Revisiting Fair-PAC Learning and the Axioms of Cardinal Welfare

Cyrus Cousins
University of Massachusetts Amherst

Abstract

Cardinal objectives like welfare and malfare have recently enjoyed increased attention in fair machine learning as computationally, statistically, and philosophically sound alternatives to constraint-based methods. Welfare *summarizes utility* over a population of g groups, whereas malfare measures *overall disutility*. Subject to standard axioms, both welfare and malfare functions are restricted to the family of $\boldsymbol{w}$*-weighted* p*-power-means*, i.e., $\mathrm{M}_p(\boldsymbol{u}; \boldsymbol{w}) \doteq \sqrt[p]{\sum_{i=1}^{g} \boldsymbol{w}_i \boldsymbol{u}_i^p}$, with $p \leq 1$ for welfare (utility $\boldsymbol{u}$), or $p \geq 1$ for malfare (disutility $\boldsymbol{u}$). This work revisits said axioms, finding that a weaker basis is sufficient to show the same, and furthermore that strengthening these axioms partition the welfare half of the spectrum (i.e., $p \leq 1$) into a few cases by further limiting p. It is known that $p \geq 1$ power means (i.e., malfare functions) are Lipschitz continuous, and thus *statistically easy* to estimate or learn (i.e., each $\boldsymbol{u}_i$ can be approximated with a *sample estimate*). We show that *all power means* are at least locally Hölder continuous, i.e., $|\mathrm{M}(\boldsymbol{u}; \boldsymbol{w}) - \mathrm{M}(\boldsymbol{u}'; \boldsymbol{w})| \leq \lambda \|\boldsymbol{u} - \boldsymbol{u}'\|^\alpha$ for some constants $\lambda > 0$, $\alpha \in (0, 1]$, and some norm $\|\cdot\|$. Furthermore, λ and $\frac{1}{\alpha}$ are bounded except as $p \to 0$ or $\min_i \boldsymbol{w}_i \to 0$, and via this analysis we bound the sample complexity of *estimating* or *optimizing* welfare functions. This yields a novel concept of *fair-PAC learning*, with dependence on the quantities $\frac{1}{|p|}$ and/or $\frac{1}{\boldsymbol{w}_{\min}}$ (which measure closeness to the challenging $p = 0$ case and inverse minimum group weight, respectively), wherein fair welfare functions are only *polynomially harder* to optimize than fair malfare functions, except when $p \approx 0$ or $\min_i \boldsymbol{w}_i \approx 0$, which may be *exponentially harder*. These challenging cases may be avoided by assuming the appropriate strengthened axioms. In all cases, we show that if a bounded quantity is learnable with finite sample complexity, then so too is the welfare of said quantity. This takes estimating and learning welfare objectives to near-parity with malfare objectives, as although sample complexity may be larger, all such objectives are *uniformly learnable*.

Keywords

Fair Machine Learning ♢ Algorithmic Fairness
Fair PAC Learning ♢ Statistical Learning Theory
Social Planner's Problem ♢ Cardinal Welfare Theory
Weighted Power Means ♢ Hölder Continuity Analysis

1 Introduction

The recent resurgence of cardinal welfare and malfare based methods in group-based fairness settings has led to increased attention as to how to *objectively quantify* fairness. Welfare summarizes utility across a population, and is thus suitable for fair utility-maximization tasks (e.g., bandit problems, reward-based reinforcement learning [Siddique et al., 2020, Cousins et al., 2022a], and recommender systems, as well as utility-based finance and economic settings), whereas malfare measures overall disutility, and is thus directly applicable to fair loss minimization tasks (arguably most machine learning tasks). The promise of statistical and computational efficiency differentiates such approaches from constraint-based fairness methods [Dwork et al., 2012, Zemel et al., 2013], which often yield hard non-convex optimization problems, requiring convex relaxations (potentially sacrificing fairness), as well as inducing statistical quandaries in estimating whether such fairness constraints generalize beyond the training set [Yona and Rothblum, 2018, Thomas et al., 2019]. The axiomatic justification for cardinal welfare and malfare functions also gives them a sense of objectivity, whereas fairness constraints are often intuitively motivated, and at times mutually incompatible [Kleinberg et al., 2017, Friedler et al., 2021]. We find a basis of cardinal welfare axioms that is weaker than the standard basis, and we then propose stronger axioms to further specify such functions, and explore the resulting classes of fair learnability.

It is now well-understood that many problems of unfairness in machine learning are caused at least in part by a lack of training data for relevant groups [Chen et al., 2018a, Mehrabi et al., 2021], e.g., an overrepresentation of images

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206. Copyright 2023 by the author(s).

of whites and males in training data for facial recognition systems [Buolamwini and Gebru, 2018, Cook et al., 2019, Cavazos et al., 2020]. While addressing such data biases is certainly a step in the right direction, we argue that *the objective itself* must be selected to appropriately compromise between the wants and needs of various groups, which may not mutually align.[1] Collecting sample sizes proportional to population frequencies of each group and minimizing empirical risk (average loss) or maximizing empirical utility seems reasonable, but this is implicitly a utilitarian perspective, as this strategy is equivalent to maximizing *empirical weighted utilitarian welfare or malfare* (as a proxy for their expected values), with the weight of each group i proportional to the number of training points from group i. Furthermore, even if one accepts the tenets of utilitarianism (a perfectly valid choice, but one that should be made *consciously*, rather than *by default*), the result is still not necessarily fair, as we risk overfitting to the proportional, but overall small, amount of data collected for smaller groups. Our axiomatic approach to fairness rigorously treats the issue of selecting and optimizing fair objectives, in particular issues of sample complexity and overfitting to fairness, whether one's chosen fairness objective be strictly utilitarian or of a more prioritarian nature. In particular, any given fair objective effectively specifies how *tradeoffs* must be made between groups when *disagreements* arise between groups during training (discussed further in section 4.3), and the utilitarian welfare or malfare is just one of many axiomatically justifiable choices.

Section 3.1 shows that some of the cardinal welfare or malfare axioms of Cousins [2021] can be relaxed or reorganized to form *a piecewise-weaker equivalent basis* (i.e., each of our axioms is no stronger than an existing standard axiom, yet their collective effect is equivalent) that is more elegant and more concise, thus our axiomatization is a more convincing premise upon which to develop a theory of fair learning. Subject to these axioms, the $\boldsymbol{w}$*-weighted* p*-power-mean family* arises as the only axiomatically justified class of fair aggregator functions, however the parameter space of this class is quite large, thus the theory does not uniquely specify an "ideal fairness concept." Many have argued that exact human-desirable fairness concepts cannot be fully specified without unjustifiable assumptions, and that variation in feasible aggregator function concepts reflects variation in human morality and social values [Awad et al., 2018, Schneider and Leland, 2021]. We do not reject this claim, however we do show in section 3.3 that additional axioms can further restrict the family of malfare or welfare functions, although such axioms may be less universal than the standard basis of cardinal welfare or malfare axioms.

[1] For example, a learned spellchecker may encounter both UK English and US English, and the decision to correct, e.g., either "color" or "colour" is a tradeoff that explicitly harms writers of one dialect while benefiting writers of the other.

In particular, we propose a stronger variant of *strict monotonicity*, which precludes both egalitarian and utilitarian welfare, as well as a relaxation of the Pigou-Dalton [Pigou, 1912, Dalton, 1920] transfer principle, an axiom which characterizes fairness by promoting equitable redistribution of (dis)utility, as well as a strengthening of it, which specifically characterizes utilitarianism as *neutral*, rather than equitable or fair. We also introduce the *zero barrier* and *finite ceiling* axioms, both of which promote a weak form of egalitarianism (prioritarianism) by restricting the behavior of welfare functions under extreme inequality.

While welfare maximization and malfare minimization appear to be two sides of the same coin, salient mathematical differences arise. We find in section 3.4 that, unlike malfare functions, welfare functions are not always Lipschitz continuous (though they are at least Hölder continuous), and not uniformly fair-PAC (FPAC) learnable in the sense of Cousins [2021]. However, section 4.3 shows that under a slightly more lenient definition of FPAC learnability, in which sample complexity (i.e., the sufficient sample size to approximately optimize an objective over some class $\mathcal{H}$, for any data distributions, with high probability) is allowed to depend on the welfare function through the *minimum group weight* reciprocal $\frac{1}{\boldsymbol{w}_{\min}}$ and/or the quantity $\frac{1}{|p|}$ (which quantifies how close the welfare function is to the challenging $p = 0$ Nash social welfare case), then if $\mathcal{H}$ is uniformly convergent with polynomial sample complexity, $\mathcal{H}$ is also FPAC learnable with polynomial sample complexity.

We then split the power-mean spectrum into regions, the boundaries of which are defined by our extended axioms, and show that each is in some sense FPAC learnable. This work culminates in theorem 4.4, which shows that if a bounded utility maximization task is PAC-learnable with polynomial sample complexity, then the corresponding *welfare maximization* task is FPAC-learnable, with possible exponential dependence on $\frac{1}{\boldsymbol{w}_{\min}}$ or $\frac{1}{|p|}$, and otherwise fully-polynomial sample complexity. We close section 4 by showing that our sample-complexity bounds can be incorporated into progressive-sampling routines, which adapt their sample consumption to the difficulty of the task at hand, with only *logarithmic multiplicative overhead*.

Our study of fair-PAC learnability is based on the Lipschitz and Hölder continuity of power means, which only coarsely describe their behavior. We thus bound *worst-case* sample complexity, as such analysis is necessarily focused on behavior under extreme inequality of per-group sentiment, where small changes to (dis)utility values are most impactful (as per the *Pigou-Dalton transfer principle*). Under equality, power-mean functions behave roughly linearly, and may thus be much easier to estimate or optimize. We argue that our worst-case bounds for sample complexity are theoretically interesting and practically significant, as they lend confidence to the proposition that a machine learning system accurately and fairly learns an objective. In section 4.4,

we argue that the actual inefficiency (excess sample consumption) stemming from our analysis is not necessarily dependent on the ratio of the best-case to worst-case sample complexity, as efficient progressive sampling algorithms (which draw successively larger samples, and terminate when the objective is estimated or optimized to within some error tolerance ε) generally depend primarily on the *true sufficient sample size*, with an excess factor of the *double-logarithm* of this ratio. Consequently, a polynomial gap between best-case mean estimation and worst case welfare-estimation bounds translates to only *double-logarithmic overhead*, although the exponential gap (dependence on $\frac{1}{\boldsymbol{w}_{\min}}$ and/or $\frac{1}{|p|}$ for $p \approx 0$) may translate to a *logarithmic* increase in sample complexity (relative to methods that are given knowledge of the optimal sample size *a priori*), even with efficient progressive sampling methods.

For the sake of brevity, all proofs are meticulously derived in appendix A. In summary, the primary contributions of this paper are as follows.
1) In section 3.1 we derive a simplified axiomatic basis for cardinal welfare and malfare functions.
2) We extend the core axiomatic basis in section 3.3 with optional axioms, which lead to convenient computational and statistical properties, while enjoying intuitive real-world interpretations, thus guiding welfare function selection.
3) In section 3.4 we analyze the Lipschitz and Hölder continuity of power means. We argue that local behavior is crucial to algorithm analysis, particularly statistical behavior and response to small input perturbations.
4) Section 4 explores the impact of additional axioms on FPAC learnability. In particular, if we allow sample complexity to depend on $\frac{1}{\boldsymbol{w}_{\min}}$ and $\frac{1}{|p|}$, additional axioms naturally split the power-mean spectrum into several regions, and show that each is in some sense FPAC learnable.

2 Related Work

In cardinal fairness learning tasks, we generally receive training data or feedback from *multiple groups*, which represents the needs or wants *of each group*, and we seek to maximize welfare or minimize malfare to *fairly compromise* among groups. Objective choice is a crucial modelling decision, as it mathematically encodes the values of the social planner [Sen, 1977, Roberts, 1980, Moulin, 2004]. Due to variation in human values and value systems, we can't uniquely characterize fairness with mathematics alone, however analysis does help to elucidate the limitations and properties of cardinal objectives. Axiomatic reasoning and analysis from the econometrics literature informs us as to the limitations and properties of cardinal objectives [Pigou, 1912, Dalton, 1920, Debreu, 1959, Gorman, 1968]. The moral philosophy literature also provides insight into social objectives, from classical *utilitarian theory* [Bentham, 1789, Mill, 1863], in which all parties are treated as equals, to *prioritarianism*, [Parfit, 1997, Arneson, 2000], where the less-fortunate are given more weight, to *egalitarian* or *Rawlsian theory* [Rawls, 1971, 2001], which considers the least-fortunate before all others. We also draw from various critiques of utilitarianism [Nozick, 1974, Hurka, 1982] in our analysis.

Prior fair learning work in computer science primarily handles the malfare case. In particular, group-DRO (distributionally robust optimization) methods minimize worst-case (over groups; i.e., egalitarian) risk [Hu et al., 2018, Oren et al., 2019, Sagawa et al., 2019, Dong and Cousins, 2022], which is also known as minimax-fair learning [Diana et al., 2021, Shekhar et al., 2021, Abernethy et al., 2022] and by other names [Martinez et al., 2020, Lahoti et al., 2020, Cortes et al., 2020, Shekhar et al., 2021], and fair-PAC learning generalizes this idea by optimizing arbitrary power-mean malfare objectives [Cousins, 2021, 2022, Cousins et al., 2022a], which derive from an axiomatic welfare theory perspective. Similar algorithmic and statistical concerns arise in the *multi-group agnostic PAC learning* setting [Blum and Lykouris, 2020, Rothblum and Yona, 2021, Cousins, 2022], wherein the goal is to minimize *regret* (over groups) w.r.t. needing to compromise on a *shared model*, rather than each group selecting their own preferred model. Some authors, e.g., Hu and Chen [2020], do discuss direct welfare optimization, however they do not treat the resulting statistical questions or bound generalization error or sample complexity, and thus the issues we identify with the statistical difficulty of welfare optimization are not addressed.

As for welfare-theoretic approaches to fair learning, Heidari et al. [2018] employ axiomatic cardinal welfare theory to introduce fairness constraints for machine learning tasks, and Cousins [2021, 2022] generalizes the axioms of cardinal welfare to allow for per-group weight values, explores computational and statistical learnability, and bounds sample complexity for the malfare ($p \geq 1$) case. Thomas et al. [2019] also introduce a concept of fair statistical learnability, termed the *Seldonian learner*, which encapsulates both constraint-based and cardinal objective-based fair learning settings, however, this framework is so general that it is difficult to establish classes of learnability. In contrast, by adopting an *axiomatic* cardinal-welfare centric viewpoint, we operate over *specific classes* of welfare or malfare functions, which allows us to broadly analyze the sample complexity of various learning problems.

Outside the sphere of fair learning, similar questions and tradeoffs arise in *fair allocation*. In particular, welfare-based fair allocators must explicitly decide how to allocate a pool of limited items (resources) to a set of agents given their valuation (utility) functions over sets of items. In a sense, a machine learning model is also a resource allocator, where the resource is the sentiment achieved by each group as a result of decisions made by the model. Fair allocation problems generally sidestep the issues of generalization error and PAC learnability by assuming known utility, but the

core problems of selecting a fair objective and optimizing a given fair objective remain. Power-means have recently seen use as fair objectives in various fair allocation settings. Barman et al. [2020] show efficient approximation algorithms for unweighted power-means under subadditive valuations, in more restricted classes of submodular valuations, Viswanathan and Zick [2023] give efficient algorithms to maximize weighted power-means, and in broader valuation classes, Cousins et al. [2023b,c] give efficient algorithms to maximize unweighted power-means.

3 On Axiomatizations of Cardinal Fairness Objectives

Herein we define *aggregator functions*, which summarize overall sentiment (either utility or disutility), expressed as a *sentiment vector* $\boldsymbol{u} \in \mathbb{R}^g_{0+}$ over a population of g groups weighted by *weights vector* $\boldsymbol{w} \in \triangle_g$, where $\triangle_g$ is the *nondegenerate probability simplex* over g groups, i.e., $\boldsymbol{w} \in (0,1)^g$ and $\|\boldsymbol{w}\|_1 = 1$. We use $\mathrm{M}(\boldsymbol{u};\boldsymbol{w})$ to denote generic aggregator functions, $\mathrm{W}(\boldsymbol{u};\boldsymbol{w})$ for welfare functions (positively-connoted sentiment), and $\mathrm{M}\!\!\mathrm{M}(\boldsymbol{u};\boldsymbol{w})$ for malfare functions (negatively-connoted sentiment). At times, we are also interested in *unweighted aggregator functions*, which take the form $\mathrm{M}(\boldsymbol{u})$, but in all cases, these will be equivalent to some weighted aggregator function under a uniform weighting, i.e. $\mathrm{M}(\boldsymbol{u}) = \mathrm{M}'_p(\boldsymbol{u}; \langle \frac{1}{g}, \ldots, \frac{1}{g} \rangle)$ for some weighted aggregator function $\mathrm{M}'(\cdot;\cdot)$. We present in section 3.1 an axiomatic basis of desirable properties that is weaker (i.e., makes fewer assumptions) than the standard basis, and in section 3.3, we give additional axioms and strengthened variants of our basis axioms, which further constrain the space of fair aggregator functions. In section 3.2, we explore the resulting classes of aggregator functions, and in section 3.4 we further explore various continuity properties and applications of these classes of aggregator functions.

