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

Cyrus Cousins

April 30, 2021

Abstract

We address an inherent difficulty in welfare-theoretic fair machine learning, proposing an equivalently-axiomatically justified alternative, and studying the resulting computational and statistical learning questions. Welfare metrics quantify *overall wellbeing* across a population of one or more groups, and welfare-based objectives and constraints have recently been proposed to incentivize *fair machine learning methods* to produce satisfactory solutions that consider the diverse needs of multiple groups. Unfortunately, many machine-learning problems are more naturally cast as *loss minimization*, rather than *utility maximization* tasks, which complicates direct application of welfare-centric methods to fair-ML tasks. In this work, we define a complementary measure, termed *malfare*, measuring overall societal harm (rather than wellbeing), with axiomatic justification via the standard axioms of cardinal welfare. We then cast fair machine learning as a direct *malfare minimization* problem, where a group's malfare is their risk (expected loss). Surprisingly, the axioms of cardinal welfare (malfare) dictate that this is not equivalent to simply defining utility as negative loss. 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 for *any* (axiomatically justified) malfare concept. We show broad conditions under which, with appropriate modifications, many standard PAC-learners may be converted to fair-PAC learners. This places fair-PAC learning on firm theoretical ground, as it yields statistical, and in some cases computational, efficiency guarantees for many well-studied machine-learning models, and is also practically relevant, as it democratizes fair ML by providing concrete training algorithms and rigorous generalization guarantees for these models.

1 Introduction

It is now well-understood that contemporary ML systems for tasks like facial recognition (Buolamwini and Gebru, 2018; Cook et al., 2019; Cavazos et al., 2020), medical settings (Mac Namee et al., 2002; Ashraf et al., 2018), and many others exhibit differential accuracy across gender, race, and other protected-group membership. This immediately yields *accessibility issues* to users of such systems, and can lead to *direct discrimination*, i.e., facial recognition in policing yielding disproportionate false-arrest rates, and ML in medical technologies yielding disproportionately bad health outcomes, exacerbating existing structural and societal inequalities impacting many minority groups. In welfare-centric ML methods, both *accuracy* and *fairness* are encoded in a single *welfare function* defined on a *group of subpopulations*. Welfare is then directly optimized (Rolf et al., 2020) or constrained (Speicher et al., 2018; Heidari et al., 2018) to promote *fair learning* across *all groups*. This addresses differential performance and bias issues across groups by ensuring that (1), each group is *seen* and *considered* during training, and (2), an outcome is incentivized that is desirable overall, ideally according to some mutually-agreed-upon welfare function. Unfortunately, welfare based metrics require a notion of (positive) utility, and we argue that this is not natural to many machine learning tasks, where we instead *minimize* some *negatively connoted* loss value (e.g., in decision-theory or with proper scoring rules). We thus define a complementary measure to *welfare*, termed *malfare*, measuring societal harm (rather than wellbeing). In particular, malfare arises naturally when one applies the standard *axioms of cardinal welfare* (with appropriate modifications) to *loss* rather than *utility*. With this framework, we then cast fair machine learning as a direct *malfare minimization* problem, where a group's malfare is their risk (expected loss).

Perhaps surprisingly, defining and minimizing a *malfare function* is *not equivalent* to defining and maximizing some *welfare function*, while taking utility to be negative loss (except in the trivial cases of egalitarian and utilitarian malfare). This is essentially because nearly every function satisfying the standard axioms of cardinal welfare requires *nonnegative* inputs, and it is not in general possible 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. More strikingly, for problems like minimizing square error (i.e., in regression), it is in general not even possible to define a complementary nonnegative gain function without changing the optimal solution, *even for a single group*.

Building upon these concepts, we develop a notion of generic fair machine learning, termed *fair-PAC learning*, where the goal is to learn models for which finite training samples may guarantee high-probability malfare bounds, for *any* (axiomatically justified) malfare concept. This definition extends Valiant's (Valiant, 1984) classic PAC-learning formalization of machine learning, and we show that, with appropriate modifications, many (standard) PAC-learners may be converted to fair-PAC learners, and argue that fair-PAC-learners are intuitive and easy to use, as one must only select a malfare concept (encoding their desired fairness concept), hypothesis class, and confidence guarantees, and then receives a provably ε-δ optimal model. In particular, we show via a constructive polynomial reduction that *realizable fair PAC-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 fair-PAC-learnability. We also show that when training is possible via *convex optimization* or efficient-enumeration of an approximate cover of the hypothesis space, then ε-δ training malfare-optimal models, like risk-optimal models, requires polynomial time.

This contrasts existing work, as no fairness tolerance parameters, demographic parity constraints, or explicit utility function definitions are required, and this simplicity also leads naturally to straightforward *generalization guarantees* on malfare (thus controlling for overfitting of both accuracy and fairness).

We briefly summarize our contributions as follows.

1. We derive in section 3 the *malfare concept*, extending welfare to measure *negatively-connoted* attributes, and show that *malfare-minimization* naturally generalizes *risk-minimization* to produce *fairness-sensitive* machine-learning objectives that consider multiple protected groups.

2. We show in section 4 that in many cases, while empirical estimates of welfare and malfare are *statistically biased*, they may be *sharply estimated* using standard finite-sample concentration-of-measure bounds.

3. Section 5 extends PAC-learning to Fair-PAC (FPAC) learning, where we consider minimization not only of risk (expected loss) objectives, but of malfare objectives. Both PAC and FPAC learning are parameterized by a *learning task* (model space and loss function), and we explore the rich hierarchy of learnability under variations of these concepts. In particular, we show that
   (a) for many loss functions, PAC and FPAC learning are *statistically equivalent* (i.e., PAC-learnability implies FPAC-learnability) in section 6 ; and
   (b) standard convexity conditions sufficient for PAC-learnability are also sufficient for FPAC-learnability in section 7.

While we explore the basic relationships between various learnability classes, many open questions remain, and we hope future work will further characterize these practically interesting and theoretically deep problems. For brevity, longer, more technical proofs are presented in the appendix.

2 Related Work

Constraint-based notions of algorithmic fairness have risen to prominence in machine learning, with the potential to ensure demographic-parity (e.g., equality of opportunity), thus correcting for some forms of data or algorithmic bias. While noble in intent and intuitive by design, fairness by demographic-parity constraints has several prominent flaws: most notably, several popular parity constraints are mutually unsatisfiable (Kleinberg et al., 2017), 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 (Hu and Chen, 2020; Kasy and Abebe, 2021) have shown that *welfare* and even *disadvantaged group utility* can decrease even as fairness constraints are tightened, calling into question whether demographic parity constraints are even beneficial to those they purport to aid. Furthermore, without a deeper axiomatic structure, it is unclear why one should choose one notion of fairness over another, and it is unsatisfying that we don't have a way of comparing and selecting between them without appealing to informal arguments.

Perhaps in response to these issues, some recent work has trended toward welfare-based fairness-concepts (Hu and Chen, 2020; Rolf et al., 2020), wherein both *accuracy* and *fairness* are encoded in a *welfare function* defined on a *group of subpopulations*. Welfare is then directly optimized (Rolf et al., 2020) or constrained (Speicher et al., 2018; Heidari et al., 2018) to promote *fair learning* across *all groups* Perhaps the most similar to our work is a method of (Hu and Chen, 2020), wherein they *directly maximize* empirical utility over linear (halfspace) classifiers; however as with other previous works, an appropriate utility function must be selected, which we avoid by instead using malfare. We argue that *empirical welfare maximization* is an effective strategy when an appropriate and natural measure of *utility* is available, but in machine learning contexts like this, there is no "correct" or clearly neutral way to convert loss to utility. Our strategy avoids this issue by working directly in terms of malfare and loss.

The most poignant contrast to existing work we can make is to the *Seldonian learner* (Thomas et al., 2019) framework, which can be thought of as extending 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 fair PAC learning is that it is sophisticated enough to handle fairness issues, with a particular axiomatically justified objective, but remains simple enough to study as a mathematical object; in particular reductions between various PAC and FPAC learnable classes are of great value in understanding FPAC learning, which would not be possible in a more general framework.

3 Quantifying Population-Level Sentiment

A generic *population mean* function $\mathrm{M}(\mathcal{S};\boldsymbol{w})$ quantifies some *sentiment value* $\mathcal{S}$, across a population Ω weighted by $\boldsymbol{w}$. In particular, $\mathcal{S}:\Omega \to \mathbb{R}_{0+}$ describes the *values* over which we take the mean, and $\boldsymbol{w}$, a probability measure over Ω, describes their *weights*. When $\mathcal{S}$ measures a *desirable quantity*, generally termed *utility*, the population mean is a measure of *cardinal welfare* (Moulin, 2004), and thus quantifies overall *well-being*. We also consider the inverse-notion of *ill-being*, termed *malfare*, in terms of an *undesirable* $\mathcal{S}$, generally *loss* or *risk*, which naturally extends the concept. We show an equivalent *axiomatic justification* for malfare, and argue that its use is more natural in many situations, particularly when considering or optimizing *loss functions*.

Definition 3.1 (Population Means: Welfare and Malfare). *A* population mean *function* $\mathrm{M}(\mathcal{S};\boldsymbol{w})$ *measures the* overall sentiment *of population* Ω*, measured by* sentiment function $\mathcal{S}:\Omega \to \mathbb{R}_{0+}$*, weighted by* probability measure $\boldsymbol{w}$ *over* Ω*. If* $\mathcal{S}$ *denotes a desirable quantity (i.e., 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 (i.e., disutility, loss, or risk), we call it a* malfare function*, written* $\mathrm{M\!\!\!\!M}(\mathcal{S};\boldsymbol{w})$*.*

For now, think of the term *population mean* as signifying that an entire population, with diverse and subjective desiderata, is considered and summarized, rather than a single objective viewpoint. As we introduce axioms and show consequent properties, the appropriateness of the term shall become more apparent. Note that 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 pre-existing concept. Coarsely speaking, the three notions are identical, all being functions of the form $(\Omega \to \mathbb{R}_{0+}) \times \textsc{Measure}(\Omega, 1) \to \mathbb{R}_{0+}$, however we shall see that in order to promote fairness, the desirable axioms of malfare and welfare functions differ slightly. The notation reflects this; $\mathrm{M}(\mathcal{S};\boldsymbol{w})$ is an M for *mean*, whereas $\mathrm{W}(\mathcal{S};\boldsymbol{w})$ is a W for *welfare*, and $\mathrm{M\!\!\!\!M}(\mathcal{S};\boldsymbol{w})$ is an $\mathrm{M\!\!\!\!M}$ (*inverted* W), to emphasize its inverted nature.

Often we are interested in *unweighted* population means, given sentiments as a *vector* $\mathcal{S} \in \mathbb{R}^g_{0+}$. Unweighted means may be defined in terms of weighted means, as

$$\mathrm{M}(\mathcal{S}) \doteq \mathrm{M}\left(i \mapsto \mathcal{S}_i; \{i \mapsto \tfrac{1}{g}\}\right) \enspace ,$$

abusing notation to concisely express the uniform measure. Indeed, it may seem antithetical to fairness to allow for weights in malfare and welfare definitions; consider however that weights can represent *differential population sizes*, and thus ensure that the welfare or malfare of *weight-preserving decompositions* of groups into subgroups with equal risk or utility remains constant.

Example 3.2 (Utilitarian Welfare)**.** *Suppose individuals reside in some space* $\mathcal{X}$*, where* distributions $\mathcal{D}_{1:g}$ *over domain* $\mathcal{X}$ *describe the distribution over individuals in* each group*. Suppose also* utility function $\mathrm{U}(x) : \mathcal{X} \to \mathbb{R}_{0+}$*, describing the* level of satisfaction *of an individual e.g., w.r.t. some particular* situation*,* allocation*, or* classifier*. We now take* sentiment function *to be the* mean utility *(per-group), i.e.,*

$$\mathcal{S}(\omega_i) \doteq \mathop{\mathbb{E}}_{x \sim \mathcal{D}_i}[\mathrm{U}(x)] = \mathop{\mathbb{E}}_{\mathcal{D}_i}[\mathrm{U}] \enspace .$$

Now, given a weights vector $\boldsymbol{w}$*, describing the* relative frequencies *of each of the g groups, we define the* utilitarian welfare *as*

$$\mathrm{W}_1(\mathcal{S}; \boldsymbol{w}) \doteq \sum_{i=1}^{g} \boldsymbol{w}(\omega_i)\mathcal{S}(\omega_i) = \mathop{\mathbb{E}}_{\omega \sim \boldsymbol{w}}[\mathcal{S}(\omega)] = \mathop{\mathbb{E}}_{\boldsymbol{w}}[\mathcal{S}] \enspace .$$

Of course, in statistical, sampling, and machine learning contexts, $\mathcal{D}_{1:g}$ and $\boldsymbol{w}$ may be unknown, so we now discuss an *empirical analogue*. Section 4 is then devoted to showing how and when empirical population means well-approximate their true counterparts.

