*).

Finally, axiom 8, the *Pigou-Dalton transfer principle* [9, 21], 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 $\boldsymbol{w}$, wherein *continuity* (3) must also hold over $\boldsymbol{w}$, 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* $\mathrm{M}_p(\cdot;\cdot)$, 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 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 $\mathcal{S} \in \mathbb{R}_{0+}^g$ 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 $\mathcal{S}, \mathcal{S}' \in \mathbb{R}_{0+}^g$ 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 in $\mathcal{S}$ 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 $\mathrm{M}(\mathcal{S};\boldsymbol{w})$, and assume arbitrary sentiment vector $\mathcal{S} \in \mathbb{R}_{0+}^g$ and probability vector $\boldsymbol{w} \in \mathbb{R}_+^g$. If $\mathrm{M}(\cdot;\cdot)$ satisfies (subsets of) the aggregator-function axioms (see definition 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. *Debreu-Gorman Factorization*: Axioms 1-5 imply $\exists p \in \mathbb{R}$, strictly-monotonically-increasing continuous $F : \mathbb{R} \to \mathbb{R}_{0+}$ s.t.

$$\mathrm{M}(\mathcal{S};\boldsymbol{w}) = F\left(\sum_{i=1}^{g} \boldsymbol{w}_i f_p(\mathcal{S}_i)\right) \ , \quad \text{with} \begin{cases} p = 0 & f_0(x) \doteq \ln(x) \\ p \neq 0 & f_p(x) \doteq \mathrm{sgn}(p)x^p \end{cases} \ .$$

3. *Power-Mean Factorization*: 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. *Fair Welfare*: Axioms 1-5 and 8 imply $p \in (-\infty, 1]$.
5. *Fair Malfare*: Axioms 1-5 and 9 imply $p \in [1, \infty)$.

Theorem 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 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 $\mathrm{W}_p(\cdot;\boldsymbol{w})$ 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 [31], 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

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

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 ML*, we define *empirical malfare minimization* (EMM), given $\mathrm{M}(\cdot;\boldsymbol{w})$, $\mathcal{D}_{1:g}$, and $\boldsymbol{z}_{1:g}$, with empirical and true optimal models

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

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

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

The *empirical Rademacher average* is a well-studied [8, 18, 26] data-dependent measurement of the *capacity to overfit* of a model class $\mathcal{H}$ w.r.t. loss function ℓ, defined as

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

where $\boldsymbol{\sigma}_{1:m}$ are drawn i.i.d. Rademacher (i.e., uniform on ± 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 malfare functions* $\mathrm{M}(\cdot;\cdot)$ (i.e., $\mathrm{M}_p(\cdot;\cdot)$ for $p \geq 1$) and *probability vectors* $\boldsymbol{w} \in \mathbb{R}_+^g$, that

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

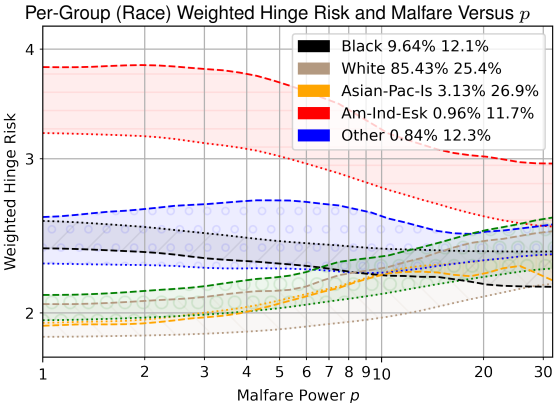


Figure 1: We minimize malfare of a *weighted hinge-loss* SVM, with $g = 5$ ethnoracial protected groups, listed in the legend with *group weight* $\boldsymbol{w}_i$ (population frequency) and *class bias* $\boldsymbol{b}_i$ (proportion with income $\geq$ \$50k/yr.). Due to existing societal inequity,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

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 $\mathrm{W}_p(\mathcal{S}; \boldsymbol{w})$ 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 samples are $\textsc{Bernoulli}(q)$-distributed for $\mathcal{S}_1$, the sample complexity of ε-δ estimating $\mathrm{W}_0(\mathcal{S}; \boldsymbol{w})$ grows unboundedly as $\boldsymbol{w}_1 \to 0$, $q \to 0$ (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 $\mathrm{M}_p(\cdot; \boldsymbol{w})$-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[2]), *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 26, chapter 26], bounds are generally vacuous for tiny subpopulations of ≈ 400 individuals. In general, overfitting increases with p, due to the higher relative-importance of small but 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 (appendix A.2), on weighted and unweighted SVM and logistic regressors with *race* and *sex* 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 [32], which treats generic learning problems.

Definition 3.3 (PAC-Learnability)**.** Suppose *hypothesis class sequence* $\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \dots$, 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. ℓ if there exists a (randomized) algorithm $\mathcal{A}$, such that for all

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 $\delta \in (0, 1)$;

[2] Note that for continuous loss functions and $g = 2$ groups, group training risks are monotonic in p, as seen in the supplementary *sex-based* experiments (see appendix A.2).

$\mathcal{A}$ 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 $\mathrm{m}(\varepsilon, \delta, d) : \left(\mathbb{R}_+ \times (0,1) \times \mathbb{Z}_+\right) \to \mathbb{Z}_+$ 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

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

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

Furthermore, if for all d, the space of $\mathcal{D}$ is restricted such that $\inf_{h \in \mathcal{H}_d} \mathrm{R}(h; \ell, \mathcal{D}) = 0$, 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 section 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. ℓ if there exists a (randomized) algorithm $\mathcal{A}$, such that for all

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 M 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., $\hat{h} \leftarrow \mathcal{A}(\mathcal{D}_{1:g}, \boldsymbol{w}, \mathrm{M}, \varepsilon, \delta, d)$, such that

1. there exists some *sample complexity* function $\mathrm{m}(\varepsilon, \delta, d, g) : \left(\mathbb{R}_+ \times (0,1) \times \mathbb{Z}_+ \times \mathbb{Z}_+\right) \to \mathbb{Z}_+$ s.t. $\mathcal{A}(\mathcal{D}_{1:g}, \boldsymbol{w}, \mathrm{M}, \varepsilon, \delta, d)$ 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

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

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 $\inf_{h \in \mathcal{H}_d} \max_{i \in 1, \dots, g} \mathrm{R}(h; \ell, \mathcal{D}_i) = 0$, 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, $\mathrm{M}(\cdot; \cdot)$ must be $\mathrm{M}_p(\cdot; \cdot)$ for some $p \in [1, \infty)$. Taking $g = 1$ implies $\boldsymbol{w} = \langle 1 \rangle$, and $\mathrm{M}_p(\mathcal{S}; \boldsymbol{w}) = \mathcal{S}_1$, 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 $\mathcal{A}(\mathcal{D}_{1:g}, \boldsymbol{w}, \mathrm{M}, \varepsilon, \delta, d)$ 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 (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 *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.

[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.

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 $\mathrm{M}_p(\mathcal{S}) = \mathrm{M}_1(\mathcal{S}) = \mathcal{S}_1$, 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 ℓ and hypothesis class $\mathcal{H}$. Then

1. $(\mathcal{H}, \ell) \in \mathrm{PAC}^0 \implies (\mathcal{H}, \ell) \in \mathrm{FPAC}^0$; and 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 $\mathcal{A}(\mathcal{D}_{1:g}, \boldsymbol{w}, \mathbb{M}, \varepsilon, \delta, d) \doteq \mathcal{A}'(\mathrm{mix}(\mathcal{D}_{1:g}), \frac{\varepsilon}{g}, \delta, d)$, where $\mathrm{mix}(\mathcal{D}_{1:g})$ denotes the uniformly-weighted mixture of distributions $\mathcal{D}_{1:g}$. 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 $\mathrm{t}_{\mathcal{A}}(\varepsilon, \delta, d, g) = \mathrm{t}_{\mathcal{A}'}(\frac{\varepsilon}{g}, \delta, d)$.

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-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

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

We stress that this definition is both uniform over ℓ 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 ε-δ uniform-convergence, where we take

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

i.e., the *minimum sufficient sample size* to ensure ε-δ 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 [26, theorem 6.2])**.** Suppose ℓ is the 0-1 loss. Then the following are equivalent:

1. $\forall d \in \mathbb{Z}_+$: $\mathcal{H}_d$ has finite Vapnik-Chervonenkis (VC) dimension [33, 34].