To concisely denote sentiment, we use functional and vector notation interchangeably, e.g., the *logarithmic utility transformation* of concave utility theory, applied to some $\boldsymbol{u}$, could be written either as $i \mapsto \ln(1 + \boldsymbol{u}_i)$ or as $\langle \ln(1 + \boldsymbol{u}_1), \ln(1 + \boldsymbol{u}_2), \ldots, \ln(1 + \boldsymbol{u}_g) \rangle$. Furthermore, indicator functions, i.e., $\mathbb{1}_a(b)$ is 1 if $b = a$ or $b \in a$, and 0 otherwise, can also be interpreted as *indicator vectors*, where $\mathbb{1}_i = j \mapsto \mathbb{1}_i(j)$ is the ith standard basis vector. This and all other potentially confusing notation pertaining to aggregator functions is summarized in table 1.

Although this work explicitly considers only group-level fairness and decision making, this is predominantly for reasons of *statistical learnability*. Section 4 assumes we can draw many samples *for each group* to estimate their sentiment while learning across a class $\mathcal{H}$ of possible models, which is reasonable when samples correspond to the *individuals* that comprise the group (thus their experiences are averaged to determine the group's sentiment), but it is often unrealistic to assume a learner has access to a large amount of individual-level data. Philosophically, operating at the group level also allows us to consider only expected outcomes over populations, which circumvents the need to reason about randomness and probabilities of individual events. These issues aside (e.g., if the outcomes of any feasible decision were known deterministically), there is no reason our framework could not be applied to individual-level fairness; we would need only substitute the word "individual" for "group," or equivalently, define the population of groups as *singletons* each consisting of exactly one individual. Note also that the *weighted aggregator functions* are implicitly motivated by differences in the relative sizes of groups, whereas at an individual level, it is generally desirable to treat all individuals equally, and thus instead operate with *unweighted aggregator functions*, but as we shall see, the axiomatic justification behind unweighted and weighted aggregator functions are quite similar.

It is important to acknowledge the implicit assumptions of this setup. Generally, sentiment vectors arise from *real-world grounded situations* that impact each group. By construction, aggregator functions are cardinal, and therefore impose a *preference ordering* over sentiment vectors (and the grounded situations that give rise to said sentiment vectors). The social planner seeks to select a grounded situation (i.e., learn a model) to optimize this preference ordering, and is thus impartial towards the grounded situation, except insofar as it impacts each group's sentiment value. This factorization simplifies the question of how to define fairness by avoiding *objective characterization*, and instead defining an *intersubjective concept* (shared welfare or malfare function) based on the *subjective experience* (sentiment value) of each group.

This is inherently an "intersubjective consequentialist" perspective on ethics, valuing the impact of decisions as perceived by the groups affected by them and agreed upon systems of compromise, and it thus draws inspiration from the philosophy of *altruistic hedonism* (i.e., it seeks to maximize the pleasure and minimize the pain of everyone). While we do not explicitly consider other moral and ethical systems, it is not difficult to modify our framework to accommodate many such philosophies. In particular, more paternalistic or prescriptive ethical systems, such as deontological or virtue ethical systems, can be modeled by modifying the definition of sentiment to reflect the feelings or opinions of the social planner (acting as a "morality arbitrator"), rather than those of the groups impacted by decisions (for example, if *truth* is a moral virtue, then a classifier should strive for accuracy, even if the decisions cause harm and leave each group in a worse position than would a less accurate model). Similarly, more solipsistic moral frameworks, e.g., *egotistical hedonism* (wherein one's concept of morality aligns with that which brings them pleasure and avoids bringing them pain),

Object	Definition or Space	Description
$\triangle_g$	$\doteq \{\boldsymbol{w} \in (0,1)^g \mid \|\boldsymbol{w}\|_1 = 1\}$	Nondegenerate probability simplex over g atoms
$i \mapsto f_i(\boldsymbol{u}_i)$	$\doteq \langle \boldsymbol{f}_1(\boldsymbol{u}_1), \dots, \boldsymbol{f}_g(\boldsymbol{u}_g) \rangle$	Functional vector notation
$\mathbb{1}_{\boldsymbol{a}}(b)$	$\doteq 1$ if $b \in \boldsymbol{a}$, 0 otherwise	Indicator vector (or function)
$\mathbb{1}_i$	$\doteq j \mapsto \mathbb{1}_{\{i\}}(j)$	One-hot indicator vector
$\mathcal{G}$	Usually $\{1, \dots, g\}$	Group identity space (e.g., race, gender, or language categories)
g	$\doteq \lvert\mathcal{G}\rvert$	Group count or cardinality (usually finite)
r	$\in [0, \infty]$	Maximum possible sentiment value (range)
$\boldsymbol{u}$	$\in [0, r]^g$	Per-group sentiment (utility or disutility), i.e., $\boldsymbol{u}_i$ pertains to group i
$\boldsymbol{w}$	$\in \triangle_g$	Probability weighting over groups (probability simplex)
$\mathrm{M}(\boldsymbol{u}), \mathrm{W}(\boldsymbol{u}), \mathrm{M}\!\!\mathrm{M}(\boldsymbol{u})$	$\in [0, r]^g \to \mathbb{R}_{0+}$	Unweighted aggregator, welfare, or malfare function
$\mathrm{M}(\boldsymbol{u}; \boldsymbol{w}), \mathrm{W}(\boldsymbol{u}; \boldsymbol{w}), \mathrm{M}\!\!\mathrm{M}(\boldsymbol{u}; \boldsymbol{w})$	$\in \big([0, r]^g \times \triangle_g\big) \to \mathbb{R}_{0+}$	Weighted aggregator, welfare, or malfare function
$\mathrm{M}_p(\boldsymbol{u}; \boldsymbol{w})$	See definition 3.4	Weighted power-mean aggregator function

Table 1: Functions, spaces, and common variables for aggregator functions.

can be modelled by changing the social planner from an unbiased abstract outsider not impacted by their decisions into a member of the population being impacted that selfishly makes decisions influenced by their own sentiment.

3.1 A Fundamental Basis of Cardinal Welfare Axioms

We now present a reduced basis of cardinal welfare axioms, which are componentwise-weaker than the standard basis, yet we find that the same Debreu-Gorman [Debreu, 1959, Gorman, 1968] type theorems and Pigou-Dalton [Pigou, 1912, Dalton, 1920] characterizations of fairness still hold.

As a prelude to the dense axiomatic mathematics that will soon follow, we briefly discuss the goals of such reasoning and analysis. It is tempting to think that the value in these axioms derives from their consequent properties, but this line of reasoning is dangerous, as it may lead to "baxiomatic" quasireligious thinking, i.e., the pressure to accept the axioms comes from the convenient or wishful thinking that, once we do so, we can reason over a convenient space, and thus normatively refrain from considering any situation outside their purview. To remain grounded, we thus argue that *the axioms themselves* must be *inherently reasonable*. This "I know it when I see it" criterion, however, does lead to a metric for of the efficacy of an axiomatization: namely, is it *simple enough* to be *easily understood* by rational thinkers, and are they likely to agree with *all parts* of it (emphasis on all, as "most" is not good enough: a single faulty premise can destroy an entire chain of reasoning)?

For these reasons, in settings of axiomatic reasoning such as this, elegance and simplicity are paramount, as they lead to interpretable and uncontroversial axioms. Formal logic plays a role as well, as showing that $A \implies B$ and $A \neq B$, or in other words, B is strictly weaker than A, tells us that B is less likely to be rejected by any rational thinker than A. We thus seek to ground our reasoning on a *minimal basis* of axioms, each of which cannot be weakened individually without changing their collective effect. To emphasize the axiomatization itself, rather than its consequent properties, we present our axioms in isolation, before stating any of their resultant properties. We encourage the reader to evaluate them in and of themselves, though occasionally we do mention specific consequences of various assumptions, in particular to illustrate concrete differences between weaker or stronger variants of some axioms.

Classical econometric theory primarily describes the *unweighted case*, wherein an aggregator function $\mathrm{M}(\boldsymbol{u})$ aggregates sentiment $\boldsymbol{u}$ over an *unweighted finite discrete* population $\mathcal{G}$, i.e., $\mathcal{G} = \{1, \dots, g\}$. We generalize this setting to the *weighted discrete* case, wherein an aggregator function $\mathrm{M}(\boldsymbol{u}; \boldsymbol{w})$ operates on a $\boldsymbol{w}$*-weighted discrete* (possibly countably infinite) population $\mathcal{G}$, i.e., $\mathcal{G} = \{1, \dots, g\}$, for $g \in \mathbb{Z}_+ \cup \{\infty\}$. We modify the axiomatization of Cousins [2021], showing that a reduced axiomatic basis is equivalent.

Axiomatization 3.1 (Weighted Aggregator Axioms)**.** We define axioms for aggregator function $\mathrm{M}(\boldsymbol{u}; \boldsymbol{w})$ below. For each item, assume (if necessary) that the axiom applies for all $\boldsymbol{u}, \boldsymbol{u}' \in \mathbb{R}_{0+}^g$ scalars $\alpha, \varepsilon \in \mathbb{R}_+$, indices $i, j \in \mathcal{G}$, and discrete probability measures $\boldsymbol{w} \in \triangle_g$ over $\mathcal{G}$.

1) *Strict Monotonicity* (SM): Suppose $\boldsymbol{u} \succ 0$, i.e., each $\boldsymbol{u}_j > 0$. Then $\mathrm{M}(\boldsymbol{u}; \boldsymbol{w}) < \mathrm{M}(\boldsymbol{u} + \varepsilon \mathbb{1}_i; \boldsymbol{w})$.

2) *Weighted Symmetry* (WS): For all permutations π over $\mathcal{G}$, it holds that $\mathrm{M}(\boldsymbol{u}; \boldsymbol{w}) = \mathrm{M}\big(\pi(\boldsymbol{u}); \pi(\boldsymbol{w})\big)$.

3) *Weighted Decomposability* (WD): Suppose $\alpha \in (0, 1)$. Then $\mathrm{M}(\boldsymbol{u}; \boldsymbol{w}) = \mathrm{M}\big(\langle \boldsymbol{u}_1, \boldsymbol{u}_1, \boldsymbol{u}_2, \dots \rangle; \langle \alpha \boldsymbol{w}_1, (1-\alpha)\boldsymbol{w}_1, \boldsymbol{w}_2, \dots \rangle\big)$.

4) *Continuity*: $\mathrm{M}(\boldsymbol{u}; \boldsymbol{w})$ is a continuous function in $\boldsymbol{u}$.

5) *Independence of Unconcerned Agents* (IUA): If $\boldsymbol{u}_i = \boldsymbol{u}'_i$, then $\mathrm{M}(\boldsymbol{u}; \boldsymbol{w}) \leq \mathrm{M}(\boldsymbol{u}'; \boldsymbol{w}) \Leftrightarrow \mathrm{M}(\boldsymbol{u} + \varepsilon \mathbb{1}_i; \boldsymbol{w}) \leq \mathrm{M}(\boldsymbol{u}' + \varepsilon \mathbb{1}_i; \boldsymbol{w})$.

6) *Multiplicative Linearity*: $\mathrm{M}(\alpha \boldsymbol{u}; \boldsymbol{w}) = \alpha \mathrm{M}(\boldsymbol{u}; \boldsymbol{w})$.

7) *Unit Scale*: $\mathrm{M}(\mathbf{1}; \boldsymbol{w}) = \mathrm{M}(i \mapsto 1; \boldsymbol{w}) = 1$.

8) *Weak Transfer Principle* (WTP): Let $i \doteq \operatorname{argmin}_i \boldsymbol{u}_i$ & $j \doteq \operatorname{argmax}_j \boldsymbol{u}_j$. If $\boldsymbol{u}_i \neq \boldsymbol{u}_j$, then *there exists* some $\varepsilon > 0$ s.t. $\boldsymbol{u}_i + \boldsymbol{w}_j \varepsilon < \boldsymbol{u}_j - \boldsymbol{w}_i \varepsilon$, and $\mathrm{W}(\boldsymbol{u} + \varepsilon \boldsymbol{w}_j \mathbb{1}_i - \varepsilon \boldsymbol{w}_i \mathbb{1}_j; \boldsymbol{w}) \geq \mathrm{W}(\boldsymbol{u}; \boldsymbol{w})$ for welfare or $\mathrm{M}\!\!\mathrm{M}(\boldsymbol{u} + \varepsilon \boldsymbol{w}_j \mathbb{1}_i - \varepsilon \boldsymbol{w}_i \mathbb{1}_j; \boldsymbol{w}) \leq \mathrm{M}\!\!\mathrm{M}(\boldsymbol{u}; \boldsymbol{w})$ for malfare.

Axioms 1–8 are generally assumed in this work, but we

present several alternatives below, to which we compare.
9) *Weighted Additivity* (WA): Suppose $g' \in \mathbb{Z}_+ \cup \{\infty\}$, $\boldsymbol{u}' \in \mathbb{R}_{0+}^{g'}$, and weights vector $\boldsymbol{w}' \in \triangle_{g'}$ such that for all $u \in \mathbb{R}_{0+}$, it holds that $\sum_{i \in \mathcal{G}} \boldsymbol{w}_i \mathbb{1}_u(\boldsymbol{u}_i) = \sum_{i \in \mathcal{G}'} \boldsymbol{w}'_i \mathbb{1}_u(\boldsymbol{u}'_i)$. Then $\mathrm{M}(\boldsymbol{u}; \boldsymbol{w}) = \mathrm{M}(\boldsymbol{u}'; \boldsymbol{w}')$.
10) *Independence of Common Scale* (ICS): $\mathrm{M}(\boldsymbol{u}; \boldsymbol{w}) \leq \mathrm{M}(\boldsymbol{u}'; \boldsymbol{w}) \implies \mathrm{M}(\alpha \boldsymbol{u}; \boldsymbol{w}) \leq \mathrm{M}(\alpha \boldsymbol{u}'; \boldsymbol{w})$.
11) *Pigou-Dalton Transfer Principle* (PDTP)[2]: If $\boldsymbol{u}_i + \boldsymbol{w}_j \varepsilon \leq \boldsymbol{u}_j - \boldsymbol{w}_i \varepsilon$, then $\mathrm{W}(\boldsymbol{u} + \boldsymbol{w}_j \varepsilon \mathbb{1}_i - \boldsymbol{w}_i \varepsilon \mathbb{1}_j; \boldsymbol{w}) \geq \mathrm{W}(\boldsymbol{u}; \boldsymbol{w})$ for welfare or $\mathrm{M}(\boldsymbol{u} + \boldsymbol{w}_j \varepsilon \mathbb{1}_i - \boldsymbol{w}_i \varepsilon \mathbb{1}_j; \boldsymbol{w}) \leq \mathrm{M}(\boldsymbol{u}; \boldsymbol{w})$ for malfare.

We now pause to discuss the rationale behind each axiom. Axioms 1–5 & 11 generalize the standard basis of *axioms of cardinal welfare* to *weighted discrete populations*, and together they imply that any aggregator function can be expressed as $\mathrm{M}(\boldsymbol{u}; \boldsymbol{w}) \doteq F\big(\sum_{i=1}^{g} \boldsymbol{w}_i f(\boldsymbol{u}_i)\big)$ for strictly monotonically increasing functions f, F. A similar basis of axioms pertains to the class of *unweighted aggregator functions*, i.e., those of the form $\mathrm{M}(\boldsymbol{u})$, where implicitly all weights are equal (which translate to the form $\mathrm{M}(\boldsymbol{u}; \langle \frac{1}{g}, \ldots, \frac{1}{g} \rangle)$ in the weighted nomenclature). In particular, unweighted variants of axioms 1 & 4–7 simply drop all weights terms, and are otherwise identical to their weighted counterparts. Unweighted variants of the remaining axioms are slightly more involved, and are discussed in turn below.

On Weighted Additivity In the unweighted case, it is standard to define symmetry as simply $\mathrm{M}(\boldsymbol{u}) = \mathrm{M}\big(\pi(\boldsymbol{u})\big)$ for all permutations π over $\mathcal{G}$. With weights, *weighted symmetry* (axiom 2), i.e., $\mathrm{M}(\boldsymbol{u}; \boldsymbol{w}) = \mathrm{M}\big(\pi(\boldsymbol{u}); \pi(\boldsymbol{w})\big)$, only requires *equal treatment* given *equal weights*. *Weighted decomposability* (axiom 3) then codifies the relative impact of weights by requiring that a group can be decomposed into two groups of equal sentiment and total weight without changing the aggregate. Prior work [Cousins, 2021, 2022] assumes *weighted additivity* (axiom 9) directly, but we argue that this axiom seems rather contrived and unintuitive, whereas axioms 2 & 3 are so natural that it would be perverse not to assume them. We now show that, despite their vastly simpler form, together axioms 2 & 3 equate to axiom 9.