Example 3.3 (Empirical Utilitarian Welfare)**.** *Now suppose* $\mathcal{D}_{1:g}$ *are unknown, but instead, we are given a* sample $\boldsymbol{x}_{1:g,1:m} \in \mathcal{X}^{g \times m}$*, where* $\boldsymbol{x}_{i,1:m} \sim \mathcal{D}_i$*. We define an* empirical analogue *of the utilitarian welfare as in example 3.2, instead taking*

$$\hat{\mathcal{S}}(\omega_i) \doteq \mathop{\hat{\mathbb{E}}}_{x \in \boldsymbol{x}_i}[\mathrm{U}(x)] \enspace , \quad \hat{\mathrm{W}}_1(\hat{\mathcal{S}}, \boldsymbol{w}) \doteq \mathop{\mathbb{E}}_{\boldsymbol{w}}[\hat{\mathcal{S}}] \enspace .$$

Similarly, if $\boldsymbol{w}$ *is unknown, but we may sample from some* $\mathcal{D}$ *over* $\Omega \times \mathcal{X}$*, we can use* empirical frequencies $\hat{\boldsymbol{w}}$ *in place of* true frequencies $\boldsymbol{w}$*, and define* $\hat{\mathcal{S}}(\omega_i)$ *as* conditional averages *over the subsample associated with group* i*.*

3.1 Axioms of Cardinal Welfare and Malfare

Definition 3.4 (Axioms of Cardinal Welfare and Malfare)**.** *We define the* population-mean axioms *for population-mean function* $\mathrm{M}(\mathcal{S}; \boldsymbol{w})$ *below. For each item, assume (if necessary) that the axiom applies* $\forall \mathcal{S}, \mathcal{S}' \in \Omega \to \mathbb{R}_{0+}$*, scalars* $\alpha, \beta \in \mathbb{R}_{0+}$*, and probability measures* $\boldsymbol{w}$ *over* Ω*.*

1. *(Strict) Monotonicity:* $\forall \boldsymbol{\varepsilon} : \Omega \to \mathbb{R}_{0+}$ *s.t.* $\int_{\boldsymbol{w}} \boldsymbol{\varepsilon}(\omega) \, \mathrm{d}(\omega) > 0$*:* $\mathrm{M}(\mathcal{S}; \boldsymbol{w}) < \mathrm{M}(\mathcal{S} + \boldsymbol{\varepsilon}; \boldsymbol{w})$*.*
2. *Symmetry:* $\forall$ *permutations* π *over* Ω*:* $\mathrm{M}(\mathcal{S}; \boldsymbol{w}) = \mathrm{M}(\pi(\mathcal{S}); \pi(\boldsymbol{w}))$*.*
3. *Continuity:* $\{\mathcal{S}' \mid \mathrm{M}(\mathcal{S}'; \boldsymbol{w}) \leq \mathrm{M}(\mathcal{S}; \boldsymbol{w})\}$ *and* $\{\mathcal{S}' \mid \mathrm{M}(\mathcal{S}'; \boldsymbol{w}) \geq \mathrm{M}(\mathcal{S}; \boldsymbol{w})\}$ *are closed sets.*
4. *Independence of unconcerned agents: Suppose subpopulation* $\Omega' \subseteq \Omega$*. Then*
$$\mathrm{M}\left(\begin{cases} p \in \Omega' : \alpha \\ p \notin \Omega' : \mathcal{S}(p) \end{cases}; \boldsymbol{w}\right) \leq \mathrm{M}\left(\begin{cases} p \in \Omega' : \alpha \\ p \notin \Omega' : \mathcal{S}'(p) \end{cases}; \boldsymbol{w}\right) \implies \mathrm{M}\left(\begin{cases} p \in \Omega' : \beta \\ p \notin \Omega' : \mathcal{S}(p) \end{cases}; \boldsymbol{w}\right) \leq \mathrm{M}\left(\begin{cases} p \in \Omega' : \beta \\ p \notin \Omega' : \mathcal{S}'(p) \end{cases}; \boldsymbol{w}\right) \enspace .$$
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:* $\mathrm{M}(\mathbf{1}; \boldsymbol{w}) = \mathrm{M}(1, \ldots, 1; \boldsymbol{w}) = 1$*.*
8. *Pigou-Dalton transfer principle: Suppose* $\mu = \mathbb{E}_{\boldsymbol{w}}[\mathcal{S}] = \mathbb{E}_{\boldsymbol{w}}[\mathcal{S}']$*, and* $\forall p \in \Omega : \left|\mu - \mathcal{S}'(p)\right| \leq \left|\mu - \mathcal{S}(p)\right|$*. Then* $\mathrm{W}(\mathcal{S}'; \boldsymbol{w}) \geq \mathrm{W}(\mathcal{S}; \boldsymbol{w})$*.*
9. *Anti-Pigou-Dalton transfer principle: Suppose as in 8, and conclude* $\mathrm{M}(\mathcal{S}'; \boldsymbol{w}) \leq \mathrm{M}(\mathcal{S}; \boldsymbol{w})$*.*

We take a moment to comment on each of these axioms, to preview their purpose and assure the reader of their necessity. Axioms 1-5 are the standard *axioms of cardinal welfarism* (1-4 are discussed by Sen (1977); Roberts (1980), and 5 by Debreu (1959); Gorman (1968)). Together, they imply (via the Debreu-Gorman theorem) that any population-mean can be decomposed as a *monotonic function* of a *sum* (over groups) of *log* or *power* functions. Axiom 6 is a natural and useful property, and ensures that *dimensional analysis* on mean functions is possible: in particular, the *units* of mean functions match those of sentiment. Note that axiom 6 implies axiom 4, and it is thus a simple strengthening of a traditional cardinal welfare axiom. This axiom also ensures that *units* of population means preserve the *units* of $\mathcal{S}$, making dimensional analysis greatly more convenient; we will also see that it is essential to show convenient *statistical* and *learnability* properties. Axiom 7 furthers this theme, as it ensures that not only do *units* of means match those of $\mathcal{S}$, but scale does as well (making comparisons like $\mathcal{S}_i$ is *above* the population-welfare meaningful), and also enabling comparison *across populations* (in the sense that comparing *averages* is more meaningful than *sums*). Finally, axiom 8 (the *Pigou-Dalton transfer principle* (see Pigou, 1912; Dalton, 1920)) 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 an expansion of inequality, but for malfare, which we generally wish to *minimize*, the opposite occurs, thus this axiom characterizes *fairness* in the context of *malfare.*

Axioms 6-4 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 population mean 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."

For context, we present an additional axiom; that of *additive separability.* For simplicity, we present it only in the unweighted discrete case, as there is some subtlety to an equivalent measure-theoretic formulation, and we derive no benefit from assuming this axiom, as it is largely incompatible with our assumptions, and presented only for comparison purposes.

Definition 3.5 (Additive Separability)**.** *Population-mean* $\mathrm{M}(\mathcal{S}_{1:g})$ *is* additively separable *if there exist functions* $f_{1:g}$ *s.t. each* $f_i \in \mathbb{R}_{0+} \to \mathbb{R}_{0+}$*, and* $\mathrm{M}(\mathcal{S}_{1:g})$ *may be decomposed as*

$$\mathrm{M}(\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_g) = \sum_{i=1}^{g} f_i(\mathcal{S}_i) \enspace .$$

While seemingly quite important to early welfare theorists (i.e., the Debreu-Gorman theorem is generally presented in additively-separable form), and it gives rise to some convenient interpretability and computational properties, we argue that these are far-outstripped by those stemming from the *multiplicative linearity* and *unit scale* axioms, with no real difference in generality.[1] Furthermore, unlike these and other standard cardinal welfare axioms, additive separability seems a bit-heavy handed, assuming something very specific that is supposedly convenient for the economist, with little justification as to why and how it serves as a fundamental property of *cardinal welfare* itself. These properties are discussed in section 3.3, and directly compared with the additively-separable form in section 3.4.

3.2 The Power Mean

We now define the *p-power mean* $\mathrm{M}_p(\cdot)$, for any $p \in \mathbb{R} \cup \pm\infty$, which we shall use to quantify both malfare and welfare. Power means arise often when analyzing population means obeying the various axioms of definition 3.4, and as we shall see in theorem 3.7, are a particularly important class of population means.

Definition 3.6 (Power-Mean Welfare and Malfare)**.** *Suppose* $p \in \mathbb{R} \cup \pm\infty$*. We first define the* unweighted

[1] In this sense, the *additive welfare functionals* are somewhat like the *natural sufficient statistics* of the exponential family, in that, while additivity is sometimes convenient, it comes only with the sacrifice of other desiderata.

power mean *of sentiment vector* $\mathcal{S} \in \mathbb{R}_{0+}^{g}$ *as*

$$\mathrm{M}_p(\mathcal{S}) \doteq \begin{cases} p \in \mathbb{R} \setminus \{0\} & \sqrt[p]{\frac{1}{g}\sum_{i=1}^{g} \mathcal{S}_i^p} \\ p = -\infty & \min_{i \in 1,\ldots,g} \mathcal{S}_i \\ p = 0 & \sqrt[g]{\prod_{i=1}^{g} \mathcal{S}_i} = \exp\left(\frac{1}{g}\sum_{i=1}^{g} \ln(\mathcal{S}_i)\right) \\ p = \infty & \max_{i \in 1,\ldots,g} \mathcal{S}_i \end{cases} .$$

We now define the weighted power mean*, given* sentiment value function $\mathcal{S} : \Omega \to \mathbb{R}_{0+}$ *and probability measure* $\boldsymbol{w}$ *over* Ω*, as*

$$\mathrm{M}_p(\mathcal{S}; \boldsymbol{w}) \doteq \begin{cases} p \in \mathbb{R} \setminus \{0\} & \sqrt[p]{\int_{\boldsymbol{w}} \mathcal{S}^p(\omega)\,\mathrm{d}(\omega)} = \sqrt[p]{\mathbb{E}_{\omega \sim \boldsymbol{w}}[\mathcal{S}^p(\omega)]} \\ p = -\infty & \min_{\omega \in \mathrm{Support}(\boldsymbol{w})} \mathcal{S}(\omega) \\ p = 0 & \exp\left(\int_{\boldsymbol{w}} \ln \mathcal{S}(\omega)\,\mathrm{d}(\omega)\right) = \exp\left(\mathbb{E}_{\omega \sim \boldsymbol{w}}[\ln \mathcal{S}(\omega)]\right) \\ p = \infty & \max_{\omega \in \mathrm{Support}(\boldsymbol{w})} \mathcal{S}(\omega) \end{cases} .$$

In both the weighted and unweighted cases, $p \in \{-\infty, 0, \infty\}$ *resolve as their (unique) limits, and for all* $p \in \mathbb{R}$*, note that* power means *are special cases of the (weighted)* generalized mean*, defined for strictly monotonic* f *as* $\mathrm{M}_f(\mathcal{S}; \boldsymbol{w}) \doteq f^{-1}\big(\mathbb{E}_{\omega \sim \boldsymbol{w}}[\ln \mathcal{S}(\omega)]\big)$*.*

Theorem 3.7 (Properties of the Power-Mean)**.** *Suppose* $\mathcal{S}, \boldsymbol{\varepsilon}$ *are loss values* $\Omega \to \mathbb{R}_{0+}$*, and* $\boldsymbol{w}$ *is a probability measures over some space* Ω*. Then*

1. *Monotonicity:* $\mathrm{M}_p(\mathcal{S}; \boldsymbol{w})$ *is weakly-monotonically-increasing in* p*, and strictly if* $\mathcal{S}$ *attains distinct* $a, b \in \mathbb{R}$ *with nonnegligible probability.*
2. *Subadditivity:* $\forall p \geq 1 : \mathrm{M}_p(\mathcal{S} + \boldsymbol{\varepsilon}; \boldsymbol{w}) \leq \mathrm{M}_p(\mathcal{S}; \boldsymbol{w}) + \mathrm{M}_p(\boldsymbol{\varepsilon}; \boldsymbol{w})$*.*
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 convex in* $\mathcal{S}$ *for* $p \in [1, \infty]$ *and concave for* $p \in [-\infty, 1]$*.*

3.3 Properties of Welfare and Malfare Functions

We now show that the axioms of definition 3.4 are sufficient to characterize many properties of welfare and malfare.

Theorem 3.8 (Population Mean Properties)**.** *Suppose population-mean function* $\mathrm{M}(\mathcal{S}; \boldsymbol{w})$*. If* $\mathrm{M}(\cdot; \cdot)$ *satisfies (subsets of) the population-mean axioms (see definition 3.4), we have that* $\mathrm{M}(\cdot; \cdot)$ *exhibits the following properties. For each, assume arbitrary sentiment-value function* $\mathcal{S} : \Omega \to \mathbb{R}_{0+}$ *and weights measure* $\boldsymbol{w}$ *over* Ω*. The following then hold.*

1. Identity*: Axioms 6 & 7 imply* $\mathrm{M}(\omega \mapsto \alpha; \boldsymbol{w}) = \alpha$*.*
2. *Axioms 1-5 imply* $\exists p \in \mathbb{R}$*, strictly-monotonically-increasing continuous* $F : \mathbb{R} \to \mathbb{R}_{0+}$ *s.t.*

$$\mathrm{M}(\mathcal{S}; \boldsymbol{w}) = F\left(\int_{\boldsymbol{w}} f_p(\mathcal{S}(\omega))\,\mathrm{d}(\omega)\right) = F\left(\mathbb{E}_{\omega \sim \boldsymbol{w}}\big[f_p(\mathcal{S}(\omega))\big]\right) \ , \ \textit{with} \begin{cases} p = 0 & f_0(x) \doteq \ln(x) \\ p \neq 0 & f_p(x) \doteq \mathrm{sgn}(p) x^p \end{cases} .$$

3. *Axioms 1-7 imply* $F(x) = f_p^{-1}(x)$*, thus* $\mathrm{M}(\mathcal{S}; \boldsymbol{w}) = \mathrm{M}_p(\mathcal{S}; \boldsymbol{w})$*.*
4. *Axioms 1-5 and 8 imply* $p \in (-\infty, 1]$*.*

5. *Axioms 1-5 and 9 imply* $p \in [1, \infty)$.

Taken together, the items of theorem 3.8 tell us that the mild conditions of axioms 1-4 (generally assumed for welfare), along with *multiplicative linearity*, imply that welfare and utility, or malfare and loss, are measured in the *same units* (e.g., *nats* or *binits* for *cross-entropy loss*, square-$\mathcal{Y}$-units for *square error*, or dollars for *income utility*), and power-mean malfare is effectively the only reasonable choice of welfare or malfare function. Even without multiplicative linearity, axioms 1-5 imply population mean functions are still *monotonic transformations* of power-mean. Furthermore, the entirely milquetoast *unit scale* axiom implies that sentiment values and population means have the same *scale*, making comparisons like "the risk of group i is above (or below) the population malfare" meaningful. Finally, we also have that $p \in [-\infty, 1)$ incentivizes redistribution of utility from better-off groups to worse-off groups, and similarly $p \in [1, \infty)$ incentivizes redistribution of harm[2] from worse-off groups to better-off groups.

3.4 A Comparison with the Additively Separable Form

In particular, assuming axioms 1-5, all population means are *monotonic transformations* of the power mean, thus whether we assume *additive separability*, and get the additively-separable form

$$\mathrm{M}(\mathcal{S}) = c \sum_{i=1}^{g} f_p(\mathcal{S}_i) = cg\mathrm{M}_p^p(\mathcal{S}) \ ,$$

for $p \in \mathbb{R} \setminus \{0\}$ (where usually $c = 1$ is taken as the canonical form), or we assume axioms 6 & 7, and get the power-mean, there is no real loss of descriptiveness, as all such population-means remain *isomorphic* under the binary comparison operator ($\leq$). With additive separability, it is straightforward to compute the welfare of a population from the welfares of *subpopulations*, but it is still computationally trivial to do this with power means. Furthermore, the limiting cases of $p \in \pm\infty$ become undefined in the additively separable form, and we lose monotonicity in p (see theorem 3.7), both of which are remedied with power means.

With additive separability, welfare summarizes population sentiment by intuitively generalizing the idea of *summation*. In contrast, in our setting, we prefer to think of welfare as a generalized *average*. This yields desirable statistical estimation and learnability properties (shown in the sequel), but is also useful in and of itself, as it allows us to, for instance compare individual sentiment values to welfare, as both the *units* and *scale* match.