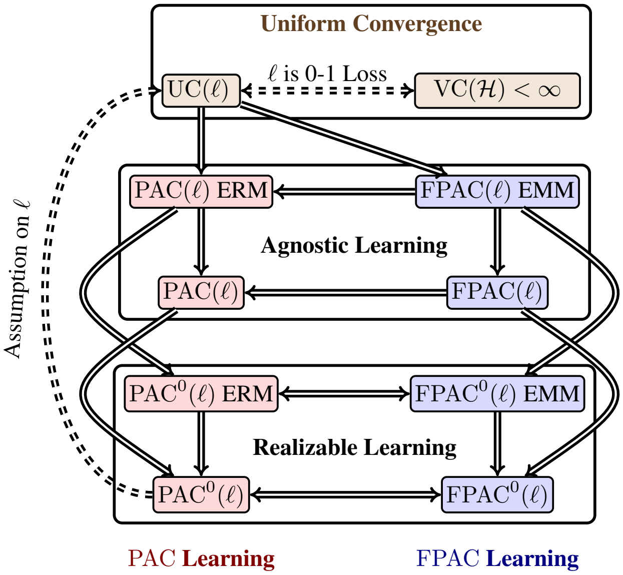


Figure 2: Implications between membership in PAC and FPAC classes. For arbitrary fixed ℓ, $\Longrightarrow$ denotes *implication of membership* of some $\mathcal{H}$ (i.e., containment); see theorems 3.7 and 3.10, and $= =\Rightarrow$ holds conditionally on ℓ. Note that when the assumption on ℓ (see theorem 3.10) holds, the hierarchy collapses, and in general, under realizability, some classes are known to coincide.

2. $\forall d \in \mathbb{Z}_+$: $(\ell, \mathcal{H}_d)$ 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 generalize to FPAC learning (summarized in figure 2).

Theorem 3.10 (Fundamental Theorem of Fair Statistical Learning)**.** Suppose ℓ 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. $\forall d \in \mathbb{Z}_+$: $(\ell, \mathcal{H}_d)$ 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 guarantees, i.e., $\mathrm{PAC}_{\mathrm{Poly}} = \mathrm{FPAC}_{\mathrm{Poly}}$. 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 $\mathrm{PAC}_{\mathrm{Poly}} = \mathrm{FPAC}_{\mathrm{Poly}}$, 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 *projected subgradient method* [27]. This result is immediately practical, and can be applied verbatim to problems like *generalized linear models* [20] and many *kernel methods*.

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 Θ_d in $\mathrm{Poly}(d)$ time. Suppose also ℓ 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. ℓ is λ_ℓ-$\|\cdot\|_{\mathcal{Y}}$-$|\cdot|$-Lipschitz in $\hat{y}$, i.e.,