Lemma 3.2 (Equivalence of Weighted Axioms)**.** Consider some aggregator function $\mathrm{M}(\cdot; \boldsymbol{w})$. It always holds that WS (axiom 2) $\wedge$ WD (axiom 3) $\Leftrightarrow$ WA (axiom 9).

On Units, Scale, and Canonical Forms Axiom 6 (*multiplicative linearity*) is a natural and useful property, and ensures that *dimensional analysis* on aggregator functions is possible; in particular, *units* of aggregator functions match those of sentiment values. Axiom 6 is also known in the *Constant Elasticity of Substitution* (CES) literature [Arrow et al., 1961, McFadden, 1963] as *homogeneity of degree 1*. Note that axiom 6 implies axiom 10, and is thus a simple strengthening of a more basic traditional cardinal welfare axiom. Axiom 7 (*unit scale*) furthers this theme, as it ensures that not only do *units* of aggregates match those of $\boldsymbol{u}$, but *scale* does as well, thus axiom 7 accords with *average utilitarianism* Hurka [1982], rather than *sum utilitarianism*, as we do not depend on the *size* of $\mathcal{G}$. Cousins [2021] shows that these axioms lead to the *power-mean* characterization of aggregator functions. Weakening axioms 6 & 7 to just axiom 10, the Debreu-Gorman theorem still implies a *monotonic transform* of the power-mean, so axioms 6 & 7 don't impact *comparisons between* aggregator values, but merely specify a convenient and elegant *canonical form* for their cardinal values.

On Equitable Redistribution and Transfer Principles The *Pigou-Dalton transfer principle* (PDTP, axiom 11) is also standard in cardinal welfare theory. It essentially states that transferring (dis)utility between two groups is *not harmful*, up to the point where the two groups have equal (dis)utility, thus it incentivizes *equitable redistribution of "wealth."* This codifies the intuition that redistributing (dis)utility towards equitability is not harmful to society.

One could argue that, while a general trend towards equality may be good, this characterization of radical equality is too strong. The *weak transfer principle* (WTP, axiom 8) is less impeachable in this regard, as it weakens the quantifier over transfer magnitude from *universal* to *existential*, i.e., it states only that transferring *some nonzero amount* of (dis)utility between the (dis)utility maximizing and minimizing groups is not harmful, making no claim about the remaining groups, or the magnitude of the transfer. We now show that, subject to the standard Debreu-Gorman axioms, the WTP and PDTP are equivalent, thus the radical equality characterization of the PDTP is not necessary in this context.

Lemma 3.3 (Transfer Principle Equivalencies)**.** Consider some aggregator function $\mathrm{M}(\cdot; \boldsymbol{w})$. The following relate properties (axioms) that $\mathrm{M}(\cdot; \boldsymbol{w})$ obeys.
1) PDTP (11) $\implies$ WTP (8); &
2) Suppose axioms 1–7. Then WTP (8) $\implies$ PDTP (11).

Note cautiously that both the PDTP and WTP are careful to claim that equitable transfers are *not harmful* to society, rather than *beneficial*. Section 3.3 presents strong variants of these axioms, which require *strict benefit* to equitable transfers. To obtain unweighted variants of the transfer principle axioms, we again simply substitute in uniform weights, but here this results in some elegant simplifications. In particular, the WTP and PDTP assume (either existentially or universally) some $\varepsilon > 0$ such that $\boldsymbol{u}_i + \boldsymbol{w}_j \varepsilon$ and $\boldsymbol{u}_j - \boldsymbol{w}_i \varepsilon$ obey some relationship. Observe that all weighting terms essentially cancel, and we now recover the guarantee for the

[2] Cousins [2021, 2022] adopts a seemingly more complicated variant of this axiom, but it is equivalent for countable populations by repeated application. In particular, they take the following: suppose $\mu = \mathbb{E}_{\boldsymbol{w}}[\boldsymbol{u}] = \mathbb{E}_{\boldsymbol{w}}[\boldsymbol{u}']$, and $\forall i \in \mathcal{G} : |\mu - \boldsymbol{u}'_i| \leq |\mu - \boldsymbol{u}_i|$. Then $\mathrm{W}(\boldsymbol{u}'; \boldsymbol{w}) \geq \mathrm{W}(\boldsymbol{u}; \boldsymbol{w})$ for welfare, or $\mathrm{W}(\boldsymbol{u}'; \boldsymbol{w}) \leq \mathrm{W}(\boldsymbol{u}; \boldsymbol{w})$ for malfare.

change-of-variables $\varepsilon' = \frac{1}{g}\varepsilon$, since $\boldsymbol{w}_j\varepsilon = \boldsymbol{w}_i\varepsilon = \varepsilon'$.

3.2 The Power Mean

We now define the class of weighted power-mean aggregator functions, and show that our aggregator function axioms are uniquely satisfied by this class.

Definition 3.4 (Power-Mean Welfare and Malfare)**.** Suppose $p \in \mathbb{R} \cup \pm\infty$. The *weighted power-mean*, given *sentiment vector* $\boldsymbol{u} \in \mathbb{R}^g_{0+}$ and *weights vector* $\boldsymbol{w} \in \triangle_g$, is defined as

$$\mathrm{M}_p(\boldsymbol{u}; \boldsymbol{w}) \doteq \lim_{\rho \to p} \lim_{\varepsilon \to 0^+} \sqrt[\rho]{\sum_{i=1}^{g} \boldsymbol{w}_i (\boldsymbol{u}_i + \varepsilon)^\rho} \ . \qquad (1)$$

Note that in most cases the limits can safely be ignored. The inner $\lim_{\varepsilon \to 0^+}$ simply avoids indeterminate forms for $p \leq 0$ while some $\boldsymbol{u}_i = 0$ while preserving continuity. The outer $\lim_{\rho \to p}$ resolves to the *weighted geometric mean* for $p = 0$, i.e., $\mathrm{M}_0(\boldsymbol{u}; \boldsymbol{w}) = \lim_{\varepsilon \to 0^+} \prod_{i=1}^{g} (\boldsymbol{u}_i + \varepsilon)^{\boldsymbol{w}_i}$, generally termed *Nash social welfare* in this context. Similarly, $\lim_{\rho \to p}$ resolves to the unweighted *maximum* and *minimum* operators for $p = \pm\infty$, also known as *egalitarian* malfare or welfare, respectively. We omit the weighting $\boldsymbol{w}$ to denote unweighted power-means, which are equivalent to weighted power-means under a uniform weighting, i.e. $\mathrm{M}_p(\boldsymbol{u}) = \mathrm{M}_p(\boldsymbol{u}; \langle \frac{1}{g}, \ldots, \frac{1}{g} \rangle)$.

We now characterize the class of fair aggregator functions in a result similar to theorem 2.4 of Cousins [2021], albeit under our reduced axiomatic basis.

Theorem 3.5 (Aggregator Function Properties)**.** Suppose aggregator function $\mathrm{M}(\boldsymbol{u}; \boldsymbol{w})$, and assume arbitrary sentiment vector $\boldsymbol{u} \in \mathbb{R}^g_{0+}$ and weights vector $\boldsymbol{w} \in \triangle_g$. The following then hold.
1) *Power-Mean Factorization*: Axioms 1–7 imply there exists some $p \in \mathbb{R}$ such that $\mathrm{M}(\boldsymbol{u}; \boldsymbol{w}) = \mathrm{M}_p(\boldsymbol{u}; \boldsymbol{w})$.
2) *Fair Welfare and Malfare*: Axioms 1–8 imply $p \in (-\infty, 1]$ for welfare and $p \in [1, \infty)$ for malfare.

3.3 Extended and Contextual Axioms

We now present new axioms and stronger variants of the axioms thus far stated. These generally extend the themes and justifications of weaker axioms, and thus require a larger concession to accept, but they have greater descriptive power and reduce the space of admissible fair aggregator functions.

Axiomatization 3.6 (Strong Axioms)**.** Suppose as in axiomatization 3.1. The following two axioms strengthen various components of axiomatization 3.1.
12) Strict Monotonicity at 0 (SM0): $\mathrm{M}(\boldsymbol{u}; \boldsymbol{w}) < \mathrm{M}(\boldsymbol{u} + \varepsilon \mathbb{1}_i; \boldsymbol{w})$.
13) *Strict Weak Transfer Principle* (SWTP): Let $i \doteq \operatorname{argmin}_i \boldsymbol{u}_i$, $j \doteq \operatorname{argmax}_j \boldsymbol{u}_j$. If $\boldsymbol{u}_i \neq \boldsymbol{u}_j$, then *there exists* some $\varepsilon > 0$ s.t. $\boldsymbol{u}_i + \boldsymbol{w}_j\varepsilon < \boldsymbol{u}_j - \boldsymbol{w}_i\varepsilon$ and $\mathrm{W}(\boldsymbol{u} + \varepsilon\boldsymbol{w}_j\mathbb{1}_i - \varepsilon\boldsymbol{w}_i\mathbb{1}_j; \boldsymbol{w}) > \mathrm{W}(\boldsymbol{u}; \boldsymbol{w})$ for welfare, or $\mathrm{M}(\boldsymbol{u} + \varepsilon\boldsymbol{w}_j\mathbb{1}_i - \varepsilon\boldsymbol{w}_i\mathbb{1}_j; \boldsymbol{w}) < \mathrm{M}(\boldsymbol{u}; \boldsymbol{w})$ for malfare.
14) *Strict PDTP* (SPDTP): Suppose $\boldsymbol{u}_i + \boldsymbol{w}_j\varepsilon < \boldsymbol{u}_j - \boldsymbol{w}_i\varepsilon$. Then $\mathrm{W}(\boldsymbol{u} + \boldsymbol{w}_j\varepsilon\mathbb{1}_i - \boldsymbol{w}_i\varepsilon\mathbb{1}_j; \boldsymbol{w}) > \mathrm{W}(\boldsymbol{u}; \boldsymbol{w})$ for welfare, or $\mathrm{M}(\boldsymbol{u} + \boldsymbol{w}_j\varepsilon\mathbb{1}_i - \boldsymbol{w}_i\varepsilon\mathbb{1}_j; \boldsymbol{w}) < \mathrm{M}(\boldsymbol{u}; \boldsymbol{w})$ for malfare.

Lemma 3.7 (Consequences of Strong Axioms)**.** Suppose power-mean aggregator function $\mathrm{M}_p(\cdot; \boldsymbol{w})$. The following then hold:
1) Strengthening the SM axiom (i.e., $1 \to 12$) implies $p > 0$.
2) Strengthening the WTP axiom (i.e., $8 \to 13$) implies $p \neq 1$, thus $p < 1$ for welfare and $p > 1$ for malfare.
3) Strict PDTP (i.e., $11 \to 14$) implies $p \neq 1$ and $p \neq \pm\infty$, thus $p \in (-\infty, 1)$ for welfare and $p \in (1, \infty)$ for malfare.

We now comment on the philosophical implications of lemma 3.7. The consequences of SM0 are immense: the "brand name" Nash social welfare ($p = 0$) is now inadmissible as a fair welfare function, and moreover we reduce the unbounded spectrum of p to just $p \in (0, 1]$. Intuitively, the "strict" aspect of SM0 encodes the idea that gains to utility should *always* be relevant, and consequently prevents a form of "minority rule," wherein a group with utility 0 ensures that welfare can not possibly improve without benefiting said group. Note that for any $p < 1$, the weighted relative impact of helping disadvantaged groups is still higher than privileged groups (as can be seen by inspecting the power-mean gradient, see lemma 3.10), but SM0 puts a sharp limit on the strength of this effect by preventing $p \leq 0$.

Under SWTP, pure utilitarianism ($p = 1$) is inadmissible: intuitively, transferring any ε utility would, by linearity, not change welfare, thus not yield strict improvement. In this sense, SWTP incentivizes equitable redistribution of wealth more strongly than does WTP. Strengthening of PDTP to strict inequality is also interesting, but it necessarily precludes both the utilitarian ($p = 1$) and egalitarian cases ($p \in \pm\infty$), and thus does not actually represent a strict preference towards egalitarianism.

The following axioms control the "degree of prioritarianism" more precisely than do our transfer principles. Each describes the behavior of a welfare function under situations of *extreme inequality*, where some group receives 0 or ∞ utility.

Axiomatization 3.8 (Extreme Axioms)**.** Suppose as in axiomatization 3.1. We now define two additional axioms.
15) *Zero Barrier* (0B): $\lim_{\boldsymbol{u}_i \to 0^+} \mathrm{W}(\boldsymbol{u}; \boldsymbol{w}) = 0$.
16) *Finite Ceiling* (FC): $\lim_{c \to \infty} \mathrm{W}(\boldsymbol{u} + c\mathbb{1}_i; \boldsymbol{w}) < \infty$.

Nozick [1974] criticizes utilitarianism via *reductio ad absurdum* by positing a "utility monster," which derives extremely high utility from some good, and thus utilitarian theory dictates we must allocate all resources to the monster. Our *zero barrier axiom* (15) promotes prioritarianism by ensuring that welfare $\mathrm{W}(\boldsymbol{u}; \boldsymbol{w}) \to 0$ as any group utility $\boldsymbol{u}_i \to 0$, a property shared by the egalitarian welfare, and furthermore, the zero barrier axiom incentivizes aiding the most needy by creating a "barrier" at 0 utility, thus disincentivizing *ex-*

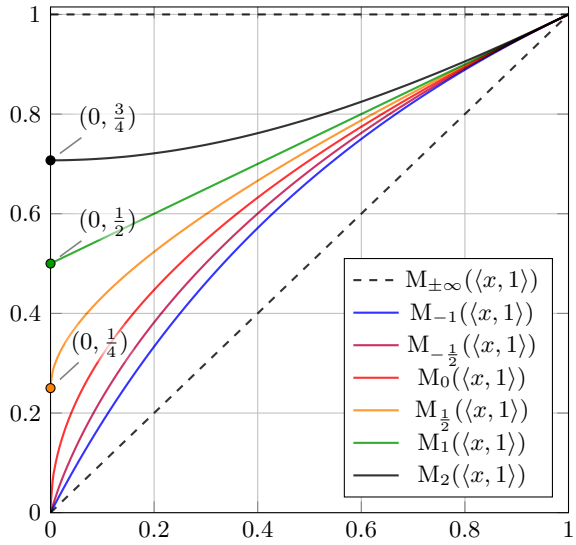


Figure 1: Plots of the unweighted power-mean (i.e., $\boldsymbol{w} \doteq \langle \frac{1}{2}, \frac{1}{2} \rangle$) for various p. Observe that the region around $x = \boldsymbol{u}_1 = 0$, wherein Lipschitz discontinuities may occur, exhibits sharp changes to welfare, as $\mathrm{M}_p(\langle x, 1\rangle)$ is very sensitive to small changes to x.

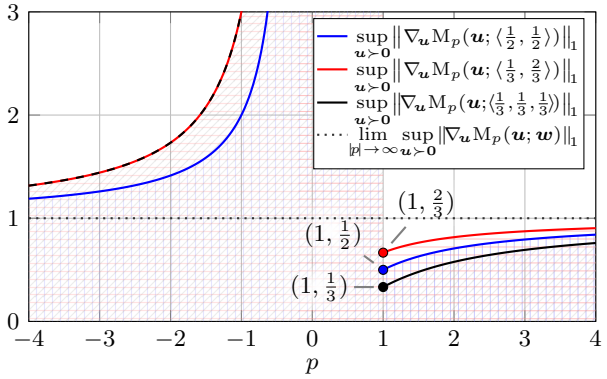


Figure 2: Plots of $\|\cdot\|_1$ Lipschitz constants of weighted power-means for various weightings as functions of p. Note that $\sup_{\boldsymbol{u}\succ\mathbf{0}} \|\nabla_{\boldsymbol{u}} \mathrm{M}_p(\boldsymbol{u}; \boldsymbol{w})\|_1 = \sup_{\boldsymbol{u},\boldsymbol{u}'\succ\mathbf{0}} \frac{|\mathrm{M}(\boldsymbol{u};\boldsymbol{w}) - \mathrm{M}(\boldsymbol{u}';\boldsymbol{w})|}{\|\boldsymbol{u}-\boldsymbol{u}'\|_1}$ describes this Lipschitz constant, as $\mathrm{M}(\cdot; \boldsymbol{w})$ is assumed to be continuous, and for any $\boldsymbol{w}$, this quantity approaches 1 as $p \to \pm\infty$ (plotted as a dashed line).The region below each Lipschitz constant plot is shaded and patterned, to emphasize that *higher values* allow for *sharper rates of change* to the power-mean functions.

treme harm to any group, and bounding the harm caused by the utility monster. Similarly, the *finite ceiling axiom* (16) ensures that even as some $\boldsymbol{u}_i \to \infty$, $\mathrm{W}(\boldsymbol{u}; \boldsymbol{w})$ *remains finite*, i.e., the disadvantaged (finite utility) groups are not "forgotten" in the monster's rush toward infinite utility (and infinite inequality).