Another reason to prefer the power-mean over the additively separable form is the potential for direct comparisons between *group sentiments*, population means of *subgroups*, and *overall population means*. In particular, the *dimensional analysis* properties of the *power mean* are convenient, as these comparisons agree in *units* (due to axiom 6) and *scale* (due to axiom 7). No such dimensional analysis is possible with the additively separable form, as, e.g., if $\mathcal{S}$ is measured in dollars, then $\mathrm{W}_2^2(\mathcal{S}; \boldsymbol{w})$ is measured in *square dollars* (a rather unintuitive unit).

4 Statistical Estimation of Malfare Values

We first illustrate the ease with which p-power means can be estimated, in contrast to the standard additive welfare formulations. Perhaps surprisingly, we find that empirical welfare and malfare are *biased estimators*, yet they admit much sharper finite-sapmle tail bounds than the additive welfare formulations, which are unbiased.

Lemma 4.1 (Statistical Estimation)**.** *Suppose probability distribution $\mathcal{D}$, population-mean* M *obeying* monotonicity*, sentiment value function $\mathcal{S}$ such that, given functions f_ω, we have $\mathcal{S}(\omega) = \mathbb{E}_{x \sim \mathcal{D}}[f_\omega(x)]$, sample $\boldsymbol{x} \sim \mathcal{D}^m$, and* empirical sentiment value estimate $\hat{\mathcal{S}} \doteq \hat{\mathbb{E}}_{x \in \boldsymbol{x}}[f_\omega(x)]$*. If it holds with probability at least $1 - \delta$ that $\forall \omega : \mathcal{S}'(\omega) - \boldsymbol{\varepsilon}(\omega) \leq \mathcal{S}(\omega) \leq \mathcal{S}'(\omega) + \boldsymbol{\varepsilon}(\omega)$, then with said probability, we have*

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

[2]Note that, mathematically speaking, it is entirely *valid* to quantify welfare with $p > 1$ or malfare with $p < 1$, and indeed such characterizations may arise in the analysis of unfair systems; however we generally advocate against intentionally creating such unfair systems.

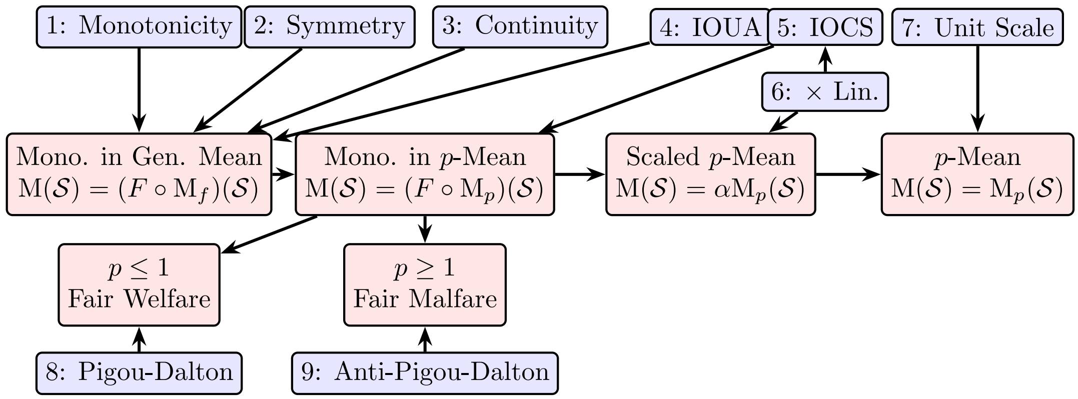

Figure 1: Relationships between population-mean axioms and properties. Assumptions and axioms shown in pastel blue, and properties shown in pastel red. Equivalent properties hold for weighted population mean functions.

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

Proof. This result follows from the assumption, and the *monotonicity* axiom (i.e., adding or subtracting $\boldsymbol{\varepsilon}$ can not decrease or increase the power mean, respectively). The minimum with 0 on the LHS is required simply because by definition, sentiment values are nonnegative, and M_p is in general undefined with negative inputs. $\square$

We now reify this result, applying the well-known Hoeffding (1963) and Bennett (1962) bounds to show concentration and derive an explicit form for $\boldsymbol{\varepsilon}$.

Corollary 4.2 (Statistical Estimation with Hoeffding and Bennett Bounds)**.** *Suppose fair power-mean malfare* $\mathrm{M}(\cdot;\cdot)$ *(i.e.,* $p \geq 1$*),* loss function $\ell : \mathcal{X} \to [0, r]$, $\mathcal{S} \in [0, r]^g$ *s.t.* $\mathcal{S}_i = \mathbb{E}_{\mathcal{D}_i}[\ell]$*, samples* $\boldsymbol{x}_i \sim \mathcal{D}_i^m$*, and take* $\hat{\mathcal{S}}_i \doteq \frac{1}{m}\sum_{j=1}^{m} \ell(\boldsymbol{x}_{i,j})$*. Then with probability at least* $1 - \delta$ *over choice of* $\boldsymbol{x}$*,*

$$\left|\mathrm{M}_p(\mathcal{S}; \boldsymbol{w}) - \mathrm{M}_p(\hat{\mathcal{S}}; \boldsymbol{w})\right| \leq r\sqrt{\frac{\ln \frac{2g}{\delta}}{2m}} \enspace .$$

Alternatively, again with probability at least $1 - \delta$ *over choice of* $\boldsymbol{x}$*, we have*

$$\left|\mathrm{M}_p(\mathcal{S}; \boldsymbol{w}) - \mathrm{M}_p(\hat{\mathcal{S}}; \boldsymbol{w})\right| \leq \frac{r \ln \frac{2g}{\delta}}{3m} + \max_{i \in 1, \ldots, g} \sqrt{\frac{2\,\mathbb{V}_{\mathcal{D}_i}[\ell] \ln \frac{2g}{\delta}}{m}} \enspace .$$

As corollary 4.2 follows directly from lemma 4.1, with Hoeffding and Bennett inequalities applied to derive $\boldsymbol{\varepsilon}$ bounds, similar results are immediately possible with arbitrary concentration inequalities. In particular, we can show *data-dependent* bounds may be shown, e.g., with *empirical Bennett bounds*, removing dependence on *a priori known variance*.

Furthermore, note that while these bounds may be used for *evaluating* the welfare or malfare of a *particular* classifier or mechanism (through $\mathcal{S}$ and $\hat{\mathcal{S}}$), they immediately extend to *learning* over a *finite family* via the union bound. Much like in standard statistical learning theory, the exponential tail bounds of corollary 4.2 allow the family to grow *exponentially*, at *linear* cost to sample complexity. Also as in standard uniform convergence analysis (generally discussed in the context of *empirical risk minimization*), we can easily handle infinite hypothesis classes and obtain much sharper bounds by considering *uniform convergence bounds* over the family, i.e., with Rademacher averages and appropriate concentration-of-measure bounds.

In learning contexts, minimizing the *malfare* among all groups immediately generalizes minimizing *risk* of a single group. These statistical estimation bounds immediately imply that the *empirical malfare-optimal* solution is a reasonable proxy for the true malfare-optimal solution, as we now formalize.

Definition 4.3 (The Empirical Malfare Minimization (EMM) Principle)**.** *Suppose hypothesis class* $\mathcal{H} \subseteq \mathcal{X} \to \mathcal{Y}$*, training samples* $\boldsymbol{z}_{1:g}$ *drawn from distributions* $\mathcal{D}_{1:g}$ *over* $\mathcal{X} \times \mathcal{Y}$*, loss function* $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{0+}$*, malfare function* M*, group weights* $\boldsymbol{w}$*. The* empirical malfare minimizer *is then defined as*

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

and the EMM principle states that $\hat{h}$ *is a reasonable proxy for the true malfare minimizer*

$$h^* \doteq \operatorname*{argmin}_{h^* \in \mathcal{H}} \mathrm{M}\left(i \mapsto \mathrm{R}(h^*; \ell, \mathcal{D}_i); \boldsymbol{w}\right) \enspace .$$

5 Statistical and Computational Efficiency Learning Guarantees

In this section, we define a formal notion of fair-learnability, termed *fair PAC-learning*, where a loss function and hypothesis class are fair PAC-learnable essentially if any distribution can be approximately learned from a polynomially-sized sample (w.h.p.). We then construct various fair-PAC learners, and relate the concept to standard PAC learning (Valiant, 1984), with the understanding that this allows the vast breadth of research of PAC-learning algorithms, and quite saliently, necessary and sufficient conditions, to be applied to fair-PAC learning. In particular, we show a hierarchy of fair-learnability via generic statistical and computational learning theoretic bounds and reductions.

Hypothesis Classes and Sequences We now define *hypothesis class sequences*, required to discuss non-trivial *computational complexity* in learning. This definition is adapted from (Shalev-Shwartz and Ben-David, 2014, Def. 8.1), which treats only *binary classification*.

Definition 5.1 (Hypothesis Class Sequence)**.** *A* hypothesis class sequence $\mathcal{H} = (\mathcal{H}_1, \mathcal{H}_2, \dots)$ *is a* nested sequence *of* hypothesis classes *mapping* $\mathcal{X} \to \mathcal{Y}$*. In other words,* $\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \dots$*.*

Usually, $\mathcal{H}_i$ is easily derived from $\mathcal{H}_j$. For instance, *linear classifiers* naturally form a sequence of families using their *dimension*:

$$\mathcal{H}_d \doteq \left\{ \vec{x} \mapsto \operatorname{sgn}\bigl(\vec{x} \cdot (\vec{w} \circ \vec{0})\bigr) \,\middle|\, \vec{w} \in \mathbb{R}^d \right\} \enspace .$$

Here each $\mathcal{H}_i$ is defined over domain $\mathbb{R}^\infty$, but it is more natural to discuss them as objects over $\mathbb{R}^d$. Similarly, unit-scale *univariate polynomial regression* naturally decomposes as

$$\mathcal{H}_d \doteq \left\{ x \mapsto (x, x^2, \dots, x^d) \cdot \vec{w} \,\middle|\, \vec{w} \in [-1, 1]^d \right\} \enspace .$$

The *hypothesis class sequence* concept is necessary, as it allows us to distinguish statistically easy problems, like learning hyperplanes in finite-dimensional space, from the statistically challenging problem of learning hyperplanes in $\mathbb{R}^\infty$. It is also important to studying computational-hardness of statistically-easy learning problems, as this essentially boils down to selecting a hypothesis class sequence such that learning *time complexity* grows *superpolynomially* (in d).

For context, we first present a generalized notion of PAC-learnability, which we then generalize to fair-PAC-learnability. Standard presentations consider only classification under 0-1 loss, but following the generalized learning setting of Vapnik (2013), some authors consider generalized notions for other learning problems (see, e.g., Shalev-Shwartz and Ben-David (2014, Definition 3.4)).

Definition 5.2 (PAC-Learnability)**.** *Suppose* nested 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* $\exists$ *a (randomized) algorithm* $\mathcal{A}$*, such that* $\forall$*:*

1. *sequence index* d*;*
2. *instance distribution* $\mathcal{D}$ *over* $\mathcal{X} \times \mathcal{Y}$*;*
3. *additive approximation error* $\varepsilon > 0$*; and*

4. *failure probability* $\delta \in (0,1)$;

it holds that $\mathcal{A}$ *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* $\mathrm{m}(\varepsilon, \delta, d) : \big(\mathbb{R}_+ \times (0,1) \times \mathbb{N}\big) \to \mathbb{N}$ *s.t.* $\mathcal{A}(\mathcal{D}, \varepsilon, \delta, d)$ *consumes no more than* $\mathrm{m}(\varepsilon, \delta, d)$ *samples from* $\mathcal{D}$ *(i.e., has finite sample complexity); and*

2. *with probability at least* $1 - \delta$ *(over randomness of* $\mathcal{A}$*),* $\hat{h}$ *obeys*

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

The class of such learning problems is denoted PAC*, thus we write* $(\mathcal{H}, \ell) \in \mathrm{PAC}$ *to denote PAC-learnability.*
Furthermore, if for all d, the space of $\mathcal{D}$ *is restricted such that*

$$\exists h \in \mathcal{H}_d \ s.t.\ \mathrm{R}(h; \ell, \mathcal{D}) = 0 \ ,$$

then $(\mathcal{H}, \ell)$ *is* realizable-PAC-learnable, *written* $(\mathcal{H}, \ell) \in \mathrm{PAC}^{\mathrm{Rlz}}$.

Observation 5.3 (On Realizable Learning)**.** *Our definition of realizability appears to differ from the standard form, in which* $\mathcal{D}$ *is a distribution over only* $\mathcal{X}$*, and* y *is simply computed as* $h^*(x)$*, for some* $h^* \in \mathcal{H}$*. We instead constrain* $\mathcal{D}$ *such that there exists a 0-risk* $h^* \in \mathcal{H}$*, which is equivalent for any loss function* ℓ *such that* $\ell(y, \hat{y}) = 0 \Leftrightarrow y = \hat{y}$*, e.g., 0-1 classification loss, absolute or square regression loss. With our definition, it is much clearer that realizable learning is a special case of agnostic learning, and furthermore, we handle a much broader class of problems, for which there may be some amount of* noise*, or wherein a ground truth may not even exist.*

For instance, in a recommender system*,* y *may represent the* set of items *that* x *will like, and* h *may predict a singleton set, and thus we take* $\ell(y, \{\hat{y}\}) = \mathbb{1}_y(\hat{y})$*. There is no ground-truth here, but rather we seek a* compatible solution *that recommends appropriate items to everyone. Similarly, in high-dimensional classification, we often treat classifier output as a* ranked list *of predictions, and the* top-k loss *is taken to be* $\ell(y, \hat{y}) = \mathbb{1}_{\hat{y}_{1:k}}(y)$*. Again, no ground truth exists, but 0-risk learning is still possible if there is not "too much" ambiguity (i.e., foxes and dogs can be confused, as long as they are separated from horses and zebras).*

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 3.3, we do not commit to any particular objective, but instead require that a fair-PAC-learner is able to maximize *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 5.4 (Fair PAC (FPAC)-Learnability)**.** *Suppose* nested 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.* group count g *and* loss function ℓ *if* $\exists$ *a (randomized) algorithm* $\mathcal{A}$*, such that* $\forall$:

1. *sequence index* d;

2. *group count* g;

3. *instance distribution* $\mathcal{D}_{1:g}$ *over* $(\mathcal{X} \times \mathcal{Y})^g$;

4. *group weights measure* $\boldsymbol{w}$ *over* $\{1, \dots, g\}$;

5. *malfare concept* $\mathrm{M}(\cdot; \cdot)$ *satisfying axioms 1-7 and 9;*

6. *additive approximation error* $\varepsilon > 0$*; and*

7. *failure probability* $\delta \in (0,1)$;

it holds that $\mathcal{A}$ *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* $\mathrm{m}(\varepsilon, \delta, d, g) : \big(\mathbb{R}_+ \times (0,1) \times \mathbb{N} \times \mathbb{N}\big) \to \mathbb{N}$ *s.t.* $\mathcal{A}(\mathcal{D}_{1:g}, \boldsymbol{w}, \mathrm{M}, \varepsilon, \delta, d)$ *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*

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

The class of such fair-learning problems is denoted FPAC*, thus we write* $(\mathcal{H},\ell) \in \mathrm{FPAC}$ *to denote fair-PAC-learnability.*

Finally, if for all d*, the space of* $\mathcal{D}$ *is restricted such that*

$$\exists h \in \mathcal{H}_d \textit{ s.t. } \max_{i\in 1,\dots,g} \mathrm{R}(h;\ell,\mathcal{D}_i) = 0 \enspace ,$$

then $(\mathcal{H},\ell)$ *is* realizable-fair-PAC-learnable*, written* $(\mathcal{H},\ell) \in \mathrm{FPAC}^{\mathrm{Rlz}}$*.*

We now observe that a few special cases are familiar learning problems, though we argue that all cases are of interest, and simply represent different ideals of fairness which may be situationally appropriate.