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

Algorithm 1 Approximate Empirical Malfare Minimization via the Projected 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
```

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

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

Finally, assume $\ell \circ \mathcal{H}_d$ exhibits ε-δ *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 $\mathrm{M}(\cdot; \cdot)$, the algorithm (see algorithm 1)

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

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*, 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 malfare properties do not hold for welfare. In particular, non-Lipschitz welfare functions, i.e., $\mathrm{W}_p(\cdot; \cdot)$ 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

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

[2] Ahmed Ashraf, Shehroz Khan, Nikhil Bhagwat, Mallar Chakravarty, and Babak Taati. Learning to unlearn: Building immunity to dataset bias in medical imaging studies. *Machine Learning for Health (ML4H) Workshop at Advances at Neural Information Processing Systems 31*, 2018.

[3] Avrim Blum and Thodoris Lykouris. Advancing subgroup fairness via sleeping experts. In *Innovations in Theoretical Computer Science Conference (ITCS)*, volume 11, 2020.

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

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

[6] Jacqueline G Cavazos, P Jonathon Phillips, Carlos D Castillo, and Alice 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.

[7] Cynthia M Cook, John J Howard, Yevgeniy B Sirotin, Jerry L Tipton, and Arun 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.

[8] Cyrus Cousins and Matteo Riondato. Sharp uniform convergence bounds through empirical centralization. *Advances in Neural Information Processing Systems*, 33, 2020.

[9] Hugh Dalton. The measurement of the inequality of incomes. *The Economic Journal*, 30(119): 348–361, 1920.

[10] Gerard Debreu. Topological methods in cardinal utility theory. 1959.

[11] Dheeru Dua and Casey Graff. UCI machine learning repository, 2021. URL `http://archive.ics.uci.edu/ml`.

[12] William M Gorman. The structure of utility functions. *The Review of Economic Studies*, 35(4): 367–390, 1968.

[13] Hoda Heidari, Claudio Ferrari, Krishna Gummadi, and Andreas 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.

[14] Lily Hu and Yiling Chen. Fair classification and social welfare. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 535–545, 2020.

[15] Maximilian Kasy and Rediet 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.

[16] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67, page 43. Schloß Dagstuhl–Leibniz-Zentrum für Informatik, 2017.

[17] Brian Mac Namee, Padraig Cunningham, Stephen Byrne, and Owen 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.

[18] Michael Mitzenmacher and Eli Upfal. *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis*. Cambridge university press, second edition, 2017.

[19] Hervé Moulin. *Fair division and collective welfare*. MIT Press, 2004.

[20] John Ashworth Nelder and Robert WM Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972.

[21] Arthur Cecil Pigou. *Wealth and welfare*. Macmillan and Company, limited, 1912.

[22] Kevin WS Roberts. Interpersonal comparability and social choice theory. *The Review of Economic Studies*, pages 421–439, 1980.

[23] Esther Rolf, Max Simchowitz, Sarah Dean, Lydia T Liu, Daniel Björkegren, Moritz Hardt, and Joshua Blumenstock. Balancing competing objectives with noisy data: Score-based classifiers for welfare-aware machine learning. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13–18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 8158–8168. PMLR, 2020.

[24] Guy N Rothblum and Gal Yona. Multi-group agnostic PAC learnability. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18–24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 9107–9115. PMLR, 2021.

[25] Amartya Sen. On weights and measures: Informational constraints in social welfare analysis. *Econometrica: Journal of the Econometric Society*, pages 1539–1572, 1977.

[26] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

[27] Naum Zuselevich Shor. *Minimization methods for non-differentiable functions*, volume 3. Springer Science & Business Media, 2012.

[28] Umer Siddique, Paul Weng, and Matthieu 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.

[29] Till Speicher, Hoda Heidari, Nina Grgić-Hlača, Krishna P Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal 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.

[30] Philip S Thomas, Bruno Castro da Silva, Andrew G Barto, Stephen Giguere, Yuriy Brun, and Emma Brunskill. Preventing undesirable behavior of intelligent machines. *Science*, 366(6468): 999–1004, 2019.

[31] Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[32] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

[33] Vladimir Vapnik and Aleksei Chervonenkis. The uniform convergence of frequencies of the appearance of events to their probabilities. In *Doklady Akademii Nauk*, volume 181, pages 781–783. Russian Academy of Sciences, 1968.

[34] Vladimir Vapnik and Aleksei Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–281, 1971.

A Experimental Setup and Extensions

A.1 Data, Preprocessing, and Experimental Setup

All experiments are conducted on the `adult` dataset, derived from the 1994 US Census database, and obtained from the UCI repository [11], where it was donated by Ronny Kohavi and Barry Becker. This dataset has $m = 48842$ instances, and we used a 90%:10% training:test split. The task has binary target variable `income`, 6 numeric features, and 8 categorical features, including `race` split into 5 ethnoracial groups, namely $\{\texttt{White}, \texttt{Black}, \texttt{Asian-Pac-Islander}, \texttt{Amer-Indian-Eskimo}\,[\textit{sic}], \texttt{Other}\}$, and `sex` split into $\{\texttt{male}, \texttt{female}\}$. In each experiment, the target and protected group are omitted from the feature set, the remaining categorical features are 1-hot encoded, and all d features are z-score normalized.

All experiments are with λ-ℓ_2-norm constrained linear predictors, i.e., the hypothesis class is

$$\mathcal{H} \doteq \left\{ h(\vec{x};\vec{\theta}) = \vec{x} \cdot \vec{\theta} \,\middle|\, \vec{\theta} \in \mathbb{R}^d, \, \|\vec{\theta}\|_2 \leq \lambda \right\} \enspace .$$

The output of this hypothesis class is real-valued, but for this binary classification task, we take $\mathcal{Y} = \pm 1$, so the loss function is selected to reify this value with a semantic classification interpretation. The 0-1 loss (for hard classification) is defined as

$$\ell_{01}(y, h(\vec{x};\vec{\theta})) = 1 - y \operatorname{sgn}(\vec{x} \cdot \vec{\theta}) \enspace ,$$

which is readily interpreted in a decision-theoretic sense, but is generally computationally intractable to optimize. The SVM objective is generally stated in terms of the *hinge loss*, which acts as a convex relaxation of the 0-1 loss. The hinge-loss is defined as

$$\ell_{\mathrm{hinge}}(y, h(\vec{x};\vec{\theta})) = \max(0, 1 - y(\vec{x} \cdot \vec{\theta})) \enspace ,$$