Lemma 3.9 (Consequences of Extreme Axioms)**.** Suppose as in lemma 3.7. The following then hold.
1) 0B (axiom 15) $\Leftrightarrow p \leq 0$. 2) FC (axiom 16) $\Leftrightarrow p < 0$.

Observe that, subject to axioms 1–8, FC $\implies$ 0B $\implies$ SWTP $\implies$ WTP. The above "egalitarian" framing of the 0B and FC axioms is complemented by a "utilitarian" framing, wherein we would require that taking the utility of any group to 0 *does not* take welfare to 0, or that taking the utility of any group to ∞ *does* unboundedly increase welfare, in each case concluding the complementary set of permissible p in lemma 3.9, though again such axioms codify intuition, but can not dictate the "correct" welfare function.

Some authors also assume variants of the *Independence of Irrelevant Alternatives* (IIA) axiom, which restricts to the Nash social welfare [Roth et al., 1977, Kaneko and Nakamura, 1979], i.e., $p = 0$. We do not discuss this axiom further, but note that in our framework, it is equivalent to jointly assuming axiom 15 and the utilitarian form of 16.

3.4 Continuity Properties of Aggregator Functions

Another axiom that can be strengthened to great effect is the *continuity* axiom. Myriad varieties of continuity exist, and we investigate the Hölder and Lipschitz varieties in great detail here. In particular, while it does make sense in principle to axiomatically assume a stronger notion of continuity, due to their parameterized nature, such characterizations lack the elegant simplicity of our axiomatization. We present the material in this manner, so as to emphasize the consequences of choice of aggregator function on continuity, rather than how the act of assuming continuity impacts the aggregator function.

Previous works bound deviations between power-mean malfare functions [Cousins, 2021, 2022], and analyze their Lipschitz continuity [Beliakov et al., 2009]. We extend this analysis to power-mean welfare functions (i.e., $p \leq 1$), showing that they are Lipschitz continuous for $p < 0$, though not for $p \in [0, 1)$, and the Lipschitz constants depend on the *minimum weight* $\boldsymbol{w}_{\min}$. This is initially surprising, as intuitively, low-weight groups should have little impact on the power mean, however we know that for $p \leq 0$, by lemma 3.9 item 1, as any group's sentiment $\boldsymbol{u}_i \to 0$, then so too must the power mean $\mathrm{M}_p(\boldsymbol{u}; \boldsymbol{w}) \to 0$, thus as $\boldsymbol{w}_i \to 0$, this must occur more rapidly, hence the dependence on $\boldsymbol{w}_{\min}$.

While it is not unreasonable to axiomatically assume a stronger notion of continuity, due to their parameterized nature, such characterizations lack the elegant simplicity of our axiomatization. We thus present continuity properties as *consequent from* choice of welfare function, rather than *vice versa*, to reflect the practical impact of this choice. We now analyze the local behavior of power means, first through their gradients, and then their Lipschitz and Hölder continuity properties. The reader is invited to reference figure 1 throughout, wherein various power-means are plotted, revealing their pathological and highly nonlinear behavior for $p \approx 0$.

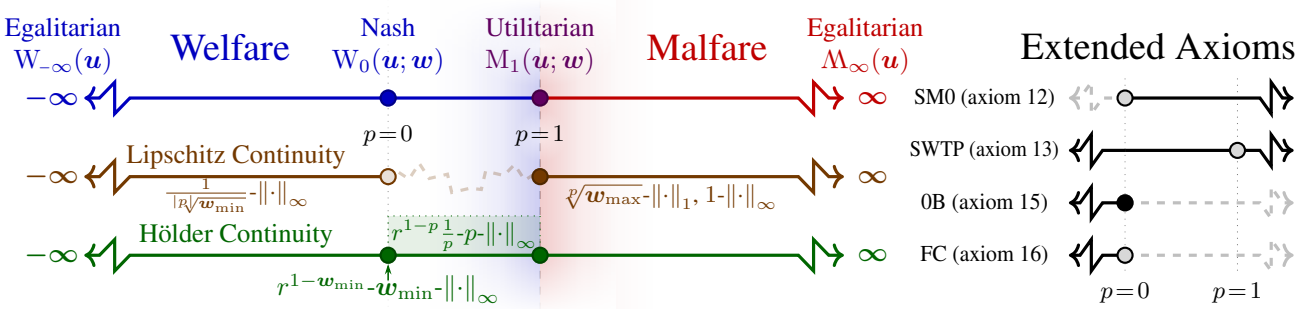


Figure 3: Illustration of power-mean properties and axioms. Solid lines and filled circles denote the values of p that concord with an axiom or property, while dashed lines and unfilled circles denote their complement. Basic properties (theorem 3.5) and continuity (lemmata 3.12 & 3.13) are plotted on the left, and the consequences of our strong and extended axioms (lemmata 3.7 & 3.9) are shown on the right.

Lemma 3.10 (Power-Mean Differentiation)**.** Suppose $\boldsymbol{u}_{\setminus i} \succ \boldsymbol{0}$, some weights vector $\boldsymbol{w} \in \triangle_g$, and $p \in \mathbb{R}$. The power mean then differentiates in $\boldsymbol{u}_i$ as follows.
1) If $\boldsymbol{u}_i > 0$, then $\frac{\partial}{\partial \boldsymbol{u}_i}\mathrm{M}_p(\boldsymbol{u};\boldsymbol{w}) = \frac{\boldsymbol{w}_i \boldsymbol{u}_i^{p-1}}{\mathrm{M}_p^{p-1}(\boldsymbol{u};\boldsymbol{w})}$.
2) If $p < 0$, then $\lim_{\boldsymbol{u}_i \to 0^+} \frac{\partial}{\partial \boldsymbol{u}_i}\mathrm{M}_p(\boldsymbol{u};\boldsymbol{w}) = {}^{-p}\sqrt{\frac{1}{\boldsymbol{w}_i}}$.
3) If $p \in [0,1)$, then $\lim_{\boldsymbol{u}_i \to 0^+} \frac{\partial}{\partial \boldsymbol{u}_i}\mathrm{M}_p(\boldsymbol{u};\boldsymbol{w}) = \infty$.

Definition 3.11 (Lipschitz and Hölder Continuity)**.** An aggregator function $\mathrm{M}(\boldsymbol{u};\boldsymbol{w})$ is *Hölder continuous* in the variable $\boldsymbol{u}$ w.r.t. some norm $\|\cdot\|_{\mathrm{M}}$ over domain $[0,r]^g$ if there exist some *scale* $\lambda \geq 0$ and *power* $\alpha \in (0,1]$, such that for all $\boldsymbol{u}, \boldsymbol{u}' \in [0,r]^g$, it holds that

$$|\mathrm{M}(\boldsymbol{u};\boldsymbol{w}) - \mathrm{M}(\boldsymbol{u}';\boldsymbol{w})| \leq \lambda \|\boldsymbol{u} - \boldsymbol{u}'\|_{\mathrm{M}}^{\alpha} \ . \qquad (2)$$

We say that such a function is λ-α-$\|\cdot\|_{\mathrm{M}}$ Hölder continuous, and if $\alpha = 1$, it is λ-$\|\cdot\|_{\mathrm{M}}$ Lipschitz continuous, and if $\lambda \leq 1$ it is $\|\cdot\|_{\mathrm{M}}$ *nonexpansive*.

Intuitively, Lipschitz continuity bounds *infinitessimal rates of change*, and Hölder continuity bounds *change over small regions*, where the relative size of the change grows larger as $\|\cdot\|_{\mathrm{M}} \to 0^+$. The relationship between Lipschitz, Hölder, and standard ε-δ limit continuity for range r $\|\cdot\|_{\mathrm{M}}$ is

$$\lambda \text{ Lipschitz} \implies \lambda r^{-\alpha}\text{-}\alpha \text{ Hölder} \implies \varepsilon\text{-}\delta \text{ Limit} \ .$$

As we assume *continuity* throughout (axiom 4), all aggregator functions of interest are tautologically ε-δ limit continuous, however we shall see that they do not all share the same Hölder and Lipschitz continuity properties, which has crucial implications for sampling and learning from data, as well as privacy and algorithmic stability. The following result (visualized in figure 2) analyzes Lipschitz continuity.

Lemma 3.12 (Power-Mean Lipschitz Continuity)**.** Suppose $p \in \mathbb{R}$, sentiment vectors $\boldsymbol{u}, \boldsymbol{u}' \in \mathbb{R}_{0+}^g$, and weights vector $\boldsymbol{w} \in \triangle_g$. The following then hold.
1) Suppose $p \geq 1$. Then $\mathrm{M}_p(\cdot;\boldsymbol{w})$ is $\sqrt[p]{\boldsymbol{w}_{\max}}$-$\|\cdot\|_1$, 1-$\mathrm{M}_p(|\cdot|;\boldsymbol{w})$, and 1-$\|\cdot\|_\infty$ Lipschitz.
2) Suppose $p < 0$. Then $\mathrm{M}_p(\cdot;\boldsymbol{w})$ is $\frac{1}{r\sqrt[p]{\boldsymbol{w}_{\min}}}$-$\|\cdot\|_\infty$ Lipschitz.

While the $p \in [0,1)$ power-means are not Lipschitz continuous (see lemma 3.10 item 3), we find that they are still Hölder continuous, which largely results in similar enough properties for our purposes.

Lemma 3.13 (Power-Mean Hölder Continuity)**.** Suppose $\boldsymbol{u} \in [0,r]^g$, group index $i \in \mathcal{G}$, weights vector $\boldsymbol{w} \in \triangle_g$, and assume where appropriate that $\boldsymbol{u}_i + \varepsilon \leq r$. The power mean then obeys the following Hölder continuity criteria.
1) *Generic Welfare Hölder Condition*: Suppose $p \leq 1$. Then $|\mathrm{M}_p(\boldsymbol{u} + \varepsilon \mathbb{1}_i;\boldsymbol{w}) - \mathrm{M}_p(\boldsymbol{u};\boldsymbol{w})| \leq r^{1-\boldsymbol{w}_i}\varepsilon^{\boldsymbol{w}_i}$, and $\mathrm{M}_p(\cdot;\boldsymbol{w})$ is $r^{1-\boldsymbol{w}_{\min}}$-$\boldsymbol{w}_{\min}$-$\|\cdot\|_\infty$ Hölder continuous.
2) *Positive Welfare Hölder Condition*: Suppose $p \in (0,1]$. Then $|\mathrm{M}_p(\boldsymbol{u} + \varepsilon \mathbb{1}_i;\boldsymbol{w}) - \mathrm{M}_p(\boldsymbol{u};\boldsymbol{w})| \leq r^{1-p}\frac{\boldsymbol{w}_i}{p}\varepsilon^p$. Furthermore, $\mathrm{M}_p(\cdot;\boldsymbol{w})$ meets the following Hölder conditions:
A) $r^{1-p}\frac{\boldsymbol{w}_{\max}}{p}$-$p$-$\|\cdot\|_1$;
B) $r^{1-p}\frac{1}{p}$-p-$\mathrm{M}_1(|\cdot|;\boldsymbol{w})$; &
C) $r^{1-p}\frac{1}{p}$-p-$\|\cdot\|_\infty$.

Applications Understanding the response of power-mean functions to small changes to sentiment, i.e., their gradients and continuity properties, is highly relevant to *privacy*, *adversarial robustness*, *algorithmic stability*, *strategy proofness*, and *statistical learnability*. Due to their commonalities, we briefly treat the first four here, while sections 4 & 5 theoretically and experimentally explore statistical learning in detail. To clarify the relationship of our axioms to these properties, we visualize them in figure 3.

We first assume that the *parameters* or *decisions made* by some algorithm are robust to small changes to the objective (welfare), and note that Lipschitz or Hölder continuity describe how robust the objective is to small changes to sampled, estimated, or queried per-group utility values.[3] Lipschitz continuity is highly relevant to *differential privacy* [Bassily et al., 2019, 2020, Wang et al., 2022, Patel et al., 2022], as differential privacy is sensitive to changes to algorithm output caused by *individual-level changes*, thus sensitivity to infinitesimal change is paramount.

For *algorithmic stability*, *adversarial robustness*, and *strat-*

[3]Note that this applies to some classes of stable algorithms and objectives. For example, algorithms maximizing *strongly concave* welfare functions can only *boundedly change* the optimal parameters or output under bounded change to the objective function.

egy proofness, if *individual-level change* is too small to cause harm, Hölder continuity is powerful, since while *cumulative change* to utility is linear in the number of colluding parties, its *impact on welfare* is sublinear, due to the α-power. In other words, here individuals may be powerful, but the cumulative effect of collusion is less than the sum of its parts.

In the context of *algorithmic stability* for convex optimization, it is well-known that if the gradient is *bounded* and *Lipschitz continuous* (i.e., the function is *smooth*), we obtain desirable bounds on the generalization error and computational cost of (stochastic) gradient descent methods. Unfortunately, even if the underlying per-group sentiment functions are smooth with bounded gradient, these properties are not necessarily preserved by the power-mean (e.g., egalitarian objectives create nondifferentiabilities where the maximum or minimum sentiment group changes). However, more recent work [Lei and Ying, 2020] shows that a Hölder continuous gradient also suffices: in particular, it suffices to have $\|\nabla_\theta \mathrm{M}(\boldsymbol{u}(\theta); \boldsymbol{w}) - \nabla_\theta \mathrm{M}(\boldsymbol{u}(\theta'); \boldsymbol{w})\|_2 \leq \lambda \|\theta - \theta'\|_2^\alpha$. This condition is a bit more subtle than that which we consider, but observe that it may be bounded for appropriate model classes (which define the space of θ) using lemmata 3.10 & 3.13 and the classical chain rule from the calculus of infinitesimals.

Outside of some convenient median-based constructs, e.g., linear regression under absolute loss [Chen et al., 2018b], true strategy-proofness in learning settings can be quite difficult to achieve [Procaccia, 2008]. However, under the *ε-truthfulness* assumption [Meir and Rosenschein, 2011], agents only lie about their labels if they gain at least ε utility (or lose at least ε disutility) by doing so. In this setting, the above robustness analysis is relevant, as we have effectively shown that, assuming bounded loss values, Hölder continuity implies that small colluding groups have small impact on the objective, and thus (assuming, e.g., strong convexity in the neighborhood of the optimal $h^\star$) lying only weakly impacts $\hat{h}$. Therefore, if the hypothesis space $\mathcal{H}$ is *continuous* (e.g., bounded linear regression, but not discrete linear classification), they can not strongly benefit from doing so. We thus conclude that robustness and stability, which are preserved by Hölder continuous objectives, are sufficient for ε-truthful strategy-proofness in some learning settings. We thus conclude that, via lemmata 3.12 & 3.13, we can analyze the privacy, stability, and statistical properties of many welfare-based algorithms for arbitrary power-mean welfare functions.

4 Generalizing Fair-PAC Learning

Suppose now that we seek to estimate or optimize some welfare function, but do not know the utility values of each group, and must instead estimate them via sampling (i.e., from data). We predominantly study the *plug-in estimator*, which approximates the welfare of expected utility values with the welfare of *empirical mean* utilities over m samples from each group's instance distribution $\mathcal{D}_i$ over *labeled instance space* $(\mathcal{X} \times \mathcal{Y})$, i.e., features $x \in \mathcal{X}$ and labels $y \in \mathcal{Y}$. We assume a hypothesis class $\mathcal{H} \subseteq \mathcal{X} \to \mathcal{Y}'$ mapping inputs $\mathcal{X}$ to prediction space $\mathcal{Y}'$, and a utility function $\mathrm{u} : \mathcal{Y}' \times \mathcal{Y} \to \mathbb{R}_{0+}$ that assesses the quality of prediction $\hat{y} \in \mathcal{Y}'$ given true label $y \in \mathcal{Y}$, and we thus express the true utility of hypothesis $h \in \mathcal{H}$ for group i as $\mathbb{E}_{\mathcal{D}_i}[\mathrm{u} \circ h]$ and the empirical estimate of utility as $\hat{\mathbb{E}}_{\hat{\mathcal{D}}_i}[\mathrm{u} \circ h]$, where $(\mathrm{u} \circ h)(x, y) \doteq \mathrm{u}(h(x), y)$. Notation introduced in this section is summarized in table 2.