Observation 5.5 (Malfare Functions and Special Cases)**.** *By assumption,* $\text{M}(\cdot;\cdot)$ *must be* $\text{M}_p(\cdot;\cdot)$ *for* $p \in [1,\infty)$*. Taking* $g=1$ *implies* $\boldsymbol{w}=(1)$*, and* $\text{M}_p(\mathcal{S};\boldsymbol{w}) = \mathcal{S}_1$*, 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, commonly encountered in adversarial and robust learning settings.*

An aside on the flexibility of fair-PAC-learnability Note that the generalized definition of (fair) PAC-learnability is sufficiently broad so as to include many *supervised*, *semi-supervised*, and *unsupervised* learning problems. While this is not immediately apparent, consider that, for instance, *k-means clustering* can be expressed as a *learning problem* where the task is to identify a set of k vectors in $\mathbb{R}^d$ where the hypothesis class maps a vector $\vec{x}$ on to the nearest cluster center, and the loss function is the square distance. This is a surprisingly natural fairness issue when cast as a *resource allocation problem*; if, for instance each cluster center represents a cellphone tower, then we seek to place towers to serve *all groups*, and to avoid serving one or more groups particularly well at the expense of others.

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. The main focus of this manuscript is sample complexity, but for completeness, we note that a similar concept of *polynomial-time fair-PAC learnability* is equally interesting, where here we assume $\mathcal{A}(\mathcal{D}_{1:g},\boldsymbol{w},\text{M},\varepsilon,\delta,d)$ 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, and thus implicitly the same bound on sample complexity. We denote these concepts $\mathrm{PAC}^{\mathrm{Agn}}_{\mathrm{Poly}}, \mathrm{PAC}^{\mathrm{Rlz}}_{\mathrm{Poly}}, \mathrm{FPAC}^{\mathrm{Agn}}_{\mathrm{Poly}}, \mathrm{FPAC}^{\mathrm{Rlz}}_{\mathrm{Poly}}$.

Some trivial reductions We first observe (immediately from definitions 5.2 and 5.4) that PAC-learning is a special case of fair-PAC-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, sample-and-compute-efficient *constructive reductions.*

Realizability We first show that in the *realizable case*, *PAC-learnability* implies *fair PAC-learnability.* In particular, we employ a simple and practical *constructive* polynomial-time reduction. More efficient reductions are possible for particular values of p, g. 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) FPAC-learning in terms of algorithms for PAC-learning.

Theorem 5.6 (Realizable Reductions)**.** *Suppose loss function* ℓ *and hypothesis class* $\mathcal{H}$*. Then*

1. $(\mathcal{H}, \ell) \in \mathrm{PAC}^{\mathrm{Rlz}} \implies (\mathcal{H}, \ell) \in \mathrm{FPAC}^{\mathrm{Rlz}}$*; and*

2. $(\mathcal{H}, \ell) \in \mathrm{PAC}^{\mathrm{Rlz}}_{\mathrm{Poly}} \implies (\mathcal{H}, \ell) \in \mathrm{FPAC}^{\mathrm{Rlz}}_{\mathrm{Poly}}$.

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* $\mathcal{A}(\mathcal{D}_{1:g}, \boldsymbol{w}, \mathrm{M}, \varepsilon, \delta, d) \doteq \mathcal{A}'(\mathrm{mix}(\mathcal{D}_{1:g}), \frac{\varepsilon}{g}, \delta, d)$*. 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 sample complexity result. Suppose $\hat{h} \leftarrow \mathcal{A}'(p, \boldsymbol{w}, \mathcal{D}_{1:g}, \varepsilon, \delta, d)$. Then, with probability at least $1 - \delta$ (by the guarantee of $\mathcal{A}$), we have

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

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. □

While mathematically correct, if somewhat trivial, unfortunately, this argument does not extend to the agnostic case. This is essentially because in general it is not possible to simultaneously satisfy all groups, and even with reweighting, the malfare of ERM solutions are highly unstable in the weights. To see this, suppose $\mathcal{Y} \doteq \{a, b\}$, group A always wants a, and group B always wants b, with symmetric preferences, and we wish to optimize egalitarian malfare. For any reweighting, the utility-optimal solution is always to produce all a or all b, except with equal weights, when all solutions are equally good. Thus for no reweighting do all reweighted-risk solutions even approximate the egalitarian-optimal solution (which is evenly split between a and b). We thus conclude that simple constructive reductions using PAC-learners as subroutines is likely to solve the general problem.

In addition to being inextensible to the agnostic case, 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 (and this solution is in fact an equilibrium). Thus unfairness and bias issues logically only arise in a world of *conflict* (i.e., zero-sum settings, or under limited resources constraints, which foster *competition* between groups). Thus while the realizable case is straightforward and solvable, we argue that in practice, the agnostic learning setting is far more relevant.

6 Characterizing Fair Statistical Learnability with FPAC-Learners

We first consider only questions of *statistical learning*. In other words, we ignore computation for now, and show only that *there exist* fair-PAC-learning algorithms of unbounded runtime. 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 fair-PAC learnability are equivalent for learning problems with a particular *no-free-lunch* guarantee.

We first define a *generalized notion* of *uniform convergence* as defined by (Shalev-Shwartz and Ben-David, 2014). In particular, our definition applies to *any bounded loss function*,[3] thus greatly generalizes the standard notion for binary classification.

[3]Boundedness should not be strictly necessary for learnability even uniform convergence, but vastly simplifies all aspects of the analysis. In many cases, it can be relaxed to moment-conditions, such as *sub-Gaussian* or *sub-exponential* assumptions.

Definition 6.1 (Uniform Convergence). *Suppose $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, r] \subseteq \mathbb{R}$ and hypothesis class $\mathcal{H} \subseteq \mathcal{X} \to \mathcal{Y}$. Let $\mathcal{F} \doteq \ell \circ \mathcal{H}$. We say $(\mathcal{H}, \ell) \in \mathrm{UC}$ or $\mathcal{F} \in \mathrm{UC}$ if*

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

Observation 6.2. *We stress that this definition is both uniform over the* function family $\mathcal{F}$ *and uniform over* all possible distributions $\mathcal{D}$. *The standard definition of* uniform convergence in probability *applies to a singular $\mathcal{D}$, however it is standard in PAC-learning and VC theory to assume uniformity over $\mathcal{D}$, so we adopt this convention.*

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 in general true that *uniform convergence* implies *PAC-learnability*; this is well-known for binary classification, but we show the generalized result for completeness. The converse is true for some learning problems, but not for others, which we shall use as a powerful tool to characterize when PAC-learnability implies fair-PAC-learnability. We first state the fundamental theorem of statistical learning (for classification) (Shalev-Shwartz and Ben-David, 2014, thms. 6.2, 29.3).

Theorem 6.3 (Fundamental Theorem of Statistical Learning (Classification)). *Suppose ℓ is the 0-1 loss for k-class classification, where $k < \infty$. Then the following are equivalent.*

1. *$\forall d \in \mathbb{N}$: $\mathcal{H}_d$ has finite Natarajan-dimension (= VC dimension for $k = 2$ classes).*
2. *$\forall d \in \mathbb{N}$: $\ell \circ \mathcal{H}_d$ has the uniform convergence property.*
3. *Any ERM rule is a successful agnostic-PAC learner for $\mathcal{H}$.*
4. *$\mathcal{H}$ is agnostic-PAC learnable.*
5. *Any ERM rule is a successful realizable-PAC learner for $\mathcal{H}$.*
6. *$\mathcal{H}$ is realizable-PAC learnable.*

It is somewhat subtle to generalize this result to arbitrary learning problems. In particular, there are PAC-learnable problems for which uniform convergence *does not hold*. However, Alon et al. (1997) show similar results for various regression problems, with the (scale-sensitive) *γ-fat-shattering dimension* playing the role of the Vapnik-Chervonenkis or Natarjan dimension in classification.

Theorem 6.4 (Fundamental Theorem of Fair Statistical Learning). *Suppose ℓ such that $\forall \mathcal{H} : (\mathcal{H}, \ell) \in \mathrm{PAC}^{\mathrm{Rlz}} \implies (\mathcal{H}, \ell) \in \mathrm{UC}$. Then, for any hypothesis class sequence $\mathcal{H}$, the following are equivalent:*

1. *$\forall d \in \mathbb{N}$: $\ell \circ \mathcal{H}_d$ has the (generalized) uniform convergence property.*
2. *Any EMM rule is a successful agnostic-fair-PAC learner for $(\ell, \mathcal{H})$.*
3. *$(\ell, \mathcal{H})$ is agnostic-fair-PAC learnable.*
4. *Any EMM rule is a successful realizable-fair-PAC learner for $(\ell, \mathcal{H})$.*
5. *$(\ell, \mathcal{H})$ is realizable-fair-PAC 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 3.7 item 3). In particular, take $m \doteq \mathrm{m}_{\mathrm{UC}}(\ell \circ \mathcal{H}_d, \frac{\varepsilon}{2}, \frac{\delta}{g})$, where $\mathrm{m}_{\mathrm{UC}}(\ell \circ \mathcal{H}_d, \varepsilon, \delta)$ upper-bounds the minimum sufficient sample size to ensure ε-δ-uniform convergence over $\ell \circ \mathcal{H}_d$. 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

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

Consequently, as $\mathrm{M}(\cdot; \boldsymbol{w})$ is 1-$\|\cdot\|_\infty$-$|\cdot|$ in risk (see lemma 4.1), it holds with probability at least $1-\delta$ that

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

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 fair-PAC 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, observe that $2 \implies 3$ and $4 \implies 5$ are almost tautological: the existence of (agnostic / realizable) fair-PAC learning algorithms imply (agnostic / realizable) fair-PAC learnability.

Now, $2 \implies 4$ and $3 \implies 5$ hold, as realizable learning is a special case of agnostic learning.

As 1 implies 2-4, which in turn each imply 5, it remains only to show that $5 \implies 1$, i.e., if $\mathcal{H}$ is realizable fair PAC learnable, then $\mathcal{H}$ has the uniform convergence property. In general, the question is rather subtle, but here the assumption "suppose ℓ such that $(\mathcal{H}, \ell) \in \mathrm{PAC}^{\mathrm{Rlz}} \implies (\mathcal{H}, \ell) \in \mathrm{UC}$" does most of the work. In particular, as PAC-learning is a special case of FPAC-learning, we have

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

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

The reductions and equivalences that compose this result are graphically depicted in fig. 2.

Observation 6.5 (The Gap Between Uniform Convergence and (fair) PAC-Learnability)**.** *Note that the assumption "suppose ℓ such that $(\mathcal{H}, \ell) \in \mathrm{PAC}^{\mathrm{Rlz}} \implies (\mathcal{H}, \ell) \in \mathrm{UC}$" does not in general hold. In many cases of interest, it is known to hold, e.g., finite-class classification under 0-1 loss, and bounded regression under square and absolute loss.*

In general, verifying this condition is a rather subtle task, and until more general techniques are developed, must be repeated for each learning problem (loss function). Regardless, this lies beyond the scope of this manuscript, as we have reduced the question of fair-*PAC learnability to one of (generic) PAC-learnability.*

7 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 PAC = 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 5.6 has already characterized the computational complexity of *realizable* FPAC-learning, so we now focus on the agnostic case. In the *agnostic case*, we show neither a generic reduction or non-constructive proof that $\mathrm{PAC}_{\mathrm{Poly}} = \mathrm{FPAC}_{\mathrm{Poly}}$, nor do we show a counterexample; rather we leave this question for future work. We do, however, show that under conditions commonly leveraged as sufficient for efficient PAC-learning, so too is efficient FPAC-learning possible. In particular, section 7.1 provides an efficient constructive reduction (i.e., an algorithm) for efficient FPAC-learning under standard *convex optimization* settings, and section 7.2 shows the same when a small cover of $\mathcal{H}$ exists and may be efficiently enumerated. The results of this section are summarized graphically in fig. 3.

In both the convex optimization and efficient enumeration settings, the proofs take the same general form: we show that *ε-approximate EMM* on m total samples is computationally efficient (in $\mathrm{Poly}(m, \varepsilon, d)$ time), and then argue that so long as *sample complexity* of FPAC-learning via EMM is polynomial, then we may construct an FPAC learner using ε-approximate EMM with polynomial *time complexity.* In particular, the proofs simply account for *optimization* and *sampling* error, and in both cases construct efficient FPAC-learners.

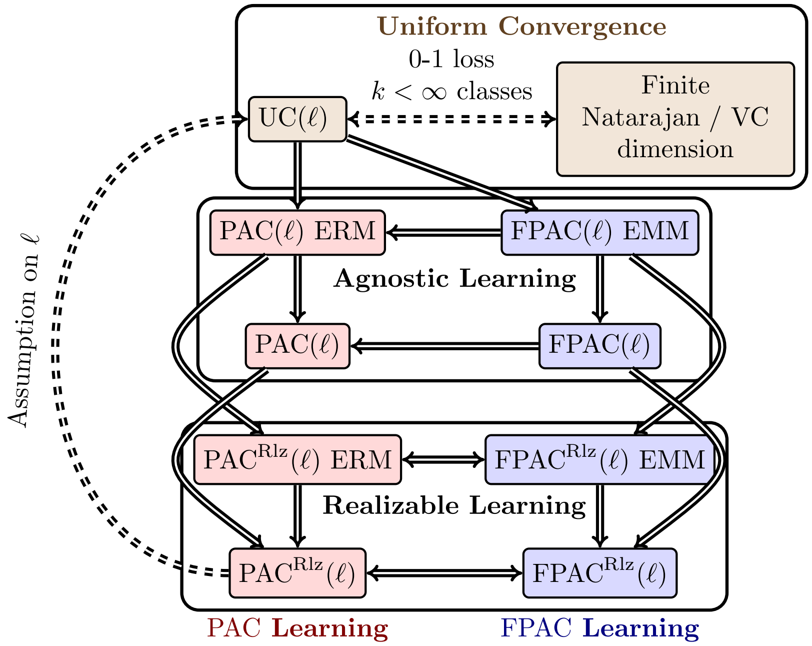

Figure 2: Implications between membership in PAC and fair-PAC classes. In particular, for arbitrary fixed ℓ, implication denotes *implication of membership* of some $\mathcal{H}$ (i.e., containment); see theorems 6.4 and 5.6. Dashed implication arrows hold conditionally on ℓ. Note that when the assumption on ℓ (see theorem 6.4) holds, the hierarchy collapses, and in general, under realizability, some classes are known to coincide.