which is of course *convex* in $\vec{\theta}$, and obeys $\ell_{01}(y, h(\vec{x};\vec{\theta})) \leq \ell_{\mathrm{hinge}}(y, h(\vec{x};\vec{\theta}))$. Finally, the logistic-regression cross-entropy loss (measured in nats) is (see, e.g., ch. 9.3 of Shalev-Shwartz and Ben-David [26])

$$\ell_{\mathrm{LRCE}}(y, h(\vec{x};\vec{\theta})) = \ln\!\big(1 + \exp\!\big(-y(\vec{x} \cdot \vec{\theta})\big)\big) \enspace ,$$

which interprets the model output as a *probabilistic classification* $\mathbb{P}(y = 1|\hat{y}) = \frac{1}{1+\exp(-\hat{y})}$. Note that for $\hat{y} \not\approx 0$, $\ell_{\mathrm{LRCE}}(y, \hat{y}) \approx \ell_{\mathrm{hinge}}(y, \hat{y})$, and logistic regression may also be viewed as a *convex relaxation* of hard classification, as $\ell_{01}(y, \hat{y}) \leq \frac{1}{\ln(2)} \ell_{\mathrm{LRCE}}$ (perhaps more naturally, the $\frac{1}{\ln(2)}$ constant vanishes if we measure cross entropy in *bits* rather than *nats*).

In all experiments with *weighted risk values*, we use regularity constraint $\lambda = 4$, and in the experiments with *unweighted risk values*, we take $\lambda = 10$.

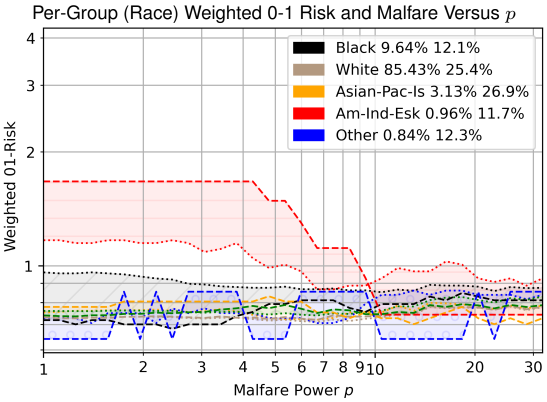
Implementation and Computational Resources Computation was not a concern on these simple convex linear models; all experiments were run on a low-end laptop with no GPU acceleration.

Theorem 3.11 analytically quantifies the computational complexity of ε-EMM, but in our experiments, we simply used standard out-of-the-box first-order methods (adaptive projected gradient descent and SLSQP), as well as derivative-free methods (COBYLA) to train all models.

A.2 Supplementary Experiments

0-1 Risk of Weighted SVM Figure A1 complements figure 1, reporting the same per-group and malfare statistics, except now on the (similarly weighted) 0-1 risk, rather than the weighted hinge risk. Here, the interpretation is that the hinge risk is a *convex proxy* for the 0-1 risk, as it would be computationally intractable to optimize the 0-1 risk directly. Because we optimize hinge risk, but report 0-1 risk, we don't expect to see monotonicity in malfare, and the discontinuity of the 0-1 risk is manifest as noise in risk values. Nevertheless, if hinge risk is a good proxy for 0-1 risk, we should still see a *general trend* of the classifier becoming fairer (improving high-risk group performance) w.r.t. 0-1 risk as it becomes fairer w.r.t. hinge risk, and we do in fact observe this with increasing p.

Figure A1: Training (dashed) and test (dotted) weighted 0-1 risk, per group, of hinge-risk SVM in the `adult` experiment, as functions of malfare power p. Training and test malfare are also plotted, again as functions of p. The model is optimized for weighted malfare of weighted hinge-risk, and is thus identical to the model reported on in figure 1. Here hinge-risk is optimized as a *proxy* for the 0-1 risk, so only the reported risk function changes in this figure.



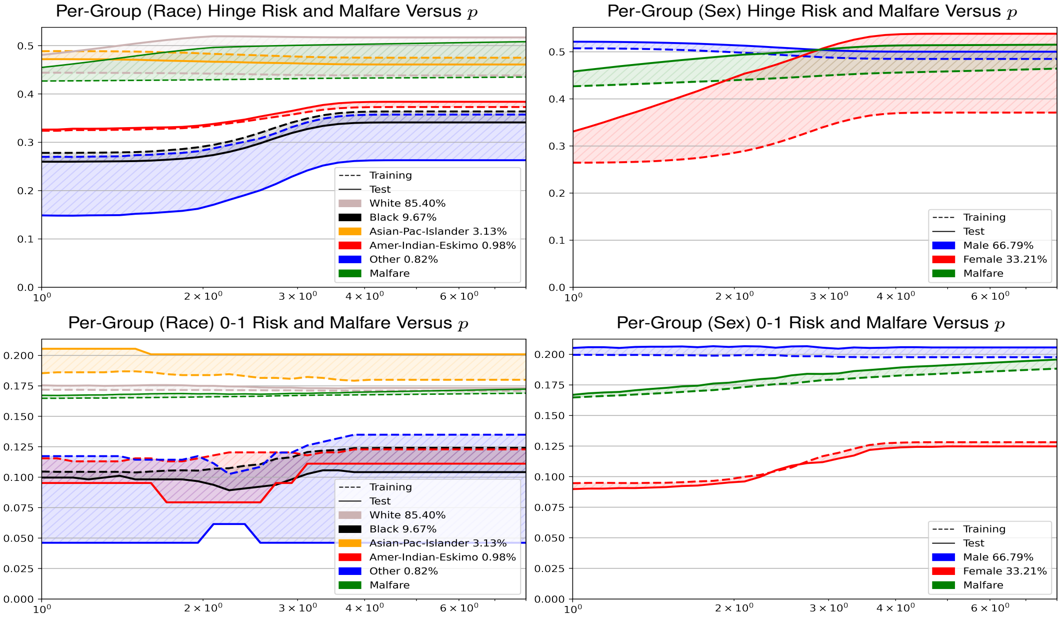


Figure A2: Unweighted linear SVM experiments on the `adult` dataset, with groups split by *race* (left) and *sex* (right), malfare and risk plotted against p. The upper row depicts hinge-risks and malfare of hinge-risks, and the lower row depicts the 0-1 risks and malfare of 0-1 risks (of the models trained on hinge-risk). All plots show training (dashed) and test (solid) per-group risk and malfare values, as functions of p, with shaded regions depicting train-test gaps.

Unweighted SVM These experiments are quite similar to those of figure 1 and figure A1, except here we optimize the malfare of, and report the values of, the *unweighted* hinge risk. In these experiments, we also take regularity constraint $\|\vec{\theta}\|_2 \leq \lambda = 10$, and report the hinge and 0-1 risks and malfares, using *race* and *sex* groups. As such, the objective is to minimize the $\mathrm{M}_p(\cdot;\boldsymbol{w})$ malfare of per-group hinge risks, using per-group-frequencies as malfare weights, i.e.,

$$\hat{h} \doteq \operatorname*{argmin}_{h\in\mathcal{H}} \mathrm{M}_p\left(i \mapsto \hat{\mathrm{R}}(h;\ell_{\mathrm{hinge}},\boldsymbol{z}_i);\boldsymbol{w}\right) \enspace .$$

With both *race* and *sex*, we see significant variation in model performance between groups. We stress that group size and affluence are not directly correlated with model accuracy; for instance, here we see that model performance on the (generally affluent) *Male*, *white*, and *Asian* populations is relatively poor, due to greater income homogeneity within these groups (in direct contrast to the *weighted* experiments).

In all cases, we see that increasing p improves the *training set performance* of the model on the high-risk (inaccurate) groups (male, white, and Asian), at the cost of significant performance degradation

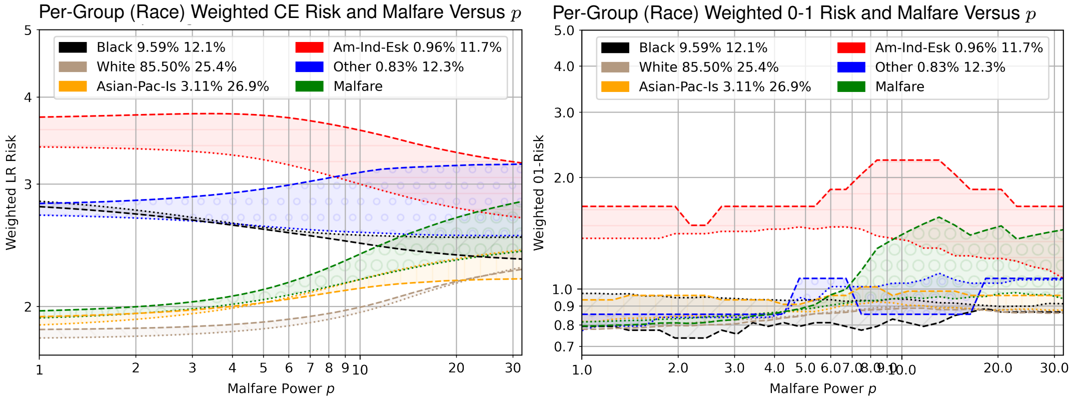


Figure A3: Experiments on the `adult` dataset, with ethnoracial protected groups, and the *weighted logistic regression* malfare objective. Training (dashed) and test (dotted) malfare and risk values are plotted as functions of p, with both weighted cross entropy risk (left) and weighted 0-1 risk (right).

for the more accurate groups. However, the trend does not always hold in *test set* performance, since raising p increases the relative importance of *high-risk subpopulations* in training, which leads to increased overfitting. This highlights the phenomenon of *overfitting to fairness*, as we see that improved training set malfare does not necessarily translate to the test set.

Logistic Regression Experiments Figure A3 complements the previous experiments, where now we optimize malfare of (weighted) *cross entropy risk* of logit predictors, where weights are chosen as in figure 1, i.e., we optimize

$$\hat{h} \doteq \operatorname*{argmin}_{h\in\mathcal{H}} \mathrm{M}_p\left(i \mapsto \frac{1}{\boldsymbol{b}_i}\hat{\mathrm{R}}(h;\ell_{\mathrm{LRCE}},\boldsymbol{z}_i);\boldsymbol{w}\right) \enspace .$$

We draw essentially the same conclusions as with the hinge risk: malfare minimization yields better training performance for high-risk (Black, native American, and other) groups, and better test performance as well, except in the *other* group, which is very small, and badly overfit.

B A Compendium of Missing Proofs

Here we present all missing proofs of results stated in the main text.

B.1 Welfare and Malfare

We now show theorem 2.4.

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 in $\mathcal{S}$ for $p \geq 1$.

Proof. We omit proof of item 1, as this is a standard property of power-means, generally termed the *power-mean inequality* [4, chapter 3].

We first show item 2. By the triangle inequality (for $p \geq 1$), we have

$$\mathrm{M}_p(\mathcal{S} + \mathcal{S}'; \boldsymbol{w}) \leq \mathrm{M}_p(\mathcal{S}; \boldsymbol{w}) + \mathrm{M}_p(\mathcal{S}'; \boldsymbol{w}) \enspace .$$

We now show item 3 First take $\boldsymbol{\varepsilon} \doteq \mathcal{S} - \mathcal{S}'$, and let $\boldsymbol{\varepsilon}_+ \doteq \mathbf{0} \vee \boldsymbol{\varepsilon}$, where $\boldsymbol{a} \vee \boldsymbol{b}$ denotes the (elementwise) maximum. Now consider