We first briefly outline computational concerns and build intuition for welfare optimization and FPAC learning in section 4.1. We then discuss asymptotic convergence and the fundamentals of statistical estimation of $\mathrm{W}(i \mapsto \mathbb{E}_{\mathcal{D}_i}[\mathrm{u} \circ h]; \boldsymbol{w})$ for a single function h in section 4.2. Section 4.3 then introduces and analyzes a concept of *fair-PAC learnability*, showing worst-case bounds on the sample complexity of estimating and optimizing welfare functions over a function family $\mathcal{H}$. Finally, section 4.4 discusses applying progressive sampling techniques to *adaptively query* to avoid worst-case sample complexity, and instead near-optimally sample to address a given optimization task.

4.1 On Compromise and Fair Learning

In standard supervised learning or risk minimization settings, given distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, we generally seek to approximate

$$h^\star \doteq \operatorname*{argmin}_{h \in \mathcal{H}} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \big[-(\mathrm{u} \circ h)(x, y)\big] \ ,$$

whereas in FPAC learning, we seek to approximate

$$h^\star \doteq \operatorname*{argmax}_{h \in \mathcal{H}} \mathrm{W}\left(i \mapsto \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_i} \big[(\mathrm{u} \circ h)(x, y)\big]; \boldsymbol{w}\right) \ .$$

The change of sign is entirely superficial, and standard first-order convex optimization techniques can be applied in either setting [Cousins, 2021], assuming appropriate structure of $\mathrm{u} \circ \mathcal{H} \doteq \{\mathrm{u} \circ h | h \in \mathcal{H}\}$ and the model parameters θ, but the impact of the welfare function and the per-group distributions on both $h^\star$ and the optimization dynamics are worth discussing.

Ideally, we would be able to optimize a model by making only local modifications that benefit all groups, but in general this is not realistic. Often decisions must be made during training that benefit some groups while harming others, which we term "disagreement." Disagreement arises *locally*, on the scale of individual changes made to the model during gradient updates, but also *globally*, in the sense that the overall model learned is a *compromise* between the wants and needs of each group. Some sources of disagreement are obvious, such as two groups disagreeing on how to classify a contiguous region of $\mathcal{X}$, but others are subtler. For

Object	Definition or Space	Description
$\mathcal{X}, \mathcal{Y}, \mathcal{Y}'$	—	Supervised prediction domain, label space, and codomain
h	$\in \mathcal{X} \to \mathcal{Y}'$	Hypothesis (possible supervised learning model)
$\mathcal{H}$	$\subseteq \mathcal{X} \to \mathcal{Y}'$	Hypothesis class (class of possible supervised learning models)
$\mathrm{u}(\hat{y}, y)$	$\in \mathcal{Y}' \times \mathcal{Y} \to [0, r]$	Utility function (score prediction $\hat{y}$ given true label y)
$\mathrm{u} \circ h$	$(\mathrm{u} \circ h)(x, y) \doteq \mathrm{u}(h(x), y), \in (\mathcal{X} \times \mathcal{Y}) \to [0, r]$	Utility composition
$\mathrm{u} \circ \mathcal{H}$	$\doteq \{\mathrm{u} \circ h \,\big\vert\, h \in \mathcal{H}\}, \subseteq (\mathcal{X} \times \mathcal{Y}) \to [0, r]$	Utility class (composition of utility function and hypothesis class)
$\mathcal{D}_i$	Over $(\mathcal{X} \times \mathcal{Y})$	Distribution over $(\mathcal{X}, \mathcal{Y})$ pairs pertaining to group i
m	$\in \mathbb{Z}_+$	Sample size

Table 2: Functions, spaces, and common variables in FPAC learning.

example, all finite learning models have learning capacity limitations, be they *finite neurons counts* in artificial neural networks, *norm constraints* on parameter vectors, or *depth bounds* in decision trees, and this paucity begets disagreement between groups, each of whom would prefer structures be learned that benefit themselves. Learners must balance between competing preferences that arise when groups have *direct disagreement*, i.e., conflicting preferences for $\mathcal{Y}$ given $\mathcal{X}$, or *indirect disagreement*, i.e., agreement on $\mathcal{Y}$ given $\mathcal{X}$, but disagreement as to what to *prioritize*, due to different distributions over $\mathcal{X}$.

It is enlightening to consider the process by which first-order optimization trains a model. Ideally, under complete agreement, each group would have the same response $f(\theta)$ in expected or empirical sentiment to the parameters θ. In this case, since $\mathrm{W}(i \mapsto f(\theta); \boldsymbol{w}) = f(\theta)$ (by axioms 6 & 7), it holds that $\nabla_\theta \mathrm{W}(i \mapsto f(\theta); \boldsymbol{w}) = \nabla_\theta f(\theta)$, thus optimization proceeds as in standard gradient ascent on expected utility. However, complete agreement is unrealistic, as different groups generally respond differently to different models.

Starting from an arbitrary parameterization θ_0, it is often possible to improve utility for all groups simultaneously by first focusing on aspects of the learning problem relevant to all groups (i.e., initially groups may agree on how to proceed in a mutually beneficial manner). However, once Pareto-dominance is achieved, we enter the realm of *disagreement*, as no decision will benefit any group without harming others. Observe now that for $p \leq_1 1$ welfare functions, the power-mean derivative $\frac{\boldsymbol{w}_i \boldsymbol{u}_i^{p-1}}{\mathrm{W}_p^{p-1}(\boldsymbol{u}; \boldsymbol{w})}$ (see lemma 3.10) *depends inversely* on group utility $\boldsymbol{u}_i$. Therefore, infinitessimal movement along the gradient favors equitable transfer from high-utility to low-utility groups, and gradient ascent steps discretely approximate this process.

The welfare-optimal model is thus an equilibrium between the forces of preferentially increasing low-utility group utilities and those utilities rising relative to their peers. This tension in training dynamics results in *compromise* between groups, and in some sense the whole process has more in common with *adversarial learning* than standard maximization tasks; indeed for $p = -\infty$, we have an explicit maximin objective, albeit one with a finite inner minimization set. For these reasons, the equilibrium model $h^\star$ that results from welfare maximization generally differs greatly from any of the models each group would select for themselves. Thus far, we have discussed only the *computational* aspects of welfare-based learning tasks, but the remainder of this work is focused on the *statistical aspect* of this setting.

4.2 The Fundamentals of Estimation

Before treating the intricacies of learning (optimization) over a *class of functions* $\mathcal{H}$, we first treat the subject of *estimating* utility values via sampling for a single function h, and we then discuss welfare estimates via the plug-in estimator. Namely, there exists some true expected utility vector $\boldsymbol{u} \doteq i \mapsto \mathbb{E}_{(x,y)\sim\mathcal{D}_i}[(\mathrm{u} \circ h)(x, y)]$, and we have some estimate $\hat{\boldsymbol{u}}$, and we ask the questions, "How well does $\hat{\boldsymbol{u}}$ approximate $\hat{\boldsymbol{u}}$?" and moreover, "How well does the plug-in welfare estimate $\mathrm{W}(\hat{\boldsymbol{u}}; \boldsymbol{w})$ approximate the true welfare $\mathrm{W}(\boldsymbol{u}; \boldsymbol{w})$?"

In estimation settings, we generally assume some *consistent estimator* $\hat{\boldsymbol{u}}$ of true utility $\boldsymbol{u} \in [0, r]^g$, i.e., some $\hat{\boldsymbol{u}}$ such that $\lim_{m\to\infty} \hat{\boldsymbol{u}} = \boldsymbol{u}$. For our purposes, the empirical mean over m samples per group will suffice. Here, for finite $\boldsymbol{u}$, using only *ε-δ limit continuity* (of welfare) and the *weak law-of-large-numbers*, we have *consistency* of the plug-in welfare estimate, i.e., $\lim_{m\to\infty} \mathrm{W}(\hat{\boldsymbol{u}}; \boldsymbol{w}) = \mathrm{W}(\boldsymbol{u}; \boldsymbol{w})$, but in the grand scheme of things, this is a rather weak guarantee. If we assume finite variance, then for any norm $\|\cdot\|_{\mathrm{W}}$ and any *failure probability* δ, by the central limit theorem, we have

$$\lim_{m\to\infty} \mathbb{P}_{\hat{\boldsymbol{u}}} \left(\tfrac{1}{\sqrt{m}} \|\boldsymbol{u} - \hat{\boldsymbol{u}}\|_{\mathrm{W}} \leq \sqrt{2v \ln \tfrac{1}{\delta}} \right) \geq 1 - \delta \ ,$$

for *variance proxy* $v \doteq \mathbb{V}_{\hat{s}}[\|\hat{s}\|_{\mathrm{W}}]$. In general, given some weighted welfare function $\mathrm{W}(\cdot; \boldsymbol{w})$, we wish to bound the *estimation error* $|\mathrm{W}(\hat{\boldsymbol{u}}; \boldsymbol{w}) - \mathrm{W}(\boldsymbol{u}; \boldsymbol{w})|$ of the plug-in estimator.

Table 3 precisely characterizes the asymptotic convergence rates of estimation error under each continuity concept studied in this work. We find root-hyperbolic rates under Lipschitz continuity, and $\frac{\alpha}{2}$ power-law rates under Hölder continuity, though in all cases, the variance proxy v (and thus the norm $\|\cdot\|_{\mathrm{W}}$) also plays a substantial role.

Despite the above positive results, significant challenges arise in estimating the welfare of even a single h. While

Continuity Concept	Deviation Bound $\lvert\mathrm{W}(\hat{\boldsymbol{u}};\boldsymbol{w})-\mathrm{W}(\boldsymbol{u};\boldsymbol{w})\rvert\leq$	Asymptotic Error Convergence Rate	Sample Complexity $\min\{m\in\mathbb{Z}_+ \mid \mathbb{P}(\lvert\mathrm{W}(\hat{\boldsymbol{u}};\boldsymbol{w})-\mathrm{W}(\boldsymbol{u};\boldsymbol{w})\rvert>\varepsilon)<\delta\}$
ε-δ Limit	—	$\lim_{m\to\infty}\varepsilon=0$	$\mathrm{m}_{\mathrm{W}}(\varepsilon,\delta)<\infty$
λ-α Hölder	$\lambda\lVert\boldsymbol{u}-\hat{\boldsymbol{u}}\rVert_{\mathrm{M}}^{\alpha}$	$\varepsilon\lesssim\lambda\left(\frac{2v\ln\frac{1}{\delta}}{m}\right)^{\alpha/2}$	$\mathrm{m}_{\mathrm{W}}(\varepsilon,\delta)\lesssim\frac{2\lambda^{2/\alpha}v\ln\frac{1}{\delta}}{\varepsilon^{2/\alpha}}$
λ Lipschitz	$\lambda\lVert\boldsymbol{u}-\hat{\boldsymbol{u}}\rVert_{\mathrm{M}}$	$\varepsilon\lesssim\lambda\sqrt{\frac{2v\ln\frac{1}{\delta}}{m}}$	$\mathrm{m}_{\mathrm{W}}(\varepsilon,\delta)\lesssim\frac{2\lambda^{2}v\ln\frac{1}{\delta}}{\varepsilon^{2}}$
Nonexpansive	$\lVert\boldsymbol{u}-\hat{\boldsymbol{u}}\rVert_{\mathrm{M}}$	$\varepsilon\lesssim\sqrt{\frac{2v\ln\frac{1}{\delta}}{m}}$	$\mathrm{m}_{\mathrm{W}}(\varepsilon,\delta)\lesssim\frac{2v\ln\frac{1}{\delta}}{\varepsilon^{2}}$

Table 3: Continuity concepts and asymptotic estimation guarantees. For each type of continuity studied herein, we bound the *estimation error* $\varepsilon=|\mathrm{M}-\hat{\mathrm{M}}|$, as well as the *sample complexity* $\mathrm{m}_{\mathrm{W}}(\varepsilon,\delta)$, which is the number of samples required to attain ε estimation error with probability at least $1-\delta$. Here $a\lesssim b$ denotes *asymptotic inequality*, which can be formalized as $\lim_{m\to\infty}\frac{a(m)}{b(m)}\leq 1$ or $\lim_{\varepsilon\to 0^+}\frac{a(\varepsilon)}{b(\varepsilon)}\leq 1$. The asymptotic approximation stems from the central limit theorem, and upper-bounds are due to continuity-based deviation bounds and Gaussian-Chernoff tail bounds.

the plug-in welfare estimator preserves the *consistency* of the utility estimator $\hat{\boldsymbol{u}}$, no unbiased estimator for welfare functions (unless $p=1$ or stringent assumptions are made on the distributions $\mathcal{D}_{1:g}$) essentially because $\mathrm{W}(\boldsymbol{u};\boldsymbol{w})$ is a nonlinear operator. There are also deep questions as to how to optimally allocate sampling effort, since the variances of each $\hat{\boldsymbol{u}}_i$ may differ, as may the impact of each $\hat{\boldsymbol{u}}_i$ on $\mathrm{W}(\hat{\boldsymbol{u}};\boldsymbol{w})$, both of which directly impact the variance proxy v. These questions are probed in detail by Cousins [2022], but in this work, we generally assume m samples for each group, which greatly simplifies the matter, and never requires more than a factor g more samples than any nonuniform allocation of sampling effort.

The above analysis is straightforward, but it does not scale well from a single h to learning over a large hypothesis class $\mathcal{H}$, and moreover the behavior of the quantity v depends intimately on $\|\cdot\|_{\mathrm{W}}$ and the sampling distributions $\mathcal{D}_{1:g}$. Going forward, we shall bound the generalization error and sample complexity of welfare estimation by first separately bounding each group's utility values, and then taking a union bound over groups. This technique introduces $\ln\frac{g}{\delta}$ terms instead of the $\ln\frac{1}{\delta}$ terms of our central limit theorem argument (see table 3), but it allows us to express our results in terms of readily-available generalization error and sample complexity bounds for per-group (scalar) utilities.

4.3 Fair-PAC Learning and Learnability

We first show that the estimation error of welfare $\mathrm{W}(\boldsymbol{u};\boldsymbol{w})$ may be bounded in terms of the error of each utility value $\boldsymbol{u}_i$. From there, we bound the sample complexity of optimization, and describe a notion of fair-PAC (FPAC) learnability for welfare functions, wherein the goal is to uniformly bound the number of samples required to learn in $\mathcal{H}$.

We abstract away the statistical details of this estimation process by assuming, for each group i, a known bound on the *supremum deviation* of the expected utility for each $h\in\mathcal{H}$, i.e., a bound of the form

$$\forall i:\sup_{h\in\mathcal{H}}\left|\mathop{\mathbb{E}}_{\mathcal{D}_i}[\mathrm{u}\circ h]-\mathop{\hat{\mathbb{E}}}_{\hat{\mathcal{D}}_i}[\mathrm{u}\circ h]\right|\leq\boldsymbol{\varepsilon}_i\ . \tag{3}$$

The details of obtaining such bounds with high probability are well-studied, and Cousins [2022] discusses them under the name *additive error vector* (AEV) bounds in group-fairness settings, showing that they can be obtained via the Chernoff method [Bennett, 1962, Hoeffding, 1963, Boucheron et al., 2013], Rademacher averages [Bartlett and Mendelson, 2002, Shalev-Shwartz and Ben-David, 2014, Mitzenmacher and Upfal, 2017, Cousins and Riondato, 2020], or other such tools.

In particular, using Rademacher averages, bounds on the supremum deviation often take the form

$$\mathbf{O}\left(\frac{r\ln\frac{g}{\delta}}{m}+\sqrt{\frac{C+r^2\ln\frac{g}{\delta}}{m}}\right)\ ,$$

where C depends on $\mathcal{D}_i$ and $\mathcal{H}$, and can generally be bounded as $v\ln n$, where $v\doteq\sup_{h\in\mathcal{H}}\mathbb{V}_{\mathcal{D}}[\mathrm{s}\circ h]$ is the *supremum variance* of utility over the class $\mathcal{H}$, and n measures the *effective size* of the hypothesis class. The details are beyond the scope of this work, but $n=|\mathcal{H}|$ suffices.[4] Observe that here v plays the same fundamental role as in single-function estimation (as discussed in section 4.2), n essentially corrects for the *multiple comparisons problem* (i.e., our bound holds across all of $\mathcal{H}$ simultaneously) with similar form to union bounds, and the $\mathbf{O}\frac{r\ln\frac{1}{\delta}}{m}$ term acts as an (asymptotically negligible) *finite-sample correction* to a tail bound that otherwise resembles a Gaussian-Chernoff bound.

With the table set, we now bound the estimation error of welfare objectives. The following result holds essentially by the definition of Hölder continuity, and can immediately be applied to any model class $\mathcal{H}$ for which bounds on the supremum deviation are available.