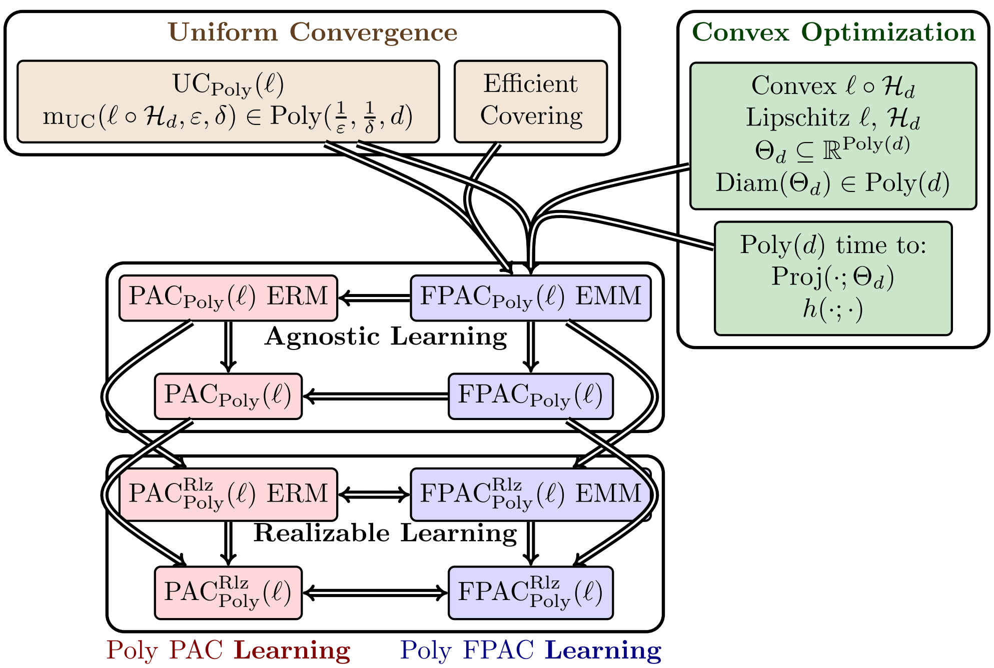

Figure 3: Implications between membership in various poly-time PAC and fair-PAC classes. In particular, for arbitrary but fixed ℓ, implication denotes *implication of membership* of some $\mathcal{H}$ (i.e., containment). See theorems 7.1 and 7.2.

7.1 Efficient FPAC Learning with Convex Optimization

Here we show concretely and constructively the existence of fair-PAC-learners under standard convex optimization assumptions via the *subgradient method* (Shor, 2012), with constants fully derived. Sharper analyses are of course possible, and potential improvements are discussed subsequently, but our result is immediately practical and can be applied verbatim to problems like *generalized linear models* (Nelder and Wedderburn, 1972) and many *kernel methods* with little analytical effort.

Algorithm 1 Approximate Empirical Malfare Minimization via the Subgradient Method

procedure $\mathcal{A}_{\mathrm{PSG}}(\ell, \mathcal{H}, \theta_0, \mathrm{m_{UC}}(\cdot,\cdot), \mathcal{D}_{1:g}, \boldsymbol{w}, \mathrm{M}(\cdot;\cdot), \varepsilon, \delta)$
  Input: λ_ℓ-Lipschitz loss function ℓ, $\lambda_{\mathcal{H}}$-Lipschitz hypothesis class $\mathcal{H}$ with parameter space Θ, initial guess $\theta_0 \in \Theta$, uniform-convergence sample-complexity bound $\mathrm{m_{UC}}(\cdot,\cdot)$, group distributions $\mathcal{D}_{1:g}$, group weights $\boldsymbol{w}$, malfare function $\mathrm{M}(\cdot;\cdot)$, solution optimality guarantee ε-δ.
  Output: ε-δ-$\mathrm{M}(\cdot;\cdot)$-optimal $h \in \mathcal{H}$ (under the conditions of theorem 7.1).
  $\mathrm{m}_{\mathcal{A}} \leftarrow \mathrm{m_{UC}}(\frac{\varepsilon}{3}, \frac{\delta}{g})$ ▷ Determine sufficient sample size
  $\boldsymbol{z}_{1:g,1:\mathrm{m}_{\mathcal{A}}} \sim \mathcal{D}_1^{\mathrm{m}_{\mathcal{A}}} \times \cdots \times \mathcal{D}_g^{\mathrm{m}_{\mathcal{A}}}$ ▷ Draw training sample for each group
  $n \leftarrow \left\lceil \left(\frac{3\,\mathrm{Diam}(\Theta)\lambda_\ell\lambda_{\mathcal{H}}}{\varepsilon} \right)^2 \right\rceil$ ▷ Iteration count
  $\alpha \leftarrow \frac{\mathrm{Diam}(\Theta)}{\lambda_\ell\lambda_{\mathcal{H}}\sqrt{n}}$ ▷ Learning rate ($\approx \frac{\varepsilon}{3\lambda_\ell^2\lambda_{\mathcal{H}}^2}$)
  $f(\theta) : \Theta \mapsto \mathbb{R}_{0+} \doteq \mathrm{M}\big(i \mapsto \hat{\mathrm{R}}(h(\cdot;\theta); \ell, \boldsymbol{z}_i); \boldsymbol{w}\big)$ ▷ 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 ε-δ optimal model
end procedure

Theorem 7.1 (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* $\forall x \in \mathcal{X}_d, \theta \in \Theta_d$*,* $h(x;\theta)$ *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.,*

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

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

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

Finally, assume $\ell \circ \mathcal{H}_d$ *exhibits* ε-δ uniform convergence *with sample complexity* $\mathrm{m_{UC}}(\varepsilon, \delta, d) \in \mathrm{Poly}(\frac{1}{\varepsilon}, \frac{1}{\delta}, d)$*.*

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)*

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

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

It is of course possible to show similar guarantees under relaxed conditions. and with sharper sample complexity and time complexity bounds. The above result merely characterizes a simple and standard convex optimization setting under which standard convex optimization risk minimization guarantees readily translate to malfare minimization.

7.2 Uniform Convergence and Efficient Covering

As we have seen in section 6, uniform convergence implies, and is often equivalent to, (fair) PAC-learnability. However, these results all consider only *statistical learning*, and to analyze *computational learning* questions, we introduce a strengthening of uniform convergence that considers *computation*. We now show sufficient conditions for polynomial-time fair PAC-learnability via *covering numbers*, which we use both to show uniform convergence and to construct an efficient training algorithm.

Here it becomes necessary to consider not the loss function and hypothesis class *in isolation*, but their *composition*, defined as

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

Given this, we show that if a *polynomially-large cover* of each $\ell \circ \mathcal{H}_d$ exists and can be efficiently enumerated, then $(\mathcal{H}, \ell) \in \mathrm{FPAC}_{\mathrm{Poly}}$. In what follows, an ℓ_2-γ-empirical-cover of loss family $(\ell \circ \mathcal{H}_d) \subseteq \mathcal{X} \mapsto \mathbb{R}_{0+}$ on a sample $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^m$ is any $\mathcal{H}_{d,\gamma}$ such that

$$\forall h \in \mathcal{H}_d : \min_{h_\gamma \in \mathcal{H}_{d,\gamma}} \sqrt{\frac{1}{m} \sum_{i=1}^{m} \big((\ell \circ h)(\boldsymbol{z}_i) - (\ell \circ h_\gamma)(\boldsymbol{z}_i)\big)^2} \leq \gamma \ .$$

We take $\mathcal{C}(\ell \circ \mathcal{H}_d, \boldsymbol{z}, \gamma)$ to denote such a cover, and $\mathcal{C}^*(\ell \circ \mathcal{H}_d, \boldsymbol{z}, \gamma)$ to denote such a cover *of minimum cardinality*. Finally, we define the *uniform covering numbers*

$$\mathcal{N}(\ell \circ \mathcal{H}_d, m, \gamma) \doteq \sup_{\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^m} \big|\mathcal{C}^*(\ell \circ \mathcal{H}_d, \boldsymbol{z}, \gamma)\big| \quad \& \quad \mathcal{N}(\ell \circ \mathcal{H}_d, \gamma) \doteq \sup_{m \in \mathbb{N}} \mathcal{N}(\ell \circ \mathcal{H}_d, m, \gamma) \ .$$

This concept is crucial to both our *uniform convergence* and *optimization efficiency* guarantees. In particular, our construction ensures that $\mathcal{N}(\ell \circ \mathcal{H}_d, m, \gamma)$ is sufficiently small so as to ensure *polynomial training time* on a *polynomially-large* training sample is sufficient to FPAC-learn $(\ell, \mathcal{H})$.

With this exposition complete, the FPAC-learning algorithm we present is simply EMM on a cover $\hat{\mathcal{C}}(\ell \circ \mathcal{H}_d, \boldsymbol{z}, \gamma)$ that is not superpolynomially larger than $\mathcal{N}(\ell \circ \mathcal{H}_d, m, \gamma)$, which our constuction ensures exists and may be efficiently enumerated, and upon which we guarantee that a polynomially-large training sample is sufficient. We now state the result, noting that full derivation with exact constants appears in the appendix.

Theorem 7.2 (Efficient FPAC-Learning by Covering)**.** *Suppose loss function ℓ of bounded codomain (i.e., $\|\ell\|_\infty$ is bounded), and hypothesis class sequence $\mathcal{H}$, s.t. $\forall m, d \in \mathbb{N}$, $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^m$, there exist*

1. *a γ-ℓ_2 cover $\mathcal{C}^*(\ell \circ \mathcal{H}_d, \boldsymbol{z}, \gamma)$, where $\big|\mathcal{C}^*(\ell \circ \mathcal{H}_d, \boldsymbol{z}, \gamma)\big| \leq \mathcal{N}(\ell \circ \mathcal{H}_d, \gamma) \in \mathrm{Poly}(\frac{1}{\gamma}, d)$; and*

2. *an algorithm to enumerate a γ-ℓ_2 cover $\hat{\mathcal{C}}(\ell \circ \mathcal{H}_d, \boldsymbol{z}, \gamma)$ of size $\mathrm{Poly}\,\mathcal{N}(\ell \circ \mathcal{H}_d, m, \gamma)$ in $\mathrm{Poly}(m, \frac{1}{\gamma}, d)$ time.*

Then $(\ell, \mathcal{H}) \in \mathrm{FPAC}_{\mathrm{Poly}}$, where (1) is sufficient to show that $(\ell, \mathcal{H})$ is fair-PAC learnable with polynomial sample complexity*, and (2) is required only to show polynomial* training time complexity*.*

This immediately implies that fixed $\mathcal{H}$ that are *finite*, or of bounded *pseudodimension* or *γ-fat-shattering dimension*[4] are fair-PAC-learnable. For instance, this includes *classifiers* such as *all possible languages* of *Boolean formulae* over (constant) d variables, or *halfspaces* (i.e., linear hard classifiers $\mathcal{H} \doteq \{\vec{x} \mapsto \mathrm{sgn}(\vec{x} \cdot \vec{w}) \mid \vec{w} \in \mathbb{R}^d\}$), as well as GLM, subject to regularity constraints to appropriately control the loss function. However, it is perhaps not as powerful as it appears; it applies to *fixed hypothesis classes*, thus each of the above linear models over $\mathbb{R}^d$ is efficiently fair-PAC learnable, but it says nothing about their performance as $d \to \infty$.

This is essentially because the *statistical analysis* to show polynomial *sample complexity* requires only that $\ln \mathcal{N}(\ell \circ \mathcal{H}_d, \gamma) \in \mathrm{Poly}(\frac{1}{\gamma}, d)$, whereas our *training algorithm* must actually *enumerate* an empirical cover, which yields the (exponentially) stronger requirement that $\mathcal{N}(\ell \circ \mathcal{H}_d, m, \gamma) \in \mathrm{Poly}(\frac{1}{\gamma}, d)$ for polynomial *time complexity*. Indeed, we see that while the covering numbers we assume imply uniform convergence with *sample complexity* polynomial in d, when covering numbers grow exponentially in d, then our algorithm

[4]The reader is invited to consult (Anthony and Bartlett, 2009) for an encyclopedic overview of various combinatorial dimensions, associated covering-number and shattering-coefficient concepts, and their applications to statistical learning theory.

yields only *exponential* time complexity in d. Consequently, the result only implies polynomial-time algorithms w.r.t. sequences that grow slowly in complexity; e.g., sequences of linear classifiers that grow only *logarithmically* in dimension $\mathcal{H}_d \doteq \{\vec{x} \mapsto \mathrm{sgn}(\vec{x} \cdot \vec{w}) \mid \vec{w} \in \mathbb{R}^{\lfloor \ln d \rfloor}\}$.

Note also that theorem 7.2 leverages *covering arguments* in both their *statistical* and *computational* capacities. Statistical bounds based on covering are generally well-regarded, particularly when strong analytical bounds on covering numbers are available, although sharper results are possible (e.g., through the *entropy integral* or majorizing measures). Furthermore, while we do construct a *polynomial time* training algorithm, in many cases, specific optimization methods (e.g., stochastic gradient descent or Newton's method) exist to perform EMM *more efficiently* and with *higher accuracy*. Worse yet, *efficient enumerability* of a cover may be non-trivial in some cases; while most covering arguments in the wild are either constructive, or compositional to the point where each component can easily be constructed, it may hold for some problems that computing or enumerating a cover is computationally prohibitive.