$$\begin{aligned}
\mathrm{M}_p(\mathcal{S}; \boldsymbol{w}) &= \mathrm{M}_p(\mathcal{S}' + \boldsymbol{\varepsilon}; \boldsymbol{w}) && \textsc{Definition of } \boldsymbol{\varepsilon} \\
&\leq \mathrm{M}_p(\mathcal{S}' + \boldsymbol{\varepsilon}_+; \boldsymbol{w}) && \textsc{Monotonicity} \\
&\leq \mathrm{M}_p(\mathcal{S}'; \boldsymbol{w}) + \mathrm{M}_p(\boldsymbol{\varepsilon}_+; \boldsymbol{w}) && \textsc{Item 2} \\
&\leq \mathrm{M}_p(\mathcal{S}'; \boldsymbol{w}) + \mathrm{M}_p(\left|\mathcal{S} - \mathcal{S}'\right|; \boldsymbol{w}) \enspace , && \textsc{Monotonicity}
\end{aligned}$$

where here MONOTONICITY refers to monotonicity of $\mathrm{M}_p(\mathcal{S}; \boldsymbol{w})$ in each $\mathcal{S}_i$, i.e., axiom 1. By symmetry, we have $\mathrm{M}_p(\mathcal{S}'; \boldsymbol{w}) \leq \mathrm{M}_p(\mathcal{S}; \boldsymbol{w}) + \mathrm{M}_p(\left|\mathcal{S} - \mathcal{S}'\right|; \boldsymbol{w})$, which implies the result.

We now show item 4. First note the special cases of $p \in \pm\infty$ follow by convexity of the maximum ($p = \infty$) and concavity of the minimum ($p = -\infty$).

Now, note that for $p \geq 1$, by concavity of $\sqrt[p]{\cdot}$, Jensen's inequality gives us

$$\mathrm{M}_1(\mathcal{S}; \boldsymbol{w}) = \sum_{i=1}^{g} \boldsymbol{w}_i \mathcal{S}_i = \underbrace{\sum_{i=1}^{g} \boldsymbol{w}_i \sqrt[p]{\mathcal{S}_i^p} \leq \sqrt[p]{\sum_{i=1}^{g} \boldsymbol{w}_i \mathcal{S}_i^p}}_{\textsc{Definition of Convexity}} = \mathrm{M}_p(\mathcal{S}; \boldsymbol{w}) \enspace ,$$

i.e., convexity, and similarly, for $p \leq 1$, $p \neq 0$, by convexity of $\sqrt[p]{\cdot}$, we have

$$\mathrm{M}_1(\mathcal{S}; \boldsymbol{w}) = \sum_{i=1}^{g} \boldsymbol{w}_i \mathcal{S}_i = \underbrace{\sum_{i=1}^{g} \boldsymbol{w}_i \sqrt[p]{\mathcal{S}_i^p} \geq \sqrt[p]{\sum_{i=1}^{g} \boldsymbol{w}_i \mathcal{S}_i^p}}_{\textsc{Definition of Concavity}} = \mathrm{M}_p(\mathcal{S}; \boldsymbol{w}) \enspace .$$

Similar reasoning, now by convexity of $\exp(\cdot)$, shows the case of $p = 0$. □

We now show theorem 2.5.

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 definition 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. *Debreu-Gorman Factorization*: Axioms 1-5 imply $\exists p \in \mathbb{R}$, strictly-monotonically-increasing continuous $F : \mathbb{R} \to \mathbb{R}_{0+}$ s.t.

$$\mathrm{M}(\mathcal{S};\boldsymbol{w}) = F\left(\sum_{i=1}^{g} \boldsymbol{w}_i f_p(\mathcal{S}_i)\right) \quad , \quad \text{with} \begin{cases} p = 0 & f_0(x) \doteq \ln(x) \\ p \neq 0 & f_p(x) \doteq \operatorname{sgn}(p) x^p \end{cases} \quad .$$

3. *Power-Mean Factorization*: 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. *Fair Welfare*: Axioms 1-5 and 8 imply $p \in (-\infty, 1]$.
5. *Fair Malfare*: Axioms 1-5 and 9 imply $p \in [1, \infty)$.

Proof. First note that item 1 is an immediate consequence of axioms 6–7 (multiplicative linearity and unit scale).

We now note that item 2 is the celebrated Debreu-Gorman theorem, extended to weighted aggregator functions. The proof strategy for this extension is essentially to define two infinite sequences of rational weightings $\boldsymbol{w}^{j\downarrow}$ and $\boldsymbol{w}^{j\uparrow}$ that both converge to $\boldsymbol{w}$, each of which may be characterized by the *unweighted* Debreu-Gorman theorem, to then show that $\mathrm{M}(\mathcal{S};\boldsymbol{w}^{j\downarrow}) \leq \mathrm{M}(\mathcal{S};\boldsymbol{w}) \leq \mathrm{M}(\mathcal{S};\boldsymbol{w}^{j\uparrow})$ for each sequence index j, and finally to apply standard continuity and limit properties to conclude the desideratum. In particular, the lower and upper weighting sequences $\boldsymbol{w}_i^{\downarrow}$ and $\boldsymbol{w}^{\uparrow}$ are composed entirely of weights that are *binary fractions*, the weighted symmetry axiom allows us to equate aggregate values with *rational weights* with aggregate values of a *larger unweighted* population, which is then characterized via the standard (unweighted) Debreu-Gorman theorem.

The (classical) Debreu-Gorman theorem describes the *unweighted case*, and is based on the *unweighted symmetry axiom*, whereas we seek to show the *weighted case*, i.e.,

$$\underbrace{\mathrm{M}(\mathcal{S};\boldsymbol{w}) = F\left(\sum_{i=1}^{g} f_p(\mathcal{S}_i)\right)}_{\text{UNWEIGHTED}} , \underbrace{\mathrm{M}(\mathcal{S};\boldsymbol{w}) = F\left(\sum_{i=1}^{g} \boldsymbol{w}_i f_p(\mathcal{S}_i)\right)}_{\text{WEIGHTED}} , \text{ with} \begin{cases} p = 0 & f_0(x) \doteq \ln(x) \\ p \neq 0 & f_p(x) \doteq \operatorname{sgn}(p) x^p \end{cases} \quad .$$

Note that the unweighted case is equivalent to taking $\boldsymbol{w} = \langle \frac{1}{g}, \frac{1}{g}, \dots, \frac{1}{g} \rangle$, but we omit these weight terms above, as they can be factored into the strictly-monotonically-increasing continuous function $F(\cdot)$. The unweighted case holds by the standard Debreu-Gorman theorem, as our weighted symmetry axiom generalizes the standard unweighted symmetry axiom, i.e., $\forall$ permutations π over $\{1, \dots, g\}$: $\mathrm{M}(\mathcal{S}) = \mathrm{M}(\pi(\mathcal{S}))$. To show the general case, a more sophisticated argument is necessary involving the *continuity* (axiom 3) of $\mathrm{M}(\mathcal{S};\boldsymbol{w})$ w.r.t. $\boldsymbol{w}$ and the *weighted symmetry* axiom (2).

We now begin the reweighting argument proper. Suppose WLOG that $(\mathcal{S};\boldsymbol{w})$ are jointly permuted s.t. $\mathcal{S}_1 \leq \mathcal{S}_2 \leq \cdots \leq \mathcal{S}_g$; this is always possible via the weighted symmetry axiom (2), e.g., for some permutation π, $\mathbb{M}(\langle 3, 2, 1\rangle, \langle \frac{1}{2}, \frac{1}{6}, \frac{1}{3} \rangle) = \mathbb{M}(\pi(\langle 3, 2, 1\rangle), \pi(\langle \frac{1}{2}, \frac{1}{6}, \frac{1}{3} \rangle)) = \mathbb{M}(\langle 1, 2, 3\rangle, \langle \frac{1}{3}, \frac{1}{6}, \frac{1}{2} \rangle)$. Now, define lower and upper weighting sequences

$$\boldsymbol{w}_i^{j\downarrow} \doteq \min\left(1 - \sum_{i'=1}^{i-1} \boldsymbol{w}_{i'}^{j\downarrow}, \frac{\lceil \boldsymbol{w}_i 2^j \rceil}{2^j}\right) \quad , \quad \text{and } \boldsymbol{w}_i^{j\uparrow} \doteq \min\left(1 - \sum_{i'=i+1}^{g} \boldsymbol{w}_{i'}^{j\uparrow}, \frac{\lceil \boldsymbol{w}_i 2^j \rceil}{2^j}\right) \quad ,$$

respectively, as illustrated in figure A4, and note that their limits obey

$$\lim_{j \to \infty} \boldsymbol{w}^{j\downarrow} = \boldsymbol{w} = \lim_{j \to \infty} \boldsymbol{w}^{j\uparrow} \quad ,$$

and furthermore, by continuity of $\mathrm{M}(\mathcal{S};\boldsymbol{w})$ in $\boldsymbol{w}$,

$$\lim_{j \to \infty} \mathrm{M}(\mathcal{S};\boldsymbol{w}^{j\downarrow}) = \mathrm{M}(\mathcal{S};\boldsymbol{w}) = \lim_{j \to \infty} \mathrm{M}(\mathcal{S};\boldsymbol{w}^{j\uparrow}) \quad .$$

Now, note that *cumulative weights* obey the simple relationship for all $i \in \{1, \dots, g\}$:

$$\sum_{i'=1}^{i} \boldsymbol{w}_{i'}^{0\downarrow} \geq \sum_{i'=1}^{i} \boldsymbol{w}_{i'}^{1\downarrow} \geq \cdots \geq \sum_{i'=1}^{i} \boldsymbol{w}_{i'}^{\infty\downarrow} = \sum_{i'=1}^{i} \boldsymbol{w}_{i'} = \sum_{i'=1}^{i} \boldsymbol{w}_{i'}^{\infty\uparrow} \geq \cdots \geq \sum_{i'=1}^{i} \boldsymbol{w}_{i'}^{1\uparrow} \geq \sum_{i'=1}^{i} \boldsymbol{w}_{i'}^{0\uparrow} \quad ,$$