[4] Furthermore, taking n to bound the cardinality of projecting $\mathrm{u}\circ\mathcal{H}$ onto $\boldsymbol{m}_i$ samples also suffices (via, e.g., the Vapnik-Chervonenkis dimension), and more sophisticated metrics of effective size, such as *covering numbers*, i.e., the smallest cover size n such that there exists some $\mathcal{H}'\subseteq\mathcal{X}\to\mathcal{Y}$ such that $|\mathcal{H}'|=n$ and each $h\in\mathcal{H}$ can be well-approximated by some $h'\in\mathcal{H}'$.

Theorem 4.1 (Hölder Continuity and Welfare Optimality)**.** Suppose $\mathrm{W}(\cdot;\boldsymbol{w})$ is λ-α-$\|\cdot\|_{\mathrm{W}}$ Hölder continuous w.r.t. some norm $\|\cdot\|_{\mathrm{W}}$, and additive error bounds $\boldsymbol{\varepsilon}$ that obey (3). Then

$$\sup_{h\in\mathcal{H}}\left|\mathrm{W}\left(i\mapsto\mathop{\mathbb{E}}_{\mathcal{D}_i}[\mathrm{u}\circ h];\boldsymbol{w}\right)-\mathrm{W}\left(i\mapsto\mathop{\hat{\mathbb{E}}}_{\hat{\mathcal{D}}_i}[\mathrm{u}\circ h];\boldsymbol{w}\right)\right|\leq\lambda\|\boldsymbol{\varepsilon}\|_{\mathrm{W}}^{\alpha}.$$

Consequently, the *empirical welfare maximizer*

$$\hat{h}\doteq\sup_{h\in\mathcal{H}}\mathrm{W}\left(i\mapsto\mathop{\hat{\mathbb{E}}}_{\hat{\mathcal{D}}_i}[\mathrm{u}\circ h];\boldsymbol{w}\right)$$

approximates the *true welfare maximizer*

$$h^{\star}\doteq\sup_{h^{\star}\in\mathcal{H}}\mathrm{W}\left(i\mapsto\mathop{\mathbb{E}}_{\mathcal{D}_i}[\mathrm{u}\circ h^{\star}];\boldsymbol{w}\right)\ ,$$

in terms of welfare-optimality, as it holds that

$$\mathrm{W}\left(i\mapsto\mathop{\mathbb{E}}_{\mathcal{D}_i}[\mathrm{u}\circ\hat{h}];\boldsymbol{w}\right)\geq\mathrm{W}\left(i\mapsto\mathop{\mathbb{E}}_{\mathcal{D}_i}[\mathrm{u}\circ h^{\star}];\boldsymbol{w}\right)-2\lambda\|\boldsymbol{\varepsilon}\|_{\mathrm{W}}^{\alpha}\ .$$

This result is easily interpreted and practically relevant. Furthermore, it is readily applied to any power-mean welfare function via lemmata 3.12 & 3.13, whereas prior work [Cousins, 2021, 2022] only handles the case of Lipschitz-continuous aggregator functions.

In PAC-learning theory, it is standard to analyze the difficulty of learning as a function of the *complexity of the model class* $\mathcal{H}$. To this end, we assume $\mathcal{H}$ is parameterized by D variables $\boldsymbol{d}\in\mathbb{R}_{0+}^{D}$, e.g., the dimension of a linear classifier [Shalev-Shwartz and Ben-David, 2014], or for neural network families, a vector of per-layer widths [Anthony and Bartlett, 2009] or per-layer norm constraints [Bartlett et al., 2017], where $\mathcal{H}_{\boldsymbol{d}}$ denotes the class parameterized by $\boldsymbol{d}$.

Furthermore, we seek to know *how much data* we need to probabilistically learn an objective to within a given error tolerance ε, rather than *how well* we can learn a concept with a given sample size. Henceforth, a *sample complexity function* $\mathrm{m}_{\mathcal{H}}(\varepsilon,\delta,r,\boldsymbol{d})$ is some function such that, for utility range r, any probability distribution $\mathcal{D}$, a sample size of at least $\mathrm{m}_{\mathcal{H}}(\varepsilon,\delta,r,\boldsymbol{d})$ ensures that $\mathrm{u}\circ\mathcal{H}$ is *uniformly convergent*, i.e., with probability at least $1-\delta$, it holds that

$$\sup_{h\in\mathcal{H}_{\boldsymbol{d}}}\left|\mathop{\mathbb{E}}_{\mathcal{D}}[\mathrm{u}\circ h]-\mathop{\hat{\mathbb{E}}}_{\hat{\mathcal{D}}}[\mathrm{u}\circ h]\right|\leq\varepsilon\ . \tag{4}$$

Similarly, we express the (per-group) sample complexity of uniformly estimating a *welfare function* $\mathrm{W}(\cdot;\boldsymbol{w})$ as $\mathrm{m}_{\mathrm{W},\mathcal{H}}(\varepsilon,\delta,g,r,\boldsymbol{d})$, requiring with probability at least $1-\delta$, it holds that

$$\sup_{h\in\mathcal{H}_{\boldsymbol{d}}}\left|\mathrm{W}\left(i\mapsto\mathop{\mathbb{E}}_{\mathcal{D}_i}[\mathrm{u}\circ h];\boldsymbol{w}\right)-\mathrm{W}\left(i\mapsto\mathop{\hat{\mathbb{E}}}_{\hat{\mathcal{D}}_i}[\mathrm{u}\circ h];\boldsymbol{w}\right)\right|\leq\varepsilon\ . \tag{5}$$

We now analyze the sample complexity of welfare estimation, and subsequently FPAC learning with welfare objectives.

Theorem 4.2 (Welfare Sample Complexity)**.** Suppose sample complexity function $\mathrm{m}_{\mathcal{H}}(\varepsilon,\delta,r,\boldsymbol{d})$ for hypothesis class $\mathcal{H}$, and some welfare function $\mathrm{W}(\cdot;\boldsymbol{w})$ that is λ-α-$\|\cdot\|_{\infty}$ Hölder continuous. Then the sample complexity function

$$\mathrm{m}_{\mathrm{W},\mathcal{H}}(\varepsilon,\delta,g,r,\boldsymbol{d})\leq\mathrm{m}_{\mathcal{H}}\left(\sqrt[\alpha]{\tfrac{\varepsilon}{\lambda}},\tfrac{\delta}{g},r,\boldsymbol{d}\right)$$

is sufficient, i.e., for at least this many samples from each of the g groups, (5) holds. Moreover, for this sample size, with probability at least $1-\delta$, the empirical welfare maximizer is 2ε-optimal.

From these uniform generalization error and sample complexity bounds, we can show that *classes of welfare functions* are FPAC learnable, defined as follows.

Definition 4.3 (Fair-PAC Learning)**.** Suppose *hypothesis class* $\mathcal{H}\subseteq\mathcal{X}\to\mathcal{Y}'$ parameterized by $\boldsymbol{d}\in\mathbb{R}_{0+}^{D}$, *utility function* $\mathrm{u}\colon\mathcal{Y}'\times\mathcal{Y}\to\mathbb{R}_{0+}$, and *welfare class* $\mathcal{W}\subseteq\mathbb{R}_{0+}^{g}\to\mathbb{R}_{0+}$. We say $\mathcal{H}$ is *fair-PAC-learnable* w.r.t. u and $\mathcal{W}$ if there exists an algorithm $\mathcal{A}$ and sample complexity function $\mathrm{m}_{\mathcal{W},\mathcal{H}}$ such that for all
1) class parameterizations $\boldsymbol{d}$;
2) group counts g;
3) per-group instance distributions $\mathcal{D}_{1:g}$, each over $(\mathcal{X}\times\mathcal{Y})$;
4) (weighted) welfare concepts $\mathrm{W}(\cdot;\boldsymbol{w})$ in $\mathcal{W}$;
5) additive approximation errors $\varepsilon>0$; &
6) failure probabilities $\delta\in(0,1)$;
it holds that $\mathcal{A}$ can identify a hypothesis $\hat{h}\in\mathcal{H}_{\boldsymbol{d}}$, i.e., $\hat{h}\leftarrow\mathcal{A}(\mathcal{D}_{1:g},\mathrm{W},\varepsilon,\delta,\boldsymbol{d})$, such that
1) for each group, $\mathcal{A}(\mathcal{D}_{1:g},\mathrm{W},\varepsilon,\delta,\boldsymbol{d})$ draws no more than $\mathrm{m}_{\mathcal{W},\mathcal{H}}(\varepsilon,\delta,\mathrm{W},g,\boldsymbol{d})$ samples; &
2) with probability at least $1-\delta$ (over randomness of $\mathcal{A}$ and sampling), $\hat{h}$ obeys

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

Furthermore, if $\mathrm{m}_{\mathcal{W},\mathcal{H}}(\varepsilon,\delta,\mathrm{W},g,\boldsymbol{d})$ can be uniformly bounded for any $\mathrm{W}(\cdot;\boldsymbol{w})\in\mathcal{W}$, then we say that $\mathcal{H}$ is *uniformly PAC learnable* over $\mathcal{W}$ w.r.t. u.

With trivial changes to convert the maximization objective to a minimization objective, this definition can also be applied to loss functions and classes of malfare functions. In particular, this definition generalizes the FPAC concept given by Cousins [2021], which was specified for the class of all malfare functions satisfying a set of axioms corresponding to $p\geq1$ weighted power-means. We also relax the definition to allow the sample complexity function to depend on the (weighted) welfare function $\mathrm{W}(\cdot;\boldsymbol{w})\in\mathcal{W}$, but our concept of *uniform FPAC-learnability* strictly generalizes that of Cousins [2021].

Theorem 4.4 (Characterizing FPAC Learnability)**.** Suppose some weighted power-mean welfare function $\mathrm{W}_p(\cdot;\boldsymbol{w})$, utility function u with range r, and hypothesis class $\mathcal{H}$ with sample complexity function $\mathrm{m}_{\mathcal{H}}(\varepsilon,\delta,r,\boldsymbol{d})\in\mathrm{Poly}(\frac{1}{\varepsilon},\log\frac{1}{\delta},r,\boldsymbol{d})$. We then bound the sample complexity $m\doteq\mathrm{m}_{\mathcal{W},\mathcal{H}}(\varepsilon,\delta,\mathrm{W},g,\boldsymbol{d})$ of FPAC learning $\mathcal{H}$ w.r.t. welfare class $\mathcal{W}\doteq\{\mathrm{W}(\cdot;\boldsymbol{w})\}$ as
1) $m\leq\mathrm{m}_{\mathcal{H}}\left(\sqrt[\alpha]{\frac{\varepsilon}{2\lambda}},\frac{\delta}{g},r,\boldsymbol{d}\right)\in\mathrm{Poly}\left(\sqrt[\alpha]{\lambda},\frac{1}{\sqrt[\alpha]{\varepsilon}},\log\frac{1}{\delta},\log g,r,\boldsymbol{d}\right)$;
2) $p\in(0,1]\Rightarrow m\in\mathrm{Poly}\left(\sqrt[p]{r},\frac{1}{\sqrt[p]{p}},\frac{1}{\sqrt[p]{\varepsilon}},\log\frac{1}{\delta},\log g,\boldsymbol{d}\right)$;
3) $p=0\Rightarrow m\in\mathrm{Poly}\left(\sqrt[\boldsymbol{w}_{\min}]{r},\frac{1}{\boldsymbol{w}_{\min}\sqrt[\boldsymbol{w}_{\min}]{\varepsilon}},\log\frac{1}{\delta},\log g,\boldsymbol{d}\right)$;

4) $p < 0 \Rightarrow m \in \mathrm{Poly}\left(\frac{1}{\varepsilon}, \frac{1}{\sqrt[p]{\boldsymbol{w}_{\min}}}, \log\frac{1}{\delta}, \log g, r, \boldsymbol{d}\right)$; & 5) for any $c \in (0,1)$, if $|p| \geq c$ and group weights obey the *nonnegligibility condition* $\boldsymbol{w}_{\min} \geq \frac{c}{g}$, then $m \in \mathrm{Poly}^{\frac{1}{c}}\left(\frac{1}{c}, \frac{1}{\varepsilon}, g, \log\frac{1}{\delta}, r, \boldsymbol{d}\right)$.

Observe that the subfamilies of power-mean welfare functions considered in items 2–4 are induced by specific axiomatic choices. In particular, item 2 follows from SM0 (axiom 12), item 3 follows from either IIA or from 0B (axiom 15) and SM0, and item 4 follows from either Lipschitz continuity and SWTP (axiom 13; to prevent $p = 1$) or from FC (axiom 16) — see figure 3 for visual explication. Similarly, item 5 follows by assuming *nonnegligibility* of weights, which holds, e.g., for *unweighted* aggregator functions, and bounding p away from 0, which may be accomplished in a variety of ways. We thus conclude that the axiomatic choices made to restrict the space of welfare functions directly impact their FPAC-learnability.

On Uniform and Polynomial FPAC-Learnability The bounds of theorem 4.4 items 1–4 imply FPAC learnability, but not uniform FPAC learnability, and exponential dependencies on α, $\frac{1}{\boldsymbol{w}_{\min}}$, or $\frac{1}{|p|}$ do appear. It is only in item 5, for any constant c, that the class is uniformly-FPAC-learnable. In contrast, using only Lipschitz continuity, it is straightforward to show that the entire class of *all fair malfare functions*, i.e., any $\mathrm{M}_p(\cdot;\boldsymbol{w})$ for $p \geq 1$, for which $\alpha = 1$ by lemma 3.12 item 1, is uniformly FPAC learnable.

In general, if α is bounded away from 0, and $\mathrm{m}_{\mathcal{H}}(\varepsilon, \delta, r, \boldsymbol{d})$ is polynomial, then the uniform sample complexity of FPAC learning is also polynomial in all parameters, and thus FPAC learning in some sense *preserves polynomial learnability*.We thus conclude that FPAC learning with malfare concepts is easier than FPAC learning with welfare concepts, however under appropriate axiomatically-motivated conditions, the gap in sample complexity between the two settings is polynomial.

4.4 Sample-Efficient Learning and Estimation with Progressive Sampling

Our study of FPAC learnability is based on the Lipschitz and Hölder continuity of power means, which only coarsely describe their behavior, yielding *worst-case* sample complexity bounds. Such bounds may be improved if *a priori* knowledge regarding relative (dis)utility values is available; for example, under near-equality, power-mean functions behave roughly linearly, and may thus be much easier to estimate or optimize. We now briefly voyage into the world of *progressive sampling* to show that, even without such *a priori* knowledge, efficient learning algorithms can adapt their sample consumption to the *inherent difficulty* of the specific task at hand, which may be substantially less than the *worst-case* sample complexity bounds of theorem 4.4 would suggest.

Without considering the delicate intricacies of probabilistic reasoning, one might naïvely assume they could iteratively draw one sample per group, and terminate when the welfare objective is uniformly estimated or approximately optimized. Unfortunately, this quickly runs into statistical errors via the *multiple comparisons problem*, as the sampling process is inherently probabilistic. Efficient progressive sampling methods take this basic idea and account for these issues, but rather than incrementing the sample size at each step, they instead increase the sample size *geometrically*. Such methods have had great impact in myriad settings, including statistical data science [Riondato and Upfal, 2015, 2018, Cousins et al., 2020, 2023d], where estimators query *a single distribution*, empirical game theoretic analysis [Viqueira et al., 2020, 2021, Cousins et al., 2022b, Mishra et al., 2022, Cousins et al., 2023a], where estimators query a noisy utility oracle at *strategy profiles* to bound various game-theoretic properties, and fair machine learning [Cousins, 2022], wherein fair objectives on model classes are estimated and optimized by sampling from *group-specific distributions*.

At a glance, assume our sample complexity bounds scale as $\boldsymbol{\Theta}(\log\frac{1}{\delta})$. Consequently, a schedule of length T with *uniformly allocated* δ (i.e., one that considers up to T sample sizes, and takes probability $1 - \frac{\delta}{T}$ tail bounds at each) can overshoot the sufficient sample size by a constant factor (due to geometric spacing), and furthermore would need a factor $\mathbf{O}(\log T)$ excess samples to correct for the multiple comparisons problem. However, aside from these factors, it is otherwise as sharp as knowing the (task-specific) minimum sufficient sample size *a priori*. Therefore, in cases where Hölder analysis only loosely bounds sample complexity, progressive sampling methods can still adaptively consume about as many samples as are actually required for the task at hand.