On Compositionality and Coverability Conditions The covers and covering numbers discussed above are of course properties of each $\ell \circ \mathcal{H}_d$, rather than ℓ and each $\mathcal{H}_d$ individually. This creates proof obligation for each loss function of interest, in contrast to theorem 7.1, wherein only Lipschitz continuity of ℓ is assumed, and the remaining analysis is on $\mathcal{H}$. Fortunately, in many cases it is still possible to analyze covers of each $\mathcal{H}_d$ in isolation, and then draw conclusions across a broad family of ℓ composed with each $\mathcal{H}_d$. In particular, via standard properties of covering numbers, if ℓ is Lipschitz-continuous w.r.t. some pseudonorm $\|\cdot\|_{\mathcal{Y}}$ over $\mathcal{Y}$, and γ-ℓ_2 covering numbers of each $\mathcal{H}_d$ w.r.t. $\|\cdot\|_{\mathcal{Y}}$ are well-behaved, it can be shown that the conditions of theorem 7.2 are met. This is useful as, for example, regression losses like *square error*, *absolute error*, and *Huber loss* are all Lipschitz continuous on bounded domains, and thus analysis on each $\mathcal{H}_d$ alone is sufficient to apply theorem 7.2 with each such loss function.

8 Conclusion

The central thrust of this paper introduces malfare minimization as a fair learning task, and shows relationships between the statistical and computational issues of malfare and risk minimization. In particular, we argue that our method is more in line with welfare-centric machine learning theory than demographic-parity theory, however malfare is better aligned to address machine learning tasks cast as loss minimization problems then is welfare. As such, the first half of this manuscript is dedicated to deriving and motivating malfare minimization, while the latter studies the problem from statistical and computational learning theoretic perspectives.

8.1 Contrasting Malfare and Welfare

With our framework now fully laid out and initial results presented, we now contrast our malfare-minimization framework with traditional welfare-maximization approaches in greater detail. We don't claim that malfare is a better or more useful concept than welfare; but rather we argue only that it is *significantly different* (with surprising non-equivalence results between power-mean welfare and malfare functions), stands on an equal axiomatic footing, and it stands to reason that the right tool (malfare) should be used for the right job (fair loss minimization). With this said, 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, with a few exceptions, e.g., the *sphericial scoring rule* from decision theory, most supervised learning problems are naturally cast as minimizing nonnegative *loss functions* (arguably via cross-entropy or KL-divergence minimization through maximum-likelihood, either as explicitly intended (Nelder and Wedderburn, 1972), or *ex-post-facto* through subsequent analysis (Cousins and Riondato, 2019)).

We are highly interested in exploring a parallel theory of fair welfare optimization, however some key malfare properties don't hold for welfare. In particular, fair welfare functions $\mathrm{W}_p(\cdot;\cdot)$ for $p \in [0,1)$ are not Lipschitz continuous; for example, the *Nash social welfare* (a.k.a. *unweighted geometric welfare*) $\mathrm{W}_0(\mathcal{S}; \omega \mapsto \frac{1}{g}) = \sqrt[g]{\prod_{i=1}^{g} \mathcal{S}_i}$ is unstable to perturbations of each $\mathcal{S}_i$ around 0, which causes difficulty in both the *statistical* and *computational* aspects of learning. Leveraging this fact, it is trivial to construct welfare-maximization learning problems for which sample complexity is unbounded, making straightforward translation of our

FPAC framework into a welfare setting rather vacuous, except in contrived, trivial, or degenerate cases. Essentially, this is because while corollary 4.2 holds for both welfare and malfare, it does not imply *uniform sample-complexity* bounds, whereas, such bounds are trivial for fair malfare (see lemma 4.1), due to the *contraction property* (theorem 3.7 item 3). It thus seems such a theory 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.

8.2 FPAC Learning: Contributions and Open Questions

After motivating the malfare-minimization machine learning setting, we introduce fair PAC-learning to study the statistical and computational difficulty of malfare minimization. As a generalization of PAC-learning, known hardness results (e.g., lower-bounds on computational and sample complexity of loss minimization) immediately apply, thus, coarsely speaking, the interesting question is whether, for some tasks, malfare minimization is harder than risk minimization. Theorem 6.4 answers this question in the negative for *sample complexity*, under appropriate conditions on the loss function, as does theorem 5.6 under realizability. The remaining cases are left open, though section 7 at least shows that many conditions sufficient for PAC-learnability are also sufficient for FPAC-learnability.

We are optimistic that our FPAC-learning definitions will motivate the community to further pursue the deep connections between various PAC and FPAC learning settings, as well as promote cross-pollination between computational learning theory and fair machine learning research. We believe that deeper inquiry into these questions will lead to both a better understanding of what is and is not FPAC-learnable, as well as more practical and efficient reductions and FPAC-learning algorithms.

References

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.

Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. Cambridge University Press, 2009.

Ahmed Ashraf, Shehroz Khan, Nikhil Bhagwat, Mallar Chakravarty, and Babak Taati. Learning to unlearn: Building immunity to dataset bias in medical imaging studies. *arXiv preprint arXiv:1812.01716*, 2018.

George Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962.

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

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.

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.

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.

Cyrus Cousins and Matteo Riondato. Cadet: interpretable parametric conditional density estimation with decision trees and forests. *Machine Learning*, 108(8):1613–1634, 2019.

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

Gerard Debreu. Topological methods in cardinal utility theory. Technical report, Cowles Foundation for Research in Economics, Yale University, 1959.

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

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.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

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.

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.

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 fuer Informatik, 2017.

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.

Pascal Massart. Some applications of concentration inequalities to statistics. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 9, pages 245–303, 2000.

Colin McDiarmid. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

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

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

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

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

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. *arXiv preprint arXiv:2003.06740*, 2020.

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

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

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

Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, 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.

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.

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

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

A Appendix: A Compendium of Missing Proofs

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

A.1 Welfare and Malfare

We now show theorem 3.7.

Theorem 3.7 (Properties of the Power-Mean). *Suppose $\mathcal{S}, \boldsymbol{\varepsilon}$ are loss values $\Omega \to \mathbb{R}_{0+}$, and $\boldsymbol{w}$ is a probability measures over some space Ω. Then*

1. *Monotonicity:* $\mathrm{M}_p(\mathcal{S}; \boldsymbol{w})$ *is weakly-monotonically-increasing in p, and strictly if $\mathcal{S}$ attains distinct $a, b \in \mathbb{R}$ with nonnegligible probability.*
2. *Subadditivity:* $\forall p \geq 1 : \mathrm{M}_p(\mathcal{S} + \boldsymbol{\varepsilon}; \boldsymbol{w}) \leq \mathrm{M}_p(\mathcal{S}; \boldsymbol{w}) + \mathrm{M}_p(\boldsymbol{\varepsilon}; \boldsymbol{w})$.
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 convex in $\mathcal{S}$ for $p \in [1, \infty]$ and concave for $p \in [-\infty, 1]$.*

Proof. We omit proof of item 1, as this is a standard property of power-means (generally termed the *power mean inequality* (Bullen, 2013, chapter 3)).

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

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

We now show item 3 First take $\boldsymbol{\varepsilon} \doteq \mathcal{S} - \mathcal{S}'$. Now consider

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

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

$$\mathrm{M}_1(\mathcal{S}; \boldsymbol{w}) = \mathop{\mathbb{E}}_{\omega \sim \boldsymbol{w}}[\mathcal{S}(\omega)] = \underbrace{\mathop{\mathbb{E}}_{\omega \sim \boldsymbol{w}}[\sqrt[p]{\mathcal{S}^p(\omega)}] \leq \sqrt[p]{\mathop{\mathbb{E}}_{\omega \sim \boldsymbol{w}}[\mathcal{S}^p(\omega)]}}_{\textsc{Definition of Convexity}} = \mathrm{M}_p(\mathcal{S}; \boldsymbol{w}) \enspace ,$$

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

$$\mathrm{M}_1(\mathcal{S}; \boldsymbol{w}) = \mathop{\mathbb{E}}_{\omega \sim \boldsymbol{w}}[\mathcal{S}(\omega)] = \underbrace{\mathop{\mathbb{E}}_{\omega \sim \boldsymbol{w}}[\sqrt[p]{\mathcal{S}^p(\omega)}] \geq \sqrt[p]{\mathop{\mathbb{E}}_{\omega \sim \boldsymbol{w}}[\mathcal{S}^p(\omega)]}}_{\textsc{Definition of Concavity}} = \mathrm{M}_p(\mathcal{S}; \boldsymbol{w}) \enspace .$$

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

We now show theorem 3.8.

Theorem 3.8 (Population Mean Properties). *Suppose population-mean function $\mathrm{M}(\mathcal{S}; \boldsymbol{w})$. If $\mathrm{M}(\cdot; \cdot)$ satisfies (subsets of) the population-mean axioms (see definition 3.4), we have that $\mathrm{M}(\cdot; \cdot)$ exhibits the following properties. For each, assume arbitrary sentiment-value function $\mathcal{S} : \Omega \to \mathbb{R}_{0+}$ and weights measure $\boldsymbol{w}$ over Ω. The following then hold.*

1. Identity*: Axioms 6 & 7 imply* $\mathrm{M}(\omega \mapsto \alpha; \boldsymbol{w}) = \alpha$.

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

$$\mathrm{M}(\mathcal{S}; \boldsymbol{w}) = F\left(\int_{\boldsymbol{w}} f_p(\mathcal{S}(\omega))\,\mathrm{d}(\omega)\right) = F\left(\mathop{\mathbb{E}}_{\omega \sim \boldsymbol{w}}\big[f_p(\mathcal{S}(\omega))\big]\right) \ , \ \textit{with} \begin{cases} p = 0 & f_0(x) \doteq \ln(x) \\ p \neq 0 & f_p(x) \doteq \mathrm{sgn}(p)x^p \end{cases} .$$

3. *Axioms 1-7 imply* $F(x) = f_p^{-1}(x)$, *thus* $\mathrm{M}(\mathcal{S}; \boldsymbol{w}) = \mathrm{M}_p(\mathcal{S}; \boldsymbol{w})$.
4. *Axioms 1-5 and 8 imply* $p \in (-\infty, 1]$.
5. *Axioms 1-5 and 9 imply* $p \in [1, \infty)$.

Proof. 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 (Debreu, 1959; Gorman, 1968), extended by continuity and measurability of $\mathcal{S}$ to the weighted case.

We now show item 3. This result is essentially a corollary of item 2, hence the dependence on axioms 1-4. Suppose $\mathcal{S}(\cdot) = 1$. By item 1, for all $p \neq 0$, we have

$$\alpha = \alpha\mathrm{M}(\mathcal{S}; \boldsymbol{w}) = \mathrm{M}(\alpha\mathcal{S}; \boldsymbol{w}) = F\left(\mathop{\mathbb{E}}_{\omega \sim \boldsymbol{w}}\big[f_p(\alpha\mathcal{S}(\omega))\big]\right) = F\left(\mathop{\mathbb{E}}_{\omega \sim \boldsymbol{w}}\big[f_p(\alpha)\big]\right) = F\big(\mathrm{sgn}(p)\alpha^p\big) \ .$$

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 $p = 0$ gets us

$$\alpha = \alpha\mathrm{M}(\mathcal{S}; \boldsymbol{w}) = F\left(\mathop{\mathbb{E}}_{\omega \sim \boldsymbol{w}}\big[\ln(\alpha\mathcal{S}(\omega))\big]\right) = F(\ln a) \ ,$$

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 4 and 5. These properties follow directly from 2, wherein f_p are defined, and Jensen's inequality. $\square$

We now show corollary 4.2.

Corollary 4.2 (Statistical Estimation with Hoeffding and Bennett Bounds)**.** *Suppose fair power-mean malfare* $\mathrm{M}(\cdot;\cdot)$ *(i.e.,* $p \geq 1$*),* loss function $\ell : \mathcal{X} \to [0, r]$, $\mathcal{S} \in [0, r]^g$ *s.t.* $\mathcal{S}_i = \mathbb{E}_{\mathcal{D}_i}[\ell]$, *samples* $\boldsymbol{x}_i \sim \mathcal{D}_i^m$, *and take* $\hat{\mathcal{S}}_i \doteq \frac{1}{m}\sum_{j=1}^m \ell(\boldsymbol{x}_{i,j})$. *Then with probability at least* $1 - \delta$ *over choice of* $\boldsymbol{x}$*,*

$$\Big|\mathrm{M}_p(\mathcal{S}; \boldsymbol{w}) - \mathrm{M}_p(\hat{\mathcal{S}}; \boldsymbol{w})\Big| \leq r\sqrt{\frac{\ln \frac{2g}{\delta}}{2m}} \ .$$

Alternatively, again with probability at least $1 - \delta$ *over choice of* $\boldsymbol{x}$*, we have*

$$\Big|\mathrm{M}_p(\mathcal{S}; \boldsymbol{w}) - \mathrm{M}_p(\hat{\mathcal{S}}; \boldsymbol{w})\Big| \leq \frac{r \ln \frac{2g}{\delta}}{3m} + \max_{i \in 1,\ldots,g} \sqrt{\frac{2\,\mathbb{V}_{\mathcal{D}_i}[\ell] \ln \frac{2g}{\delta}}{m}} \ .$$

Proof. This result is a corollary of lemma 4.1, applied to $\boldsymbol{\varepsilon}$, where we note that for $p \geq 1$, by theorem 3.7 item 3 (contraction) it holds that

$$\mathrm{M}_p(\hat{\mathcal{S}} + \boldsymbol{\varepsilon}; \boldsymbol{w}) \leq \mathrm{M}_p(\hat{\mathcal{S}}; \boldsymbol{w}) + \|\boldsymbol{\varepsilon}\|_\infty \ \& \ \mathrm{M}_p(\boldsymbol{0} \vee (\hat{\mathcal{S}} - \boldsymbol{\varepsilon}); \boldsymbol{w}) \leq \mathrm{M}_p(\hat{\mathcal{S}}; \boldsymbol{w}) - \|\boldsymbol{\varepsilon}\|_\infty \ .$$

Now, for the first bound, note that we take $\boldsymbol{\varepsilon}_i \doteq r\sqrt{\frac{\ln \frac{2g}{\delta}}{2m}}$, and by Hoeffding's inequality and the union bound, for $\Omega = \{1, \ldots, n\}$, we have $\forall \omega : \mathcal{S}'(\omega) - \boldsymbol{\varepsilon}(\omega) \leq \mathcal{S}(\omega) \leq \mathcal{S}'(\omega) + \boldsymbol{\varepsilon}(\omega)$ with probability at least $1 - \delta$. The result then follows via the *power-mean contraction* (theorem 3.7 item 3) property.