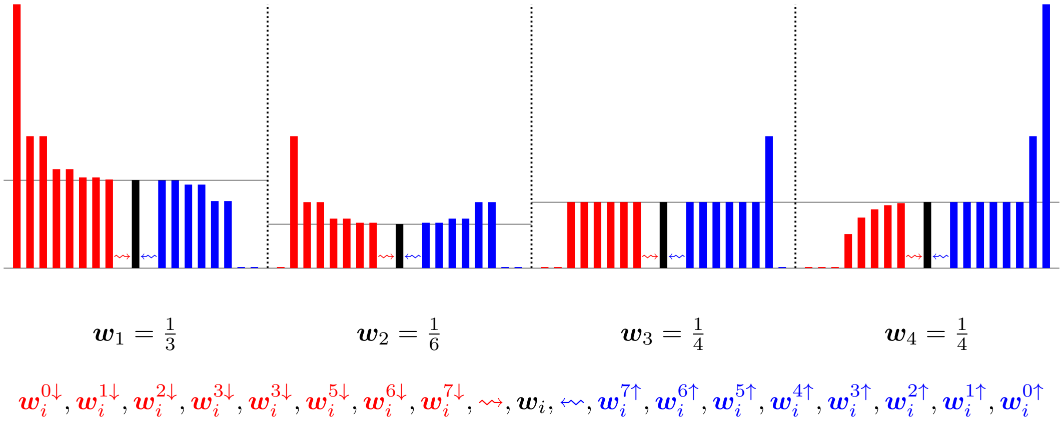


Figure A4: Illustration of lower and upper reweighting sequences $\boldsymbol{w}^{j\downarrow}$ and $\boldsymbol{w}^{j\uparrow}$, for weights vector $\boldsymbol{w} = \langle \frac{1}{2}, \frac{1}{6}, \frac{1}{4}, \frac{1}{4} \rangle$. Each $\boldsymbol{w}_i$ is shown in black, separated by vertical dotted lines, alongside the lower and upper reweighting sequences $\boldsymbol{w}_i^{j\downarrow}$ and $\boldsymbol{w}_i^{j\uparrow}$, for various values of j. Note that with increasing j, the binary fractional weights of the reweighting sequences more accurately approximate $\boldsymbol{w}$, and even by $j = 7$, the difference between $\boldsymbol{w}_i^{j\downarrow} \approx \boldsymbol{w} \approx \boldsymbol{w}^{j\uparrow}$ is essentially imperceptible above.

which is equivalent to *stochastic dominance* when each weights vector is viewed as a discrete random variable over $\{1, \dots, g\}$. Intuitively, this is because the lower and upper weights sequences round weights progressively more accurately, in such a way that the lower sequence $\boldsymbol{w}_i^{j\downarrow}$ always transfers weight to higher group indices i as sequence index j increases, whereas the upper sequence $\boldsymbol{w}_i^{j\uparrow}$ always transfers weight to lower group indices i as sequence index j increases.

Now, by the weighted symmetry axiom (2), letting $\boldsymbol{w}^{j\uparrow\downarrow}$ refer generically to either $\boldsymbol{w}^{j\uparrow}$ or $\boldsymbol{w}^{j\downarrow}$,

$$\forall j: \ \mathrm{M}(\mathcal{S}; \boldsymbol{w}^{j\uparrow\downarrow}) = \mathrm{M}\left(\Big\langle \underbrace{\mathcal{S}_1, \dots, \mathcal{S}_1}_{2^j \boldsymbol{w}_1^{j\uparrow\downarrow} \text{ Copies}}, \underbrace{\mathcal{S}_2, \dots, \mathcal{S}_2}_{2^j \boldsymbol{w}_2^{j\uparrow\downarrow} \text{ Copies}}, \dots, \underbrace{\mathcal{S}_g, \dots, \mathcal{S}_g}_{2^j \boldsymbol{w}_g^{j\uparrow\downarrow} \text{ Copies}} \Big\rangle, \underbrace{i \mapsto \frac{1}{2^j}}_{\text{Uniform Weights}} \right) .$$

From here, by the monotonicity axiom (1), as well as the above argument regarding cumulative weights, and the fact that $\mathcal{S}$ was assumed to be ascending (and remains so above), we have

$$\mathrm{M}(\mathcal{S}; \boldsymbol{w}^{0\downarrow}) \leq \mathrm{M}(\mathcal{S}; \boldsymbol{w}^{1\downarrow}) \leq \dots \leq \mathrm{M}(\mathcal{S}; \boldsymbol{w}^{\infty\downarrow}) = \mathrm{M}(\mathcal{S}; \boldsymbol{w}) = \mathrm{M}(\mathcal{S}; \boldsymbol{w}^{\infty\uparrow}) \leq \dots \leq \mathrm{M}(\mathcal{S}; \boldsymbol{w}^{1\uparrow}) \leq \mathrm{M}(\mathcal{S}; \boldsymbol{w}^{0\uparrow}) .$$

In particular, monotonicity and weighted symmetry imply the inequalities relating the lower and upper weighting sequences, and continuity in $\boldsymbol{w}$ implies the equalities between the malfare values of (limits of) the weighting sequences and $\boldsymbol{w}$.

Now, by the (unweighted) Debreu-Gorman theorem, we have that there exists some $p \in \mathbb{R}$, and some strictly-monotonically-increasing continuous function $F(\cdot)$, s.t. for all j,

$$\mathrm{M}(\mathcal{S}^{j\uparrow\downarrow}; \boldsymbol{w}) = F\big(\mathrm{M}_p(\mathcal{S}^{j\uparrow\downarrow}; \boldsymbol{w})\big) .$$

Substituting this result into the above yields the desideratum.

We now show item 3. This result is essentially a corollary of item 2, hence the dependence on axioms 1–5. In particular, we need only derive the canonical form of $F(\cdot)$ under this axiomatization, which is quite straightforward when one considers the case of $\mathcal{S} = \mathbf{1}$. By item 1 (as we now assume axioms 6–7), and axiom 6 itself, for all $p \neq 0$, we have

$$\alpha = \alpha \mathrm{M}(\mathbf{1}; \boldsymbol{w}) = \mathrm{M}(\alpha \mathbf{1}; \boldsymbol{w}) = F\left(\sum_{i=1}^{g} \boldsymbol{w}_i f_p(\alpha) \right) = F\big(\mathrm{sgn}(p)\alpha^p\big) .$$

From here, we have $\alpha = F\big(\mathrm{sgn}(p)\alpha^p\big)$, thus $F^{-1}(u) = \mathrm{sgn}(p)u^p$, and consequently, $F(v) = \sqrt[p]{\mathrm{sgn}(p)v}$.

Taking $p = 0$, and applying similar logic to the above, gets us

$$\alpha = \alpha\mathrm{M}(\mathbf{1}; \boldsymbol{w}) = \alpha\mathrm{M}(\alpha\mathbf{1}; \boldsymbol{w}) = F\left(\sum_{i=1}^{g} \boldsymbol{w}_i \ln(\alpha)\right) = F(\ln\alpha) \enspace ,$$

from which it is clear that $F^{-1}(u) = \ln(u) \implies F(v) = \exp(v)$.

For all values of $p \in \mathbb{R}$, substituting the values of f_p and $F(\cdot)$ into item 2 yields $\mathrm{M}(\mathcal{S}; \boldsymbol{w}) = \mathrm{M}_p(\mathcal{S}; \boldsymbol{w})$ by definition.

We now show items 4 and 5. These properties follow directly from item item 3 and item 4, alongside the observation that the Pigou-Dalton transfer principle (axiom 8) implies $\mathrm{M}(\mathcal{S}; \boldsymbol{w})$ is *concave* in $\mathcal{S}$, whereas the negated Pigou-Dalton transfer principle (axiom 9) implies $\mathrm{M}(\mathcal{S}; \boldsymbol{w})$ is *convex* in $\mathcal{S}$. □

B.2 FPAC-Learning

We now work towards showing theorem 3.2. We first show some lemmas in service of the main result.

Lemma B.1 (Statistical Estimation)**.** Suppose probability distributions $\mathcal{D}_{1:g}$, samples $\boldsymbol{z}_{1:g,1:m} \sim \mathcal{D}_1^m \times \cdots \times \mathcal{D}_g^m$, aggregator function $\mathrm{M}(\cdot; \cdot)$ obeying *monotonicity*, probability vector $\boldsymbol{w} \in \mathbb{R}_+^g$, and some sentiment-function f. Now take *sentiment vector* $\mathcal{S} \in \mathbb{R}_{0+}^g$ such that $\mathcal{S}_i = \mathbb{E}_{\mathcal{D}_i}[f]$, and *empirical sentiment value estimate* $\hat{\mathcal{S}} \in \mathbb{R}_{0+}^g$ such that $\hat{\mathcal{S}}_i \doteq \hat{\mathbb{E}}_{\boldsymbol{z}_i}[f]$. If it holds with probability at least $1 - \delta$ that $\forall i \in 1, \ldots, g : \hat{\mathcal{S}}_i - \boldsymbol{\varepsilon}_i \leq \mathcal{S}_i \leq \hat{\mathcal{S}}_i + \boldsymbol{\varepsilon}_i$, then with said probability, we have

$$\mathrm{M}_p(\mathbf{0} \vee (\hat{\mathcal{S}} - \boldsymbol{\varepsilon}); \boldsymbol{w}) \leq \mathrm{M}_p(\mathcal{S}; \boldsymbol{w}) \leq \mathrm{M}_p(\hat{\mathcal{S}} + \boldsymbol{\varepsilon}; \boldsymbol{w}) \enspace ,$$

where $\boldsymbol{a} \vee \boldsymbol{b}$ denotes the (elementwise) maximum.

Proof. This result follows from the assumption, and the *monotonicity* axiom (i.e., adding/subtracting $\boldsymbol{\varepsilon}$ can not decrease/increase the aggregator function value, respectively). The minimum with 0 on the LHS is *valid* simply because, by definition, sentiment values are nonnegative, and is *necessary*, as $\mathrm{M}_p(\cdot; \boldsymbol{w})$ is in general undefined with negative inputs. □

We now state a standard result[4] in uniform convergence theory (see, e.g., [18, 26]).