The question remains, "How long must the sampling schedule be?" In other words, "How large must T be to guarantee a sufficient sample size is reached?" Here our sample complexity bounds prove invaluable: a geometric sampling schedule must have length logarithmic in the ratio of maximum to minimum sufficient sample sizes, both of which are $\boldsymbol{\Theta}(\log\frac{Tg}{\delta})$, thus solving for a minimal sufficient T is straightforward, e.g., with Hoeffding's inequality, a doubling schedule admits

$$T = \left\lceil \log_2 \frac{\lceil \frac{1}{2}\left(\frac{2r\lambda}{\varepsilon}\right)^{\frac{2}{\alpha}} \ln\frac{2|\mathcal{H}|Tg}{\delta} \rceil}{\lceil \frac{1}{2}\ln\frac{2|\mathcal{H}|Tg}{\delta}\rceil} \right\rceil \in \boldsymbol{\Theta}\left(\tfrac{1}{\alpha}\log\tfrac{r\lambda}{\varepsilon}\right)$$

via theorem 4.4 item 1. In general, this progressive sampling strategy induces an *overhead cost factor* of

$$\mathbf{O}\log\log T \subseteq \mathbf{O}\log\log \frac{\mathrm{m}_{\mathrm{W},\mathcal{H}}\left(\sqrt[\alpha]{\frac{\varepsilon}{\lambda}}, \frac{\delta}{T}, g, r, \boldsymbol{d}\right)}{\log\frac{Tg}{\delta}} \subset \log\mathrm{Poly}\left(\tfrac{1}{|p|}, \tfrac{1}{\boldsymbol{w}_{\min}}, \log\mathrm{Poly}\left(\tfrac{1}{\varepsilon}, g, r, \boldsymbol{d}\right)\right) \quad (6)$$

relative to the (unknown) task-specific sufficient sample size. Note that in (6), even terms *exponential* in $\frac{1}{|p|}$ and $\frac{1}{\boldsymbol{w}_{\min}}$ in

the FPAC sample complexity bound become *logarithmic*, due to the double-logarithm. Thus while welfare functions may be *inherently difficult* to estimate, the *statistical overhead* of progressive sampling, as compared to drawing a task-specific sufficient sample, is quite negligible.

5 Experiments

To demonstrate practical relevance, we present a synthetic experiment in a *welfare maximization 1-armed bandit* setting, and study the sample complexity and sensitivity of welfare estimates to various parameters. In particular, we assume each pull of the bandit arm gives a single *utility sample* for each group, and from *empirical mean utilities* $\hat{\boldsymbol{u}}$, we wish to estimate the welfare $\mathrm{W}(\boldsymbol{u}; \boldsymbol{w})$ as $\mathrm{W}(\hat{\boldsymbol{u}}; \boldsymbol{w})$. Note that this is a key step towards regret-optimally selecting among k arms.

We assume groups $\{\boldsymbol{g}_1, \boldsymbol{g}_2\}$, where $\boldsymbol{g}_1$ is the majority group and $\boldsymbol{g}_2$ is the minority group, and the weight and/or expected utility of $\boldsymbol{g}_1$ is no less than that of $\boldsymbol{g}_2$. In this experiment, utility samples are $\textsc{Uniform}(\boldsymbol{u}_i - \frac{1}{2}, \boldsymbol{u}_i + \frac{1}{2})$ i.i.d. random variables (mean $\boldsymbol{u}_i$) for each group i. Figure 4 varies the key parameters of welfare p, minority mean $\boldsymbol{u}_2$, minority weight $\boldsymbol{w}_2$, and sample size m, in order to study welfare estimation around the particularly challenging $p \approx 0$ and $\boldsymbol{w}_{\min} \approx 0$ domains. Here empirical utilities $\hat{\boldsymbol{u}}_i$ have variance $\frac{1}{12m}$, but similar results are shown for *Bernoulli* and *beta* noise models with more complicated variance structure in appendix B. We present the true welfare and approximate Gaussian $\pm 1\sigma$ (68.27%) confidence intervals on empirical welfare, i.e., $\mathrm{W}_p(\boldsymbol{u}; \boldsymbol{w}) \pm \lambda(\frac{1}{12m})^{\frac{\alpha}{2}}$, where λ, α are as in lemmata 3.12 & 3.13. Using 5000 trials over sampled utilities, we also plot average empirical welfare, and a 68.27% empirical confidence band on empirical welfare, i.e., $\mathrm{W}_p(\hat{\boldsymbol{u}}; \boldsymbol{w})$.

Note that our theoretical Gaussian confidence intervals should ideally contain *at least* 68.27% of the empirical welfare samples (i.e., the empirical band), and indeed this is the case, despite the error due to sampling and the Gaussian approximation. We don't necessarily expect the empirical 68.27% confidence intervals to contain the true welfare, although this does always occur in these experiments. Moreover, empirical confidence bands may be substantially smaller than Gaussian confidence bands, as they are estimated via sampling, and thus not susceptible to the looseness of Lipschitz and Hölder continuity bounds. We see exactly this, and the gap between the empirical and theoretical confidence bounds varies with the parameters.

Figure 4a shows the impact of changing p on the welfare. Observe that $\mathrm{W}_p(\cdot; \boldsymbol{w})$ is monotonic in p, and it is thus no surprise that all measures of welfare are increasing in p. The interesting portion of the experiment is that both the 68.27% approximate Gaussian and empirical confidence intervals are very wide for $p \approx 0$, and narrow as $|p|$ increases. This concords with the theory of section 3.4, as despite the small variances of per-group utility estimates, $\mathrm{W}_p(\cdot; \boldsymbol{w})$ for $p \approx 0$ remains difficult to estimate, due to high sensitivity to near-0 utility values. Observe that the Hölder and Lipschitz analyzes are not sharp, particularly around $p = 0$, because $\boldsymbol{u}_2 > 0$, and indeed the approximate Gaussian confidence bands contain the empirical percentiles.[5] Figure 4b then varies the minority utility $\boldsymbol{u}_2$, and we find that as $\boldsymbol{u}_2 \to 0$, empirical confidence intervals sharply diverge, due to high sensitivity to minimum utility, i.e., $\hat{\boldsymbol{u}}_2 \approx 0$. In figure 4c, we vary the weight of the minority group $\boldsymbol{w}_2$, and find extremely wide confidence intervals as $\boldsymbol{w}_2 \to 0$, since $\mathrm{W}_0(\hat{\boldsymbol{u}}; \boldsymbol{w})$ is very sensitive to $\hat{\boldsymbol{u}}_2 \approx 0$ when $\boldsymbol{w}_2 \approx 0$, but as $\boldsymbol{w}_2 \to \frac{1}{2}$, the estimate of welfare becomes much more stable, as higher $\boldsymbol{w}_2$ means smaller overall welfare, but less sensitivity to $\hat{\boldsymbol{u}}_2$.

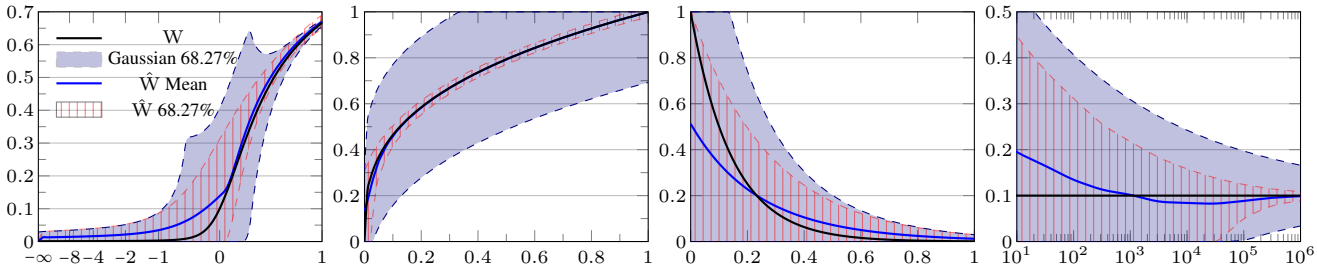
Figures 4a–4c all show the pathologically large estimation error of welfare functions, which when left unchecked causes models to *overfit* to disadvantaged groups, then exhibit bias against them when applied *ex vitro*. In figure 4d, we study a mitigation to this problem, by observing the impact of sample size m on $\mathrm{W}_0(\hat{\boldsymbol{u}}; \boldsymbol{w})$. Note that by theorem 4.2 the sample complexity of $\mathrm{W}_0(\cdot; \boldsymbol{w})$ estimation is $\mathbf{O}(\frac{1}{\varepsilon})^{\frac{2}{\boldsymbol{w}_2}}$, whereas for $\mathrm{W}_1(\cdot; \boldsymbol{w})$ it is only $\mathbf{O}(\frac{\boldsymbol{w}_2}{\varepsilon})^2$, and this asymptotically larger sample complexity is visually manifest as slower convergence rates for both the Gaussian and empirical confidence intervals. In all cases, we conclude that, as the theory suggests, an understanding of the continuity properties of power-mean functions is crucial to understanding the sample complexity and estimation error of practical welfare objectives.

6 Conclusion

We show an alternative axiomatic basis for fair aggregator functions, which we argue is simpler, weaker, and more fundamental than prior art. We also draw interdisciplinary connections to moral philosophy and econometric theory to establish stronger axioms, which intuitively guide modellers on fair objective selection, and theoretically distinguish between natural classes of welfare functions. In particular, our (strict) *weak transfer principle*, *zero barrier*, and *finite ceiling* axioms strengthen arguments for prioritarian (i.e., more egalitarian than utilitarian) fairness concepts by assuming less and/or concluding more, and our axioms handling group weights $\boldsymbol{w}$ simplify existing theory.

We then perform a detailed analysis of the Lipschitz and Hölder continuity of classes of power-mean welfare functions that satisfy our axioms. In particular, we find that our extended axioms naturally partition the class of power-mean functions into classes, each of which share Lipschitz or Hölder continuity properties, which is visually depicted in

[5] Still, the Lipschitz continuity bounds used for sufficiently negative $p < 0$ closely follow the Lipschitz constants plotted in figure 2. Thus while Lipschitz continuity describes the behavior of power means reasonably well here, our Hölder continuity analysis is needed to understand behavior about $p \approx 0$ and for $p > 0$.



(a) Welfare vs. $p \in [-\infty, 1]$ (b) Welfare vs. minority mean $\boldsymbol{u}_2$ (c) Welfare vs. minority weight $\boldsymbol{w}_2$ (d) Welfare vs. sample size m

Figure 4: Estimating the Welfare of a 1-Armed Bandit with Uniform Noise. Each plot studies the response of welfare to one parameter, and the remaining parameters are selected from $p = 0$, $\boldsymbol{u} = \langle 0.999, 0.001 \rangle$, $\boldsymbol{w} = \langle \frac{2}{3}, \frac{1}{3} \rangle$, and $m = 100$. All axes are linear, except 4a, which plots $p \in [-\infty, 1]$ by transforming $x = \frac{1}{\pi} \arctan(1 - p)$, and 4d, which is logarithmic in x.

figure 3. We also argue that Lipschitz continuity itself may be viewed as a form of *stability* axiom, with consequent resistance to *minority rule*. We follow with a discussion of applications in privacy, algorithmic stability, adversarial robustness, and strategy proofness, finding that Lipschitz and/or Hölder continuity of welfare are often sufficient to show these properties, and we later experimentally study the relationship between choice of axioms, welfare function, data distributions, and the difficulty of estimation.

Finally, we generalize the concept of fair-PAC learning to arbitrary families of welfare functions. We then show conditions under which fair-PAC learning welfare objectives has polynomial sample complexity, and is nearly as efficient as fair-PAC learning malfare objectives, improving the state of the art in utility-based and econometric learning settings. Moreover, prior work handles only the continuity and sample-complexity analysis of the $p \geq 1$ case (malfare), and we show that while the $p < 1$ case is more challenging, the difficulty of learning actually increases smoothly as $p \to 0$ *from both directions*, yielding intuitive, mathematically rigorous, and practically actionable understanding of learning and estimation problems over the entire power-mean spectrum. Furthermore, specific *axiomatic choices* regarding the class of welfare functions specify *discrete classes* with interpretable properties and desirable fair-PAC learning guarantees, thus establishing a hierarchy of fair learning settings.

We hope also that these results will be generalized to related settings. For instance, similar analysis is clearly beneficial in *regret minimization* (over groups), also termed "*multi-group agnostic PAC learning*" [Blum and Lykouris, 2020, Rothblum and Yona, 2021, Cousins, 2022], wherein the task is to minimize malfare of *differences between* utility or risk of each group's preferred model and some compromise model. We envision similar applications in *improvement maximization*, wherein we seek to maximize the welfare of *differences between* utility or risk of the learned compromise model and some fixed reference model, which addresses issues raised by Thomas et al. [2019], Estornell et al. [2023] as to how fair intervention models may be perceived as unfair by groups that preferred the reference model over the fair intervention.

Acknowledgments

We thank Yair Zick for insightful discussion regarding esoteric axioms for aggregator functions, and Indra Kumar, Justin Payan, and Vignesh Viswanathan for helpful feedback on drafts of this work. The authors also wish to acknowledge the support of the Center for Data Science at the University of Massachusetts Amherst, where part of this work was performed under a postdoctoral fellowship.

References

Jacob D Abernethy, Pranjal Awasthi, Matthäus Kleindessner, Jamie Morgenstern, Chris Russell, and Jie Zhang. Active sampling for min-max fairness. In *International Conference on Machine Learning*, volume 162, 2022.

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

Richard J Arneson. Luck egalitarianism and prioritarianism. *Ethics*, 110(2):339–349, 2000.

Kenneth J Arrow, Hollis B Chenery, Bagicha S Minhas, and Robert M Solow. Capital-labor substitution and economic efficiency. *The review of Economics and Statistics*, pages 225–250, 1961.

Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. The moral machine experiment. *Nature*, 563(7729):59–64, 2018.

Siddharth Barman, Umang Bhaskar, Anand Krishna, and Ranjani G Sundaram. Tight approximation algorithms for p-mean welfare under subadditive valuations. *Leibniz International Proceedings in Informatics, LIPIcs*, 173, 2020.

Peter L Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural net-

works. *Advances in neural information processing systems*, 30, 2017.

Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Guha Thakurta. Private stochastic convex optimization with optimal rates. *Advances in neural information processing systems*, 32, 2019.

Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 33:4381–4391, 2020.

Gleb Beliakov, Tomasa Calvo, and Simon James. Some results on Lipschitz quasi-arithmetic means. In *European Society for Fuzzy Logic and Technology Conference*, pages 1370–1375, 2009.

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

Jeremy Bentham. An introduction to the principles of morals and legislation. *University of London: the Athlone Press*, 1789.

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

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 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.

Irene Chen, Fredrik D Johansson, and David Sontag. Why is my classifier discriminatory? *Advances in Neural Information Processing Systems*, 31, 2018a.

Yiling Chen, Chara Podimata, Ariel D Procaccia, and Nisarg Shah. Strategyproof linear regression in high dimensions. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 9–26, 2018b.

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.

Corinna Cortes, Mehryar Mohri, Javier Gonzalvo, and Dmitry Storcheus. Agnostic learning with multiple objectives. *Advances in Neural Information Processing Systems*, 33, 2020.

Cyrus Cousins. An axiomatic theory of provably-fair welfare-centric machine learning. In *Advances in Neural Information Processing Systems*, 2021.

Cyrus Cousins. Uncertainty and the social planner's problem: Why sample complexity matters. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022.

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

Cyrus Cousins, Shahrzad Haddadan, and Eli Upfal. Making mean-estimation more efficient using an MCMC trace variance approach: DynaMITE. *arXiv:2011.11129*, 2020.

Cyrus Cousins, Kavosh Asadi, and Michael L. Littman. Fair E^3: Efficient welfare-centric fair reinforcement learning. In *5th Multidisciplinary Conference on Reinforcement Learning and Decision Making (RLDM)*, 2022a.

Cyrus Cousins, Bhaskar Mishra, Enrique Areyan Viqueira, and Amy Greenwald. Computational and data requirements for learning generic properties of simulation-based games. *arXiv:2208.06400*, 2022b.

Cyrus Cousins, Bhaskar Mishra, Enrique Areyan Viqueira, and Amy Greenwald. Learning properties in simulation-based games. In *Proceedings of the 22nd International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, 2023a.

Cyrus Cousins, Vignesh Viswanathan, and Yair Zick. Dividing good and better items among agents with submodular valuations. *arXiv:2302.03087*, 2023b.

Cyrus Cousins, Vignesh Viswanathan, and Yair Zick. The good, the bad and the submodular: Fairly allocating mixed manna under order-neutral submodular preferences. *arXiv:2307.12516*, 2023c.

Cyrus Cousins, Chloe Wohlgemuth, and Matteo Riondato. BAVarian: Betweenness centrality approximation with variance-aware Rademacher averages. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 17(6):1–47, March 2023d. ISSN 1556-4681. doi: 10.1145/3577021.

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. *Cowles Foundation Discussion Papers*, 76, 1959.

Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, and Aaron Roth. Minimax group fairness: Algorithms and experiments. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 66–76, 2021.

Evan Dong and Cyrus Cousins. Decentering imputation: Fair learning at the margins of demographics. In *Queer in AI Workshop @ ICML*, 2022.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

Andrew Estornell, Sanmay Das, Brendan Juba, and Yevgeniy Vorobeychik. Popularizing fairness: Group fairness and individual welfare. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7485–7493, 2023.

Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making. *Communications of the ACM*, 64(4): 136–143, 2021.

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.

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.

Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning*, pages 2029–2037. PMLR, 2018.