Similarly, for the second bound, note that we take $\boldsymbol{\varepsilon}_i \doteq \frac{r \ln \frac{2g}{\delta}}{3m} + \sqrt{\frac{2\,\mathbb{V}_{\mathcal{D}_i}[\ell] \ln \frac{2g}{\delta}}{m}}$, which this time follows via Bennett's inequality and the union bound. Now, we again apply lemma 4.1, noting that $\mathrm{M}(\boldsymbol{\varepsilon}) \leq \mathrm{M}_\infty(\boldsymbol{\varepsilon}) = \|\boldsymbol{\varepsilon}\|_\infty$ (by *power-mean monotonicity*, theorem 3.7 item 1), and the rest follows as in the Hoeffding case. $\square$

A.2 Efficient FPAC-Learning

We now show theorem 7.1.

Theorem 7.1 (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 $\forall x \in \mathcal{X}_d, \theta \in \Theta_d$, $h(x;\theta)$ 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)

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

fair-PAC-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 $(\mathcal{H}, \ell) \in \mathrm{FPAC}^{\mathrm{Agn}}_{\mathrm{Poly}}$.*

Proof. We now show that this 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 fair-PAC-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 subgradient-update steps, with appropriate step size, on a sufficiently large training set, to yield the appropriate guarantees, and that each step of the subgradient method, of which there are polynomially many, itself requires polynomial time.

First, note that by theorem 3.8 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 Θ_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 quantity evaluated on each of g training sets. By theorem 3.7 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 3.7 item 3), and thus by Lipschitz composition, we have Lipschitz property

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

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 $(\dim(\Theta_d) + 1)m$ 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 subgradient method produces $\tilde{\theta}$ approximating the empirically-optimal $\hat{\theta}$ such that (see Shor, 2012)

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

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

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

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

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

As each iteration requires $m \cdot \mathrm{Poly}(d) \subseteq \mathrm{Poly}(\frac{1}{\varepsilon}, \frac{1}{\delta}, d, g)$ time, the 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 $\frac{\varepsilon}{3}$-$\frac{\delta}{g}$ 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

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

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

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

We may thus conclude that $\mathcal{A}$ fair-PAC 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 $(\mathcal{H}, \ell) \in \mathrm{PAC}^{\mathrm{Agn}}_{\mathrm{Poly}}$. ☐

We now work towards proof of theorem 7.2. We begin with a technical lemma deriving relevant properties of the cover employed in the main result.

Lemma A.1 (Group Cover Properties)**.** *Suppose loss function ℓ of bounded codomain (i.e., $\ell : (\mathcal{Y} \times \mathcal{Y}) \to [0, \|\ell\|_\infty]$), hypothesis class $\mathcal{H} \subseteq \mathcal{X} \to \mathcal{Y}$, and per-group samples $\boldsymbol{z}_{1:g,1:m} \in (\mathcal{X} \times \mathcal{Y})^{g \times m}$, letting $\bigcirc_{i=1}^{g} \boldsymbol{z}_i$ denote their* concatenation. *Now define*

$$\hat{\mathcal{C}}_{\cup(1:g)} \doteq \bigcup_{i=1}^{g} \hat{\mathcal{C}}\left(\ell \circ \mathcal{H}, \boldsymbol{z}_i, \gamma\right) \quad \& \quad \hat{\mathcal{C}}_{\circ(1:g)} \doteq \hat{\mathcal{C}}\left(\ell \circ \mathcal{H}, \bigcirc_{i=1}^{g} \boldsymbol{z}_i, \tfrac{\gamma}{\sqrt{g}}\right) \ .$$

Then, letting $\hat{\mathcal{C}}_?$ refer generically to either $\hat{\mathcal{C}}_{\cup(1:g)}$ or $\hat{\mathcal{C}}_{\circ(1:g)}$, the following hold.

1. *If $\hat{\mathcal{C}}_?$ is of* minimal cardinality*, then*

$$\left|\hat{\mathcal{C}}_{\cup(1:g)}\right| \leq g\mathcal{N}(\ell \circ \mathcal{H}, m, \gamma) \quad \& \quad \left|\hat{\mathcal{C}}_{\circ(1:g)}\right| \leq \mathcal{N}(\ell \circ \mathcal{H}, gm, \tfrac{\gamma}{\sqrt{g}}) \ .$$

2. $\sup_{\mathcal{D} \text{ over } \mathcal{X}\times\mathcal{Y}} \mathfrak{R}_m(\ell \circ \mathcal{H}, \mathcal{D}) \leq \inf_{\gamma \geq 0} \gamma + \|\ell\|_\infty \sqrt{\frac{\ln \mathcal{N}(\ell \circ \mathcal{H}, \gamma)}{2m}}$.

3. *Suppose* $\ln \mathcal{N}(\ell \circ \mathcal{H}, \gamma) \in \mathrm{Poly}(\frac{1}{\gamma})$. *Then the uniform-convergence sample-complexity of* $\ell \circ \mathcal{H}$ *over* g *groups obeys*

$$\begin{aligned}\mathrm{m}_{\mathrm{UC}}(\varepsilon, \delta, g) &\doteq \operatorname{argmin}\left\{ m \;\middle|\; \mathbb{P}\left(\max_{i \in 1,\dots,g} \sup_{h \in \mathcal{H}} \left| \mathop{\mathbb{E}}_{\mathcal{D}_i}[\ell \circ h] - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{z}_i \sim \mathcal{D}_i^m}[\ell \circ h] \right| > \varepsilon \right) \leq \delta \right\} \\ &\leq \frac{8\|\ell\|_\infty^2 \ln\left(\sqrt[4]{\frac{2g}{\delta}} \mathcal{N}(\ell \circ \mathcal{H}, \frac{\varepsilon}{4}) \right)}{\varepsilon^2} \\ &\in \mathbf{O}\left(\frac{\ln \frac{g\mathcal{N}(\ell \circ \mathcal{H}, \varepsilon)}{\delta}}{\varepsilon^2} \right) \subset \mathrm{Poly}\left(\frac{1}{\varepsilon}, \exp \frac{1}{\delta}, \exp g \right) \enspace .\end{aligned}$$

4. *For the sample* $\boldsymbol{z}_i$ *associated with each group* $i \in 1, \dots, g$, $\hat{\mathcal{C}}_?$ *is a* γ*-uniform-approximation of* empirical risk $\hat{\mathrm{R}}(h; \ell, \boldsymbol{z}_i)$, *and a* γ-ℓ_2 *cover of the* loss family $\ell \circ \mathcal{H}$, *as*

$$\max_{i \in 1,\dots,g} \min_{h_\gamma \in \hat{\mathcal{C}}_?} \left| \hat{\mathrm{R}}(h; \ell, \boldsymbol{z}_i) - \hat{\mathrm{R}}(h_\gamma; \ell, \boldsymbol{z}_i) \right| \leq \max_{i \in 1,\dots,g} \min_{h_\gamma \in \hat{\mathcal{C}}_?} \sqrt{\frac{1}{m} \sum_{j=1}^m \big((\ell \circ h)(\boldsymbol{z}_{i,j}) - (\ell \circ h_\gamma)(\boldsymbol{z}_{i,j}) \big)^2} \leq \gamma \enspace .$$

5. $\hat{\mathcal{C}}_{\circ(1:g)}$, *but not necessarily* $\hat{\mathcal{C}}_{\cup(1:g)}$, simultaneously *(across all groups)* γ*-uniformly-approximates* empirical risk, *and is a* γ-ℓ_2 *cover of the* loss family $\ell \circ \mathcal{H}$, *as*

$$\min_{h_\gamma \in \hat{\mathcal{C}}_{\circ(1:g)}} \max_{i \in 1,\dots,g} \left| \hat{\mathrm{R}}(h; \ell, \boldsymbol{z}_i) - \hat{\mathrm{R}}(h_\gamma; \ell, \boldsymbol{z}_i) \right| \leq \min_{h_\gamma \in \hat{\mathcal{C}}_{\circ(1:g)}} \max_{i \in 1,\dots,g} \sqrt{\frac{1}{m} \sum_{j=1}^m \big((\ell \circ h)(\boldsymbol{z}_{i,j}) - (\ell \circ h_\gamma)(\boldsymbol{z}_{i,j}) \big)^2} \leq \gamma \enspace .$$

Proof. We first show items 1 to 3, followed by a key intermediary relating risk values and ℓ_2 distances, and close by showing items 4 and 5.

We begin with item 1. Both bounds follow directly from the definition of uniform covering numbers.

We now show item 2. This result follows via a standard sequence of operations over the Rademacher average. In particular, observe

$$\begin{aligned}\sup_{\mathcal{D} \text{ over } \mathcal{X}\times\mathcal{Y}} \mathfrak{R}_m(\ell \circ \mathcal{H}, \mathcal{D}) &= \sup_{\mathcal{D} \text{ over } \mathcal{X}\times\mathcal{Y}} \mathop{\mathbb{E}}_{\boldsymbol{z} \sim \mathcal{D}^m}\left[\hat{\mathfrak{R}}_m(\ell \circ \mathcal{H}, \boldsymbol{z}) \right] && \textsc{Definition of } \mathfrak{R} \\ &\leq \sup_{\mathcal{D} \text{ over } \mathcal{X}\times\mathcal{Y}} \mathop{\mathbb{E}}_{\boldsymbol{z} \sim \mathcal{D}^m}\left[\inf_{\gamma \geq 0} \gamma + \hat{\mathfrak{R}}_m(\mathcal{C}^*(\ell \circ \mathcal{H}, \boldsymbol{z}, \gamma), \boldsymbol{z}) \right] && \textsc{Discretization} \\ &\leq \inf_{\gamma \geq 0} \gamma + \sup_{\mathcal{D} \text{ over } \mathcal{X}\times\mathcal{Y}} \mathop{\mathbb{E}}_{\boldsymbol{z} \sim \mathcal{D}^m}\left[\|\ell\|_\infty \sqrt{\frac{\ln|\mathcal{C}^*(\ell \circ \mathcal{H}, \boldsymbol{z}, \gamma)|}{2m}} \right] && \textsc{Massart's Inequality} \\ &\leq \inf_{\gamma \geq 0} \gamma + \|\ell\|_\infty \sqrt{\frac{\ln \mathcal{N}(\ell \circ \mathcal{H}, \gamma)}{2m}} \enspace , && \textsc{Definition of } \mathcal{N}\end{aligned}$$

where the Massart's Inequality step follows via *Massart's finite class inequality* (Massart, 2000, lemma 1), and the Discretization step via *Dudley's discretization argument.*

We now show item 3. By the *symmetrization inequality*, and a 2-tailed application of McDiarmid's bounded difference inequality (McDiarmid, 1989), where changing any $\boldsymbol{z}_{i,j}$ has bounded difference $\frac{\|\ell\|_\infty}{m}$, we have that

$$\forall i: \; \mathbb{P}\left(\sup_{h \in \mathcal{H}} \left| \mathop{\mathbb{E}}_{\mathcal{D}_i}[\ell \circ h] - \mathop{\hat{\mathbb{E}}}_{\boldsymbol{z}_i \sim \mathcal{D}_i^m}[\ell \circ h] \right| > 2\mathfrak{R}_m(\ell \circ \mathcal{H}, \mathcal{D}_i) + \|\ell\|_\infty \sqrt{\frac{\ln \frac{2}{\delta}}{2m}} \right) \leq \delta$$

thus by union bound over g groups, we have

$$\mathbb{P}\left(\max_{i\in 1,\ldots,g}\sup_{h\in\mathcal{H}}\left|\mathop{\mathbb{E}}_{\mathcal{D}_i}[\ell\circ h]-\mathop{\hat{\mathbb{E}}}_{\boldsymbol{z}_i\sim\mathcal{D}_i^m}[\ell\circ h]\right|>\sup_{\mathcal{D}\text{ over }\mathcal{X}\times\mathcal{Y}}2\mathfrak{R}_m(\ell\circ\mathcal{H},\mathcal{D})+\|\ell\|_\infty\sqrt{\frac{\ln\frac{2g}{\delta}}{2m}}\right)\leq\delta\ .$$

Now, let *estimation error* bound $\epsilon_{\text{est}}\doteq\sup_{\mathcal{D}\text{ over }\mathcal{X}\times\mathcal{Y}}2\mathfrak{R}_m(\ell\circ\mathcal{H},\mathcal{D})+\|\ell\|_\infty\sqrt{\frac{\ln\frac{2g}{\delta}}{2m}}$, and observe that via item 2,

$$\epsilon_{\text{est}}\leq\inf_{\gamma\geq 0}2\gamma+2\|\ell\|_\infty\sqrt{\frac{\ln\mathcal{N}(\ell\circ\mathcal{H},\gamma)}{2m}}+\|\ell\|_\infty\sqrt{\frac{\ln\frac{2g}{\delta}}{2m}}\ .$$

From here, we solve for an upper-bound on sample-size m to get

$$\begin{aligned}\mathrm{m}_{\mathrm{UC}}(\varepsilon,\delta,g)&\leq\inf_{\gamma\geq 0}\frac{4\|\ell\|_\infty^2\ln\mathcal{N}(\ell\circ\mathcal{H},\gamma)+\|\ell\|_\infty^2\ln\frac{2g}{\delta}}{2(\varepsilon-2\gamma)^2}=\frac{2\|\ell\|_\infty^2\ln\left(\sqrt[4]{\frac{2g}{\delta}}\mathcal{N}(\ell\circ\mathcal{H},\gamma)\right)}{(\varepsilon-2\gamma)^2}\\&\leq\frac{8\|\ell\|_\infty^2\ln\left(\sqrt[4]{\frac{2g}{\delta}}\mathcal{N}(\ell\circ\mathcal{H},\frac{\varepsilon}{4})\right)}{\varepsilon^2}&&\textsc{Set }\gamma=\frac{\varepsilon}{4}\\&\in\mathbf{O}\left(\frac{\ln\frac{g\mathcal{N}(\ell\circ\mathcal{H},\varepsilon)}{\delta}}{\varepsilon^2}\right)\subset\mathrm{Poly}\left(\frac{1}{\varepsilon},\exp\frac{1}{\delta},\exp g\right)\ .&&\mathcal{N}(\ell\circ\mathcal{H},\varepsilon)\in\mathrm{Poly}\frac{1}{\varepsilon}\end{aligned}$$

We now show an intermediary which immediately implies the left inequalities of both items 4 and 5. In particular, we may relate these empirical risk gaps to (size-normalized) ℓ_2 distance, as (for each i) we have $\forall h\in\mathcal{H},h_\gamma\in\hat{\mathcal{C}}_?$ that

$$\left|\hat{\mathrm{R}}(h;\ell,\boldsymbol{z}_i)-\hat{\mathrm{R}}(h_\gamma;\ell,\boldsymbol{z}_i)\right|\leq\frac{1}{m}\sum_{j=1}^m\left|(\ell\circ h)(\boldsymbol{z}_{i,j})-(\ell\circ h_\gamma)(\boldsymbol{z}_{i,j})\right|\leq\sqrt{\frac{1}{m}\sum_{j=1}^m\big((\ell\circ h)(\boldsymbol{z}_{i,j})-(\ell\circ h_\gamma)(\boldsymbol{z}_{i,j})\big)^2}\ .$$

Here the last inequality holds since we divide m *inside the* $\sqrt{\cdot}$. The opposite inequality holds for standard ℓ_1 and Euclidean distance, where m is not divided, essentially because the ℓ_1 and ℓ_2 distances differ by up to a factor $\sqrt{m}$, but this form may be familiar as the relationship between the *mean* and *root mean square* errors. The unconvinced reader may note that this size-normalized ℓ_2 distance is in fact the (unweighted) $p=2$ power-mean, and thus this step follows via theorem 3.7 item 1.