Lemma B.2 (Emprical Rademacher Bounds)**.** Suppose $\ell : (\mathcal{Y} \times \mathcal{Y}) \to [0, r]$, and $\mathcal{H} \subseteq \mathcal{X} \to \boldsymbol{y}$. Suppose also $\boldsymbol{z} \sim \mathcal{D}^m$. Then, with probability at least $1 - \delta$ over choice of $\boldsymbol{z}$, we have

$$\sup_{h \in \mathcal{H}} \left| \hat{\mathrm{R}}(h; \ell, \boldsymbol{z}) - \mathrm{R}(h; \ell, \mathcal{D}) \right| \leq 2\hat{\mathfrak{R}}_m(\ell \circ \mathcal{H}, \boldsymbol{z}) + 3r\sqrt{\tfrac{\ln\frac{1}{\delta}}{2m}} \enspace .$$

With these results, we are now ready to show theorem 3.2.

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 malfare functions* $\rotatebox{180}{W}(\cdot; \cdot)$ (i.e., $\rotatebox{180}{W}_p(\cdot; \cdot)$ for $p \geq 1$) and *probability vectors* $\boldsymbol{w} \in \mathbb{R}_+^g$, that

$$\sup_{h \in \mathcal{H}} \left| \rotatebox{180}{W}\big(i \mapsto \mathrm{R}(h; \ell, \mathcal{D}_i); \boldsymbol{w}\big) - \rotatebox{180}{W}\big(i \mapsto \hat{\mathrm{R}}(h; \ell, \boldsymbol{z}_i); \boldsymbol{w}\big) \right| \leq \rotatebox{180}{W}\left(i \mapsto 2\hat{\mathfrak{R}}_m(\ell \circ \mathcal{H}, \boldsymbol{z}_i) + 3r\sqrt{\tfrac{\ln\frac{g}{\delta}}{2m}}; \boldsymbol{w}\right) \enspace .$$

Proof. This result follows from the above statistical lemmas, and the contraction property of malfare. In particular, note that by lemma B.2, with probability at least $1 - \delta$, by union bound it holds that

$$\forall i \in 1, \ldots, g : \forall h \in \mathcal{H} : \left| \hat{\mathrm{R}}(h; \ell, \boldsymbol{z}_i) - \mathrm{R}(h; \ell, \mathcal{D}) \right| \leq 2\hat{\mathfrak{R}}_m(\ell \circ \mathcal{H}, \boldsymbol{z}_i) + 3r\sqrt{\frac{\ln\frac{g}{\delta}}{2m}} \enspace .$$

The desideratum then follows via lemma B.1 and the *contraction property* of malfare, i.e., theorem 2.4 item 3. Note that the above statement on risk holds *in probability*, but the rest of the argument holds *deterministically*, hence the *with probability at least* $1 - \delta$ qualifier holds over all fair malfare functions and weighting probability vectors *simultaneously*. □

[4] Note cautiously that we bound the *absolute supremum deviation*, which incurs no union bound penalty, as we define the *empirical Rademacher average* itself with the absolute value built in.

We now show theorem 3.7.

Theorem 3.7 (Realizable Reductions)**.** Suppose loss function ℓ and hypothesis class $\mathcal{H}$. Then
1. $(\mathcal{H}, \ell) \in \mathrm{PAC}^0 \implies (\mathcal{H}, \ell) \in \mathrm{FPAC}^0$; and 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 $\mathcal{A}(\mathcal{D}_{1:g}, \boldsymbol{w}, \mathrm{M}, \varepsilon, \delta, d) \doteq \mathcal{A}'(\mathrm{mix}(\mathcal{D}_{1:g}), \frac{\varepsilon}{g}, \delta, d)$, where $\mathrm{mix}(\mathcal{D}_{1:g})$ denotes the uniformly-weighted mixture of distributions $\mathcal{D}_{1:g}$. 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 $\mathrm{t}_{\mathcal{A}}(\varepsilon, \delta, d, g) = \mathrm{t}_{\mathcal{A}'}(\frac{\varepsilon}{g}, \delta, d)$.

Proof. We first show the *correctness* of the constructed FPAC-learner $\mathcal{A}$. Suppose $\hat{h} \leftarrow \mathcal{A}(\mathcal{D}_{1:g}, \boldsymbol{w}, \mathrm{M}, \varepsilon, \delta, d)$. Then, with probability at least $1 - \delta$, by the PAC-learnability guarantee of $\mathcal{A}'$, we have

$$\begin{aligned} \mathrm{M}_p\big(i \mapsto \mathrm{R}(h; \ell, \mathcal{D}_i); \boldsymbol{w}\big) &\leq \mathrm{M}_\infty\big(i \mapsto \mathrm{R}(h; \ell, \mathcal{D}_i); i \mapsto \tfrac{1}{g}\big) \\ &\leq g\mathrm{M}_1\big(i \mapsto \mathrm{R}(h; \ell, \mathcal{D}_i); i \mapsto \tfrac{1}{g}\big) \\ &= g\mathrm{R}\big(h; \ell, \mathrm{mix}(\mathcal{D}_{1:g})\big) \leq g\tfrac{\varepsilon}{g} = \varepsilon \ . \end{aligned}$$

We thus may conclude that $(\mathcal{H}, \ell)$ is efficiently *realizable-PAC-learnable* by $\mathcal{A}'$, with *sample complexity* $\mathrm{m}_{\mathcal{A}}(\varepsilon, \delta, d, g) = \mathrm{m}_{\mathcal{A}'}(\frac{\varepsilon}{g}, \delta, d)$. Similarly, if $\mathcal{A}$ has polynomial runtime, then so too does $\mathcal{A}'$, thus we may also conclude efficiency. □

We now show theorem 3.10.

Theorem 3.10 (Fundamental Theorem of Fair Statistical Learning)**.** Suppose ℓ 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. $\forall d \in \mathbb{Z}_+$: $(\ell, \mathcal{H}_d)$ 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.

Proof. First note that $1 \implies 2$ is a rather straightforward consequence of the definition of uniform convergence and the contraction property of fair malfare functions (theorem 2.4 item 3). In particular, take $m \doteq \mathrm{m}_{\mathrm{UC}}(\ell \circ \mathcal{H}_d, \frac{\varepsilon}{2}, \frac{\delta}{g})$. By union bound, this implies that with probability at least $1 - \delta$, taking samples $\boldsymbol{z}_{1:g,1:m} \sim \mathcal{D}_1^m \times \cdots \times \mathcal{D}_g^m$, we have

$$\forall i \in \{1, \dots, g\} : \sup_{h \in \mathcal{H}_d} \Big| \mathrm{R}(h; \ell, \mathcal{D}_i) - \hat{\mathrm{R}}(h; \ell, \boldsymbol{z}_i) \Big| \leq \frac{\varepsilon}{2} \ .$$

Consequently, as $\mathrm{M}(\cdot; \boldsymbol{w})$ is 1-$\|\cdot\|_\infty$-$|\cdot|$-Lipschitz in risk (see lemma B.1), it holds with probability at least $1 - \delta$ that

$$\forall h \in \mathcal{H}_d : \Big| \mathrm{M}\big(i \mapsto \hat{\mathrm{R}}(h; \ell, \boldsymbol{z}_i); \boldsymbol{w}\big) - \mathrm{M}\big(i \mapsto \mathrm{R}(h; \ell, \mathcal{D}_i); \boldsymbol{w}\big) \Big| \leq \frac{\varepsilon}{2} \ .$$

Now, for EMM-optimal $\hat{h}$, and malfare-optimal h^*, we apply this result twice to get

$$\begin{aligned} \mathrm{M}\big(i \mapsto \mathrm{R}(\hat{h}; \ell, \mathcal{D}_i); \boldsymbol{w}\big) &\leq \mathrm{M}\big(i \mapsto \hat{\mathrm{R}}(\hat{h}; \ell, \boldsymbol{z}_i); \boldsymbol{w}\big) + \tfrac{\varepsilon}{2} \\ &\leq \mathrm{M}\big(i \mapsto \hat{\mathrm{R}}(h^*; \ell, \boldsymbol{z}_i); \boldsymbol{w}\big) + \tfrac{\varepsilon}{2} \\ &\leq \mathrm{M}\big(i \mapsto \mathrm{R}(h^*; \ell, \mathcal{D}_i); \boldsymbol{w}\big) + \varepsilon \ . \end{aligned}$$

Therefore, under uniform convergence, the EMM algorithm agnostic FPAC learns $(\mathcal{H}, \ell)$ with finite sample complexity $\mathrm{m}_{\mathcal{A}}(\varepsilon, \delta, d, g) = g \cdot \mathrm{m}_{\mathrm{UC}}(\ell \circ \mathcal{H}_d, \frac{\varepsilon}{2}, \frac{\delta}{g})$, completing $1 \implies 2$.

Now, $2 \implies 3$ holds, as realizable learning is a special case of agnostic learning.

It remains only to show that $3 \implies 1$, i.e., if $\mathcal{H}$ is realizable FPAC learnable, then $\mathcal{H}$ has the uniform convergence property. In general, the question is rather subtle, but here the assumption "suppose ℓ such that $(\mathcal{H}, \ell) \in \mathrm{PAC}^0 \implies (\mathcal{H}, \ell) \in \mathrm{UC}$" does most of the work. In particular, as PAC-learning is a special case of FPAC-learning, we have

$$(\mathcal{H}, \ell) \in \mathrm{FPAC} \implies (\mathcal{H}, \ell) \in \mathrm{PAC} \ ,$$

then applying the assumption yields $(\mathcal{H}, \ell) \in \mathrm{UC}$. □

B.3 Efficient FPAC-Learning

We now show theorem 3.11.

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 Θ_d in $\mathrm{Poly}(d)$ time. Suppose also ℓ 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. ℓ is λ_ℓ-$\|\cdot\|_{\mathcal{Y}}$-$|\cdot|$-Lipschitz in $\hat{y}$, i.e.,

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

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

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

Finally, assume $\ell \circ \mathcal{H}_d$ exhibits ε-δ *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 $\mathrm{M}(\cdot;\cdot)$, the algorithm (see algorithm 1)

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

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 $(\mathcal{H}, \ell) \in \mathrm{FPAC}_{\mathrm{Poly}}$.

Proof. We now show that this projected-subgradient method construction of $\mathcal{A}$ requires $\mathrm{Poly}(\frac{1}{\varepsilon}, \frac{1}{\delta}, d, g)$ time to identify an ε-δ-$\mathrm{M}_p(\cdot;\cdot)$-optimal $\tilde{\theta} \in \Theta_d$, and thus FPAC-learns $(\mathcal{H}, \ell)$. This essentially boils down to showing that (1) the empirical malfare objective is *convex* and *Lipschitz continuous*, and (2) that algorithm 1 runs sufficiently many projected-subgradient update steps, with appropriate step size, on a sufficiently large training set, to yield the appropriate guarantees, and that each step of the projected-subgradient method, of which there are polynomially many, itself requires polynomial time.

First, note that by theorem 2.5 items 3 and 5, we may assume that $\mathrm{M}(\cdot;\cdot)$ can be expressed as a p-power-mean with $p \geq 1$; thus henceforth we refer to it as $\mathrm{M}_p(\cdot;\cdot)$. Now, recall that the empirical malfare objective (given $\theta \in \Theta_d$ and training sets $\boldsymbol{z}_{1:g}$) is defined as

$$\mathrm{M}_p\big(i \mapsto \hat{\mathrm{R}}(h(\cdot;\theta); \ell, \boldsymbol{z}_i); \boldsymbol{w}\big) \ .$$

We first show that empirical malfare is convex in $\theta \in \Theta_d$. By assumption and positive linear closure, $\hat{\mathrm{R}}(h(\cdot;\theta'); \ell, \boldsymbol{z}_i)$ is convex in $\theta \in \Theta_d$. The objective of interest is the composition of $\mathrm{M}_p(\cdot;\boldsymbol{w})$ with this empirical risk evaluated on each of the g training sets $\boldsymbol{z}_{1:g}$. By theorem 2.4 item 4, $\mathrm{M}_p(\cdot;\boldsymbol{w})$ is convex $\forall p \in [1,\infty]$ in $\mathbb{R}^g_{0+}$, and by the monotonicity axiom, it is monotonically increasing. Composition of a monotonically increasing convex function on $\mathbb{R}^g_{0+}$ with convex functions on Θ_d yields a convex function, thus we conclude the empirical malfare objective is convex in Θ_d.

We now show that empirical malfare is Lipschitz-continuous. Now, note that for any $p \geq 1$, $\boldsymbol{w}$,

$$\forall \mathcal{S}, \mathcal{S}' : \left|\mathrm{M}_p(\mathcal{S};\boldsymbol{w}) - \mathrm{M}_p(\mathcal{S}';\boldsymbol{w})\right| \leq 1\left\|\mathcal{S} - \mathcal{S}'\right\|_\infty \ ,$$

i.e., $\mathrm{M}_p(\cdot;\boldsymbol{w})$ is 1-$\|\cdot\|_\infty$-$|\cdot|$-Lipschitz in *empirical risks* (see theorem 2.4 item 3), and thus by Lipschitz composition, we have Lipschitz property

$$\forall \theta, \theta' \in \Theta_d : \Big|\mathrm{M}_p\big(i \mapsto \hat{\mathrm{R}}(h(\cdot;\theta); \ell, \boldsymbol{z}_i); \boldsymbol{w}\big) - \mathrm{M}_p\big(i \mapsto \hat{\mathrm{R}}(h(\cdot;\theta'); \ell, \boldsymbol{z}_i); \boldsymbol{w}\big)\Big| \leq \lambda_\ell \lambda_{\mathcal{H}} \left\|\theta - \theta'\right\|_2 \ .$$

We now show that algorithm 1 FPAC-learns $(\mathcal{H}, \ell)$. As above, take $m \doteq \mathrm{m}_{\mathrm{UC}}(\frac{\varepsilon}{3}, \frac{\delta}{g}, d)$. Our algorithm shall operate on a training sample $\boldsymbol{z}_{1:g,1:m} \sim \mathcal{D}_1^m \times \cdots \times \mathcal{D}_g^m$.

First note that evaluating a subgradient (via forward finite-difference estimation or automated subdifferentiation) requires $(\mathrm{Dim}(\Theta_d) + 1)m$ evaluations of $h(\cdot;\cdot)$, which by assumption is possible in $\mathrm{Poly}(d, m) = \mathrm{Poly}(\frac{1}{\varepsilon}, \frac{1}{\delta}, d, g)$ time.

The projected subgradient method produces $\tilde{\theta}$ approximating the empirically-optimal $\hat{\theta}$ such that [see 27]

$$f(\tilde{\theta}) \leq f(\hat{\theta}) + \frac{\|\theta_0 - \hat{\theta}\|_2^2 + \Lambda^2\alpha^2 n}{2\alpha n} \leq \frac{\mathrm{Diam}^2(\Theta_d) + \Lambda^2\alpha^2 n}{2\alpha n} \enspace ,$$

for Λ-$\|\cdot\|_2$-$|\cdot|$-Lipschitz objective f, thus taking $\alpha \doteq \frac{\mathrm{Diam}(\Theta_d)}{\Lambda\sqrt{n}}$ yields

$$f(\tilde{\theta}) - f(\hat{\theta}) \leq \frac{\mathrm{Diam}(\Theta_d)\Lambda}{\sqrt{n}} \enspace .$$

As shown above, $\Lambda = \lambda_\ell \lambda_{\mathcal{H}}$, thus we may guarantee *optimization error*

$$\varepsilon_{\mathrm{opt}} \doteq f(\hat{\theta}) - f(\theta^*) \leq \frac{\varepsilon}{3}$$

if we take iteration count

$$n \geq \frac{9\,\mathrm{Diam}^2(\Theta_d)\lambda_\ell^2\lambda_{\mathcal{H}}^2}{\varepsilon^2} = \left(\frac{3\,\mathrm{Diam}(\Theta_d)\lambda_\ell\lambda_{\mathcal{H}}}{\varepsilon}\right)^2 \in \mathrm{Poly}(\tfrac{1}{\varepsilon}, d) \enspace .$$

As each iteration requires $m \cdot \mathrm{Poly}(d) \subseteq \mathrm{Poly}(\frac{1}{\varepsilon}, \frac{1}{\delta}, d, g)$ time, the projected-subgradient method identifies an $\frac{\varepsilon}{3}$-empirical-malfare-optimal $\tilde{\theta} \in \Theta_d$ in $\mathrm{Poly}(\frac{1}{\varepsilon}, \frac{1}{\delta}, d, g)$ time.

As m was selected to ensure $\frac{\epsilon}{3}$-$\frac{\delta}{g}$ uniform convergence, we thus have that by uniform convergence, and union bound (over g groups), with probability at least $1 - \delta$ over choice of $\boldsymbol{z}_{1:g}$, we have

$$\forall i \in \{1, \dots, g\}, \theta \in \Theta_d : \left| \mathrm{M}_p(i \mapsto \hat{\mathrm{R}}(h(\cdot;\theta); \ell, \boldsymbol{z}_i); \boldsymbol{w}) - \mathrm{M}_p(i \mapsto \mathrm{R}(h(\cdot;\theta); \ell, \mathcal{D}_i); \boldsymbol{w}) \right| \leq \frac{\varepsilon}{3} \enspace .$$

Combining *estimation* and *optimization* errors, we get that with probability at least $1 - \delta$, the approximate-EMM-optimal $h(\cdot;\tilde{\theta})$ obeys

$$\begin{aligned} \mathrm{M}_p(i \mapsto \mathrm{R}(h(\cdot;\tilde{\theta}); \ell, \mathcal{D}_i); \boldsymbol{w}) &\leq \mathrm{M}_p(i \mapsto \hat{\mathrm{R}}(h(\cdot;\tilde{\theta}); \ell, \boldsymbol{z}_i); \boldsymbol{w}) + \tfrac{\varepsilon}{3} \\ &\leq \mathrm{M}_p(i \mapsto \hat{\mathrm{R}}(h(\cdot;\hat{\theta}); \ell, \boldsymbol{z}_i); \boldsymbol{w}) + \tfrac{2\varepsilon}{3} \\ &\leq \mathrm{M}_p(i \mapsto \hat{\mathrm{R}}(h(\cdot;\theta^*); \ell, \boldsymbol{z}_i); \boldsymbol{w}) + \tfrac{2\varepsilon}{3} \\ &\leq \mathrm{M}_p(i \mapsto \mathrm{R}(h(\cdot;\theta^*); \ell, \mathcal{D}_i); \boldsymbol{w}) + \varepsilon \enspace . \end{aligned}$$

We may thus conclude that $\mathcal{A}$ FPAC learns $\mathcal{H}$ with sample complexity $gm = g \cdot \mathrm{m}_{\mathrm{UC}}(\frac{\varepsilon}{3}, \frac{\delta}{g}, d)$. Furthermore, as the entire operation requires polynomial time, we have $(\mathcal{H}, \ell) \in \mathrm{PAC}_{\mathrm{Poly}}$. □