We now show the right inequality of item 4. Note that for the case of $\hat{\mathcal{C}}_{\cup(1:g)}$, the result is almost tautological, as it holds per group by the union-based construction of $\hat{\mathcal{C}}_{\cup(1:g)}$. The case of $\hat{\mathcal{C}}_{\circ(1:g)}$ is more subtle, but we defer its proof to the final item, as it then follows as an immediate consequence of the *max-min inequality*, i.e., $\forall h\in\mathcal{H}$,

$$\max_{i\in 1,\ldots,g}\min_{h_\gamma\in\hat{\mathcal{C}}_{\circ(1:g)}}\sqrt{\frac{1}{m}\sum_{j=1}^m\big((\ell\circ h)(\boldsymbol{z}_{i,j})-(\ell\circ h_\gamma)(\boldsymbol{z}_{i,j})\big)^2}\leq\min_{h_\gamma\in\hat{\mathcal{C}}_{\circ(1:g)}}\max_{i\in 1,\ldots,g}\sqrt{\frac{1}{m}\sum_{j=1}^m\big((\ell\circ h)(\boldsymbol{z}_{i,j})-(\ell\circ h_\gamma)(\boldsymbol{z}_{i,j})\big)^2}\ .$$

We now show item 5. Suppose (by way of contradiction) that there exists some $h\in\mathcal{H}$ such that

$$\min_{h_\gamma\in\hat{\mathcal{C}}_{\circ(1:g)}}\max_{i\in 1,\ldots,g}\sqrt{\frac{1}{m}\sum_{j=1}^m\big((\ell\circ h)(\boldsymbol{z}_{i,j})-(\ell\circ h_\gamma)(\boldsymbol{z}_{i,j})\big)^2}>\gamma\ .$$

One then need only consider the summands associated with a maximal i to observe that this implies

$$\min_{h_\gamma \in \hat{\mathcal{C}}_{\circ(1:g)}} \sqrt{\frac{1}{mg}\sum_{i=1}^{g}\sum_{j=1}^{m}\Big((\ell \circ h)(\boldsymbol{z}_{i,j}) - (\ell \circ h_\gamma)(\boldsymbol{z}_{i,j})\Big)^2} > \frac{\gamma}{\sqrt{g}} \enspace ,$$

thus $\hat{\mathcal{C}}_{\circ(1:g)}$ is not a $\frac{\gamma}{\sqrt{g}}$-ℓ_2 cover of $\bigcirc_{i=1}^{g} \boldsymbol{z}_i$, which contradicts its very definition. We thus conclude

$$\forall h \in \mathcal{H} : \min_{h_\gamma \in \hat{\mathcal{C}}_{\circ(1:g)}} \max_{i \in 1,\dots,g} \sqrt{\frac{1}{m}\sum_{j=1}^{m}\Big((\ell \circ h)(\boldsymbol{z}_{i,j}) - (\ell \circ h_\gamma)(\boldsymbol{z}_{i,j})\Big)^2} \leq \gamma \enspace .$$

□

With lemma A.1 in hand, we are now ready to show theorem 7.2.

Theorem 7.2 (Efficient FPAC-Learning by Covering)**.** *Suppose loss function ℓ of bounded codomain (i.e., $\|\ell\|_\infty$ is bounded), and hypothesis class sequence $\mathcal{H}$, s.t. $\forall m, d \in \mathbb{N}$, $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^m$, there exist*

1. *a γ-ℓ_2 cover $\mathcal{C}^*(\ell \circ \mathcal{H}_d, \boldsymbol{z}, \gamma)$, where $\left|\mathcal{C}^*(\ell \circ \mathcal{H}_d, \boldsymbol{z}, \gamma)\right| \leq \mathcal{N}(\ell \circ \mathcal{H}_d, \gamma) \in \mathrm{Poly}(\frac{1}{\gamma}, d)$; and*
2. *an algorithm to enumerate a γ-ℓ_2 cover $\hat{\mathcal{C}}(\ell \circ \mathcal{H}_d, \boldsymbol{z}, \gamma)$ of size $\mathrm{Poly}\,\mathcal{N}(\ell \circ \mathcal{H}_d, m, \gamma)$ in $\mathrm{Poly}(m, \frac{1}{\gamma}, d)$ time.*

Then $(\ell, \mathcal{H}) \in \mathrm{FPAC}_{\mathrm{Poly}}$, where (1) is sufficient to show that $(\ell, \mathcal{H})$ is fair-PAC learnable with polynomial sample complexity*, and (2) is required only to show polynomial* training time complexity.

Proof. We now constructively show the existence of a fair-PAC-learner $\mathcal{A}$ for $(\ell, \mathcal{H})$ over domain $\mathcal{X}$ and codomain $\mathcal{Y}$. As in theorem 7.1, we first note that by theorem 3.8 items 3 and 5, under the conditions of FPAC learning, this reduces to showing that we can learn any malfare concept $\mathrm{M}_p(\cdot;\cdot)$ that is a p-power mean with $p \geq 1$.

We first assume a training sample $\boldsymbol{z}_{1:g,1:m} \sim \mathcal{D}_1^m \times \cdots \times \mathcal{D}_g^m$, i.e., a collection of m draws from each of the g groups. In particular, we shall select m to guarantee that the *estimation error* for the malfare does not exceed $\frac{\varepsilon}{3}$ with probability at least $1 - \delta$, i.e., we require that with said probability,

$$\epsilon_{\mathrm{est}} \doteq \left|\mathrm{M}_p(i \mapsto \hat{\mathrm{R}}(h; \ell, \boldsymbol{z}_i); \boldsymbol{w}) - \mathrm{M}_p(i \mapsto \mathrm{R}(h; \ell, \mathcal{D}_i); \boldsymbol{w})\right| \leq \frac{\varepsilon}{3} \enspace .$$

Now, note that by theorem 3.7 item 3 (contraction), we have

$$\left|\mathrm{M}_p(i \mapsto \hat{\mathrm{R}}(h; \ell, \boldsymbol{z}_i); \boldsymbol{w}) - \mathrm{M}_p(i \mapsto \mathrm{R}(h; \ell, \mathcal{D}_i); \boldsymbol{w})\right| \leq \max_{i \in 1,\dots,g} \sup_{h \in \mathcal{H}_d} \left|\mathrm{R}(h; \ell, \mathcal{D}_i); \boldsymbol{w}) - \hat{\mathrm{R}}(h; \ell, \boldsymbol{z}_i); \boldsymbol{w})\right| \enspace ,$$

and by lemma A.1 item 3, a sample of size

$$m = \left\lceil \frac{81\|\ell\|_\infty^2 \ln\left(\sqrt[4]{\frac{2g}{\delta}}\mathcal{N}(\ell \circ \mathcal{H}, \frac{\varepsilon}{12})\right)}{\varepsilon^2} \right\rceil \in \mathbf{O}\left(\frac{\ln \frac{g\mathcal{N}(\ell \circ \mathcal{H}, \varepsilon)}{\delta}}{\varepsilon^2}\right) \subset \mathrm{Poly}\left(\frac{1}{\varepsilon}, \exp\frac{1}{\delta}, \exp g\right)$$

suffices to ensure that

$$\mathbb{P}\left(\max_{i \in 1,\dots,g} \sup_{h \in \mathcal{H}_d} \left|\mathrm{R}(h; \ell, \mathcal{D}_i); \boldsymbol{w}) - \hat{\mathrm{R}}(h; \ell, \boldsymbol{z}_i); \boldsymbol{w})\right| > \frac{\varepsilon}{3}\right) \leq \delta \enspace ,$$

thus guaranteeing the stated estimation error bound.

With our sample size and *estimation error* guarantee, we now define the *learning algorithm* and bound its *optimization error*. Take *cover precision* $\gamma \doteq \frac{\varepsilon}{3\sqrt{g}}$. By assumption, for each $d \in \mathbb{N}$, we may enumerate a γ-cover $\hat{\mathcal{C}}(\ell \circ \mathcal{H}_d, \bigcirc_{i=1}^{g} \boldsymbol{z}_i, \gamma)$, where $\bigcirc_{i=1}^{g} \boldsymbol{z}_i$ denotes the *concatenation* of each $\boldsymbol{z}_i$, in $\mathrm{Poly}(gm, \frac{1}{\gamma}, d) = \mathrm{Poly}(\frac{1}{\varepsilon}, \frac{1}{\delta}, d, g)$ time. For the remainder of this proof, we refer to this cover as $\hat{\mathcal{C}}$.

Now, we take the learning algorithm to be *empirical malfare minimization* over $\hat{\mathcal{C}}$. Let

$$\hat{h} \doteq \operatorname*{argmin}_{h_\gamma \in \hat{\mathcal{C}}} \mathrm{M}_p(i \mapsto \hat{\mathrm{R}}(h_\gamma; \ell, \boldsymbol{z}_i); \boldsymbol{w}) \quad \& \quad \tilde{h} \doteq \operatorname*{argmin}_{h \in \mathcal{H}_d} \mathrm{M}_p(i \mapsto \hat{\mathrm{R}}(h; \ell, \boldsymbol{z}_i); \boldsymbol{w}) \ ,$$

where ties may be broken arbitrarily. Note that via standard covering properties, that the *optimization error* is bounded as

$$\begin{aligned}
\varepsilon_{\mathrm{opt}} &\doteq \mathrm{M}_p\big(i \mapsto \hat{\mathrm{R}}(\hat{h}; \ell, \boldsymbol{z}_i); \boldsymbol{w}\big) - \mathrm{M}_p\big(i \mapsto \hat{\mathrm{R}}(\tilde{h}; \ell, \boldsymbol{z}_i); \boldsymbol{w}\big) && \textsc{Definition} \\
&= \sup_{h \in \mathcal{H}_d} \min_{h_\gamma \in \hat{\mathcal{C}}} \mathrm{M}_p\big(i \mapsto \hat{\mathrm{R}}(h_\gamma; \ell, \boldsymbol{z}_i); \boldsymbol{w}\big) - \mathrm{M}_p\big(i \mapsto \hat{\mathrm{R}}(h; \ell, \boldsymbol{z}_i); \boldsymbol{w}\big) && \begin{array}{r}\textsc{Properties of Suprema} \\ \textsc{Definition of } \hat{h}, \tilde{h}\end{array} \\
&\leq \sup_{h \in \mathcal{H}_d} \min_{h_\gamma \in \hat{\mathcal{C}}} \max_{i \in \{1, \dots, g\}} \left| \hat{\mathrm{R}}(h_\gamma; \ell, \boldsymbol{z}_i) - \hat{\mathrm{R}}(h; \ell, \boldsymbol{z}_i) \right| && \textsc{Item 3 (Contraction)} \\
&\leq \sqrt{g}\gamma = \frac{\varepsilon}{3} \ . && \textsc{Lemma A.1 Item 5}
\end{aligned}$$

This controls for *optimization error* between the true and approximate EMM solutions $\tilde{h}$ and $\hat{h}$.

We now combine the optimization and estimation error inequalities, letting

$$h^* \doteq \operatorname*{argmin}_{h \in \mathcal{H}_d} \mathrm{M}_p(i \mapsto \mathrm{R}(h; \ell, \mathcal{D}_i); \boldsymbol{w}) \ ,$$

denote the true *malfare optimal* solution, over *distributions* rather than *samples*, breaking ties arbitrarily. We then derive

$$\begin{aligned}
&\mathrm{M}_p(i \mapsto \mathrm{R}(h^*; \ell, \mathcal{D}_i); \boldsymbol{w}) - \mathrm{M}_p(i \mapsto \mathrm{R}(\hat{h}; \ell, \mathcal{D}_i); \boldsymbol{w}) \\
&\quad = \Big(\mathrm{M}_p(i \mapsto \mathrm{R}(h^*; \ell, \mathcal{D}_i); \boldsymbol{w}) - \mathrm{M}_p(i \mapsto \hat{\mathrm{R}}(h^*; \ell, \boldsymbol{z}_i); \boldsymbol{w})\Big) && \leq \epsilon_{\mathrm{est}} \\
&\quad\quad + \Big(\mathrm{M}_p(i \mapsto \hat{\mathrm{R}}(h^*; \ell, \boldsymbol{z}_i); \boldsymbol{w}) - \mathrm{M}_p(i \mapsto \hat{\mathrm{R}}(\tilde{h}; \ell, \boldsymbol{z}_i); \boldsymbol{w})\Big) && \leq 0 \\
&\quad\quad + \Big(\mathrm{M}_p(i \mapsto \hat{\mathrm{R}}(\tilde{h}; \ell, \boldsymbol{z}_i); \boldsymbol{w}) - \mathrm{M}_p(i \mapsto \hat{\mathrm{R}}(\hat{h}; \ell, \boldsymbol{z}_i); \boldsymbol{w})\Big) && \leq \epsilon_{\mathrm{opt}} \\
&\quad\quad + \Big(\mathrm{M}_p(i \mapsto \hat{\mathrm{R}}(\hat{h}; \ell, \boldsymbol{z}_i); \boldsymbol{w}) - \mathrm{M}_p(i \mapsto \mathrm{R}(\hat{h}; \ell, \mathcal{D}_i); \boldsymbol{w})\Big) && \leq \epsilon_{\mathrm{est}} \\
&\quad \leq \epsilon_{\mathrm{opt}} + 2\epsilon_{\mathrm{est}} = \frac{\varepsilon}{3} + \frac{2\varepsilon}{3} = \varepsilon \ . && \textsc{See Above}
\end{aligned}$$

We thus conclude that this algorithm produces an ε-$\mathrm{M}_p(\cdot; \cdot)$ optimal solution with probability at least $1 - \delta$, and furthermore both the sample complexity and time complexity of this algorithm are $\mathrm{Poly}(\frac{1}{\varepsilon}, \frac{1}{\delta}, g, d)$. Hence, as we have constructed a polynomial-time fair-PAC learner for $(\mathcal{H}, \ell)$, we may conclude $(\mathcal{H}, \ell) \in \mathrm{FPAC}_{\mathrm{Poly}}$. $\square$