Thomas M Hurka. Average utilitarianisms. *Analysis*, 42(2): 65–69, 1982.

Mamoru Kaneko and Kenjiro Nakamura. The Nash social welfare function. *Econometrica: Journal of the Econometric Society*, pages 423–435, 1979.

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)*, volume 67, page 43. Schloß Dagstuhl–Leibniz-Zentrum für Informatik, 2017.

Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems*, 33:728–740, 2020.

Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning*, pages 5809–5819. PMLR, 2020.

Natalia Martinez, Martin Bertran, and Guillermo Sapiro. Minimax Pareto fairness: A multi objective perspective. In *International Conference on Machine Learning*, pages 6755–6764. PMLR, 2020.

Daniel McFadden. Constant elasticity of substitution production functions. *The Review of Economic Studies*, 30 (2):73–83, 1963.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

Reshef Meir and Jeffrey S Rosenschein. Strategyproof classification. *ACM SIGecom Exchanges*, 10(3):21–25, 2011.

John Stuart Mill. *Utilitarianism*. Parker, Son, and Bourn, London, 1863.

Bhaskar Mishra, Cyrus Cousins, and Amy Greenwald. Regret pruning for learning equilibria in simulation-based games. *arXiv:2211.16670*, 2022.

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

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

Robert Nozick. *Anarchy, state, and utopia*. Basic Books, 1974.

Yonatan Oren, Shiori Sagawa, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust language modeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4227–4237, 2019.

D Parfit. Equality and priority. *Ratio (Oxford)*, 10(3):202–221, 1997.

Neel Patel, Reza Shokri, and Yair Zick. Model explanations with differential privacy. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1895–1904, 2022.

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

Arieal D Procaccia. Towards a theory of incentives in machine learning. *ACM SIGecom Exchanges*, 7(2):1–5, 2008.

John Rawls. *A theory of justice*. Harvard University Press, 1971.

John Rawls. *Justice as fairness: A restatement*. Harvard University Press, 2001.

Matteo Riondato and Eli Upfal. Mining frequent itemsets through progressive sampling with Rademacher averages. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1005–1014. ACM, 2015.

Matteo Riondato and Eli Upfal. ABRA: Approximating betweenness centrality in static and dynamic graphs with Rademacher averages. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 12(5):1–38, 2018.

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

Alvin E Roth et al. Independence of irrelevant alternatives, and solutions to Nash's bargaining problem. *Journal of Economic Theory*, 16(2):247–251, 1977.

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

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2019.

Mark Schneider and Jonathan W Leland. Salience and social choice. *Experimental Economics*, 24(4):1215–1241, 2021.

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.

Shubhanshu Shekhar, Greg Fields, Mohammad Ghavamzadeh, and Tara Javidi. Adaptive sampling for minimax fair classification. *Advances in Neural Information Processing Systems*, 34, 2021.

Umer Siddique, Paul Weng, and Matthieu Zimmer. Learning fair policies in multi-objective (deep) reinforcement learning with average and discounted rewards. In *International Conference on Machine Learning*, pages 8905–8915. PMLR, 2020.

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.

Enrique Areyan Viqueira, Cyrus Cousins, and Amy Greenwald. Improved algorithms for learning equilibria in simulation-based games. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 79–87, 2020.

Enrique Areyan Viqueira, Cyrus Cousins, and Amy Greenwald. Learning competitive equilibria in noisy combinatorial markets. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, 2021.

Vignesh Viswanathan and Yair Zick. A general framework for fair allocation under matroid rank valuations. In *Proceedings of the 24th ACM Conference on Economics and Computation*, pages 1129–1152, 2023.

Puyu Wang, Yunwen Lei, Yiming Ying, and Hai Zhang. Differentially private SGD with non-smooth losses. *Applied and Computational Harmonic Analysis*, 56:306–336, 2022.

Gal Yona and Guy Rothblum. Probably approximately metric-fair learning. In *International Conference on Machine Learning*, pages 5680–5688. PMLR, 2018.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR, 2013.

Appendices

We derive every result stated in the main body in appendix A, and provide supplementary experiments in appendix B.

A Proof Compendium

We now derive all lemmas, theorems, corollaries, and other results stated in the paper body. This appendix is broken into two subappendices, the first (appendix A.1) shows the properties and relationships between the various axioms of section 3, and the second (appendix A.2) shows all results in section 4 related to sample complexity and FPAC learning.

A.1 Properties of Axiomatic Aggregator Functions

Before delving into the proofs of all results given in the paper body, we first state a standard result that shows the power-mean $\mathrm{M}_p(\boldsymbol{u}; \boldsymbol{w})$ is monotonically increasing in p, usually referred to as the *power-mean inequality*.

Lemma A.1 (Power-Mean Inequality)**.** Suppose $-\infty \leq p \leq q \leq \infty$. Then for any $\boldsymbol{u} \in \mathbb{R}_{0+}^{g}, \boldsymbol{w} \in \triangle_g$, it holds

$$\mathrm{M}_p(\boldsymbol{u}; \boldsymbol{w}) \leq \mathrm{M}_q(\boldsymbol{u}; \boldsymbol{w}) \ .$$

We now show lemma 3.2.

Lemma 3.2 (Equivalence of Weighted Axioms)**.** Consider some aggregator function $\mathrm{M}(\cdot; \boldsymbol{w})$. It always holds that WS (axiom 2) $\wedge$ WD (axiom 3) $\Leftrightarrow$ WA (axiom 9).

Proof. Recall the axioms in question:

Weighted Symmetry (WS): For all permutations π over $\mathcal{G}$, it holds that $\mathrm{M}(\boldsymbol{u}; \boldsymbol{w}) = \mathrm{M}\big(\pi(\boldsymbol{u}); \pi(\boldsymbol{w})\big)$.

Weighted Decomposability (WD): Suppose $\alpha \in (0, 1)$. Then $\mathrm{M}(\boldsymbol{u}; \boldsymbol{w}) = \mathrm{M}\big(\langle \boldsymbol{u}_1, \boldsymbol{u}_1, \boldsymbol{u}_2, \dots \rangle; \langle \alpha \boldsymbol{w}_1, (1-\alpha)\boldsymbol{w}_1, \boldsymbol{w}_2, \dots \rangle\big)$.

Weighted Additivity (WA): Suppose $g' \in \mathbb{Z}_+ \cup \{\infty\}$, $\boldsymbol{u}' \in \mathbb{R}_{0+}^{g'}$, and weights vector $\boldsymbol{w}' \in \triangle_{g'}$ such that for all $u \in \mathbb{R}_{0+}$, it holds that $\sum_{i \in \mathcal{G}} \boldsymbol{w}_i \mathbb{1}_u(\boldsymbol{u}_i) = \sum_{i \in \mathcal{G}'} \boldsymbol{w}'_i \mathbb{1}_u(\boldsymbol{u}'_i)$. Then $\mathrm{M}(\boldsymbol{u}; \boldsymbol{w}) = \mathrm{M}(\boldsymbol{u}'; \boldsymbol{w}')$.

We first show the reverse direction, i.e., WS (axiom 2) $\wedge$ WD (axiom 3) $\Leftarrow$ WA (axiom 9). Note that the WD holds by definition, as the two weight terms $\alpha \boldsymbol{w}_1$ and $(1-\alpha)\boldsymbol{w}_1$ that share sentiment $\boldsymbol{u}_1$ in WD are combined within the summation of WA. Now, observe that WS holds by commutativity of summation over countable sets, thus the LHS and RHS summations in the WA definitions both remain invariant under arbitrary permutation.

We now show the forward direction, i.e., WS (axiom 2) $\wedge$ WD (axiom 3) $\implies$ WA (axiom 9). This result is less direct, but observe that together, (WS) and (WD) can be used to consolidate the weights of all $\boldsymbol{u}_i, \boldsymbol{u}_j$ s.t. $\boldsymbol{u}_i = \boldsymbol{u}_j$. In particular for each unique $\boldsymbol{u}_i$, we can produce some *unique minimal reduction* $\boldsymbol{u}^\star$ and $\boldsymbol{w}^\star$ over population $\mathcal{G}^\star$ such that
(1) $\boldsymbol{u}^\star_1 < \boldsymbol{u}^\star_2 < \boldsymbol{u}^\star_3 < \dots$;
(2) for all group indices $i \in \mathcal{G}^\star$, there exists some $j \in \mathcal{G}$ such that $\boldsymbol{u}^\star_i = \boldsymbol{u}_j$; &
(3) for all group indices $i \in \mathcal{G}^\star$, it holds $\boldsymbol{w}^\star_k = \sum_{j \in \mathcal{G}} \boldsymbol{w}_j \mathbb{1}_u(\boldsymbol{u}_j)$ for some $u \in \mathbb{R}_{0+}$.
Now, observe that *exactly the same* $\boldsymbol{u}^\star$ and $\boldsymbol{w}^\star$ are produced by repeating this process for $\boldsymbol{u}'$ and $\boldsymbol{w}'$ over population $\mathcal{G}'$, thus we may conclude that for all $u \in \mathbb{R}_{0+}$, it holds that

$$\overbrace{\sum_{i \in \mathcal{G}} \boldsymbol{w}_i \mathbb{1}_u(\boldsymbol{u}_i) = \underbrace{\sum_{i \in \mathcal{G}^\star} \boldsymbol{w}^\star_i \mathbb{1}_u(\boldsymbol{u}^\star_i)}_{\text{Unique Minimal Reduction for } \boldsymbol{u}', \boldsymbol{w}'} = \sum_{i \in \mathcal{G}'} \boldsymbol{w}'_i \mathbb{1}_u(\boldsymbol{u}'_i)}^{\text{Unique Minimal Reduction for } \boldsymbol{u}, \boldsymbol{w}} \ .$$

We may thus conclude WA. □

We now show lemma 3.3.

Lemma 3.3 (Transfer Principle Equivalencies)**.** Consider some aggregator function $\mathrm{M}(\cdot; \boldsymbol{w})$. The following relate properties (axioms) that $\mathrm{M}(\cdot; \boldsymbol{w})$ obeys.
1) PDTP (11) $\implies$ WTP (8); &

2) Suppose axioms 1–7. Then WTP (8) $\implies$ PDTP (11).

Proof. We first show item 1. Observe that the PDTP (11) $\implies$ WTP (8) follows directly, as the PDTP requires that a broad class of equitable (dis)utility transfers are favorable, whereas the WTP requires only that *there exist* some (dis)utility transfer between two particular groups that is favorable.

The conditional reverse implication of item 2 is a bit more subtle. Suppose axioms 1–7. Then, by theorem 3.5 item 1,[6] we may conclude $\mathrm{M}(\boldsymbol{u};\boldsymbol{w}) = \mathrm{M}_p(\boldsymbol{u};\boldsymbol{w})$ for some $p \in \mathbb{R}$. Now, suppose for the sake of argument (shown below) that axiom 8 does not hold for welfare functions with $p > 1$, or for malfare functions with $p < 1$. We may thus conclude $p \leq 1$ in the welfare case, and $p \geq 1$ in the malfare case, in either case for which axiom 11 is known to hold, which completes the result.

The remaining step is to show that axiom 8 *does not hold* if $p > 1$ for welfare functions, or $p < 1$ for malfare functions. First, observe that for $p \neq 0$, the monotonic transformation $p\mathrm{M}_p^p(\boldsymbol{u};\boldsymbol{w}) = p\sum_{i=1}^g \boldsymbol{w}_i \boldsymbol{u}_i^p$ of $\mathrm{M}_p(\boldsymbol{u};\boldsymbol{w})$ is *convex* for $p > 1$, and *concave* for $0 \neq p < 1$. Now, for any $\boldsymbol{u} \in \mathbb{R}_{0+}^g$, let $i \doteq \operatorname{argmin}_i \boldsymbol{u}_i$ and $j \doteq \operatorname{argmax}_j \boldsymbol{u}_j$, and suppose some $\varepsilon > 0$ s.t. $\boldsymbol{u}_i + \boldsymbol{w}_j\varepsilon < \boldsymbol{u}_j - \boldsymbol{w}_i\varepsilon$. Any "equitable transfer" of the form $\mathrm{W}(\boldsymbol{u} + \varepsilon\boldsymbol{w}_j\mathbb{1}_i - \varepsilon\boldsymbol{w}_i\mathbb{1}_j;\boldsymbol{w})$ obeys $\mathrm{W}(\boldsymbol{u} + \varepsilon\boldsymbol{w}_j\mathbb{1}_i - \varepsilon\boldsymbol{w}_i\mathbb{1}_j;\boldsymbol{w}) > \mathrm{W}(\boldsymbol{u};\boldsymbol{w})$ for welfare if $p > 1$. Similarly, an "equitable transfer" of the form $\mathbb{M}(\boldsymbol{u} + \varepsilon\boldsymbol{w}_j\mathbb{1}_i - \varepsilon\boldsymbol{w}_i\mathbb{1}_j;\boldsymbol{w})$ obeys $\mathbb{M}(\boldsymbol{u} + \varepsilon\boldsymbol{w}_j\mathbb{1}_i - \varepsilon\boldsymbol{w}_i\mathbb{1}_j;\boldsymbol{w}) < \mathbb{M}(\boldsymbol{u};\boldsymbol{w})$ for malfare if $p < 1$. Both cases are apparent from the monotonic transform, as the $\frac{\boldsymbol{w}_i}{\boldsymbol{w}_i}$ and $\frac{\boldsymbol{w}_j}{\boldsymbol{w}_j}$ weighting terms cancel, leaving only transfers along the curvature of the $(\cdot)^p$ power function. In either case, the WTP is violated, thus we may conclude $p \leq 1$ for welfare functions, and $p \geq 1$ for malfare functions. Finally, note that similar logic applies for the case of $p = 0$, instead using a logarithmic monotonic transform. □

We now show theorem 3.5.

Theorem 3.5 (Aggregator Function Properties)**.** Suppose aggregator function $\mathrm{M}(\boldsymbol{u};\boldsymbol{w})$, and assume arbitrary sentiment vector $\boldsymbol{u} \in \mathbb{R}_{0+}^g$ and weights vector $\boldsymbol{w} \in \triangle_g$. The following then hold.
1) *Power-Mean Factorization*: Axioms 1–7 imply there exists some $p \in \mathbb{R}$ such that $\mathrm{M}(\boldsymbol{u};\boldsymbol{w}) = \mathrm{M}_p(\boldsymbol{u};\boldsymbol{w})$.
2) *Fair Welfare and Malfare*: Axioms 1–8 imply $p \in (-\infty, 1]$ for welfare and $p \in [1, \infty)$ for malfare.

Proof. The key to showing this result is to note that, as mentioned in the text, theorem 2.4 of Cousins [2021] draws the same *conclusion*, but under different *assumptions*. The proof strategy is thus to show that our seemingly weaker assumptions actually imply (in fact, are equivalent to) the assumptions of the aforementioned result. In particular, for item 1, it suffices to conclude axioms 1, 4–7 & 9, and for item 2, we need only additionally conclude axiom 11.

We now show item 1. Observe that we assume axioms 1–7 directly, leaving only axiom 9 (WA), which by lemma 3.2, is implied by axioms 2 & 3. This concludes item 1.

We now show item 2. Observe that after assuming our axioms, we need only show axiom 11 (PDTP), which by lemma 3.3 item 2, is implied by the assumed axioms 1–7, in conjunction with axiom 8, which is also assumed. This concludes item 2. □

We now show lemma 3.7.

Lemma 3.7 (Consequences of Strong Axioms)**.** Suppose power-mean aggregator function $\mathrm{M}_p(\cdot;\boldsymbol{w})$. The following then hold:
1) Strengthening the SM axiom (i.e., $1 \to 12$) implies $p > 0$.
2) Strengthening the WTP axiom (i.e., $8 \to 13$) implies $p \neq 1$, thus $p < 1$ for welfare and $p > 1$ for malfare.
3) Strict PDTP (i.e., $11 \to 14$) implies $p \neq 1$ and $p \neq \pm\infty$, thus $p \in (-\infty, 1)$ for welfare and $p \in (1, \infty)$ for malfare.

Proof. We first show item 1. First note that the desideratum follows directly from the following claim: "If $\boldsymbol{u} \neq \mathbf{0}$ and $\min_{i\in\mathcal{G}} \boldsymbol{u}_i = 0$, then $(\mathrm{M}_p(\boldsymbol{u};\boldsymbol{w}) = 0) \Leftrightarrow (p \leq 0)$." In particular, here $\mathrm{M}(\boldsymbol{u};\boldsymbol{w}) = \mathrm{M}(\mathbf{0};\boldsymbol{w}) = 0$ for $\boldsymbol{u} \neq \mathbf{0}$ would violate axiom 12, thus by contraposition, axiom 12 implies $p > 0$. We thus need only show this claim, which follows via analysis of the power mean.

It is straightforward to see that since $\boldsymbol{u} \neq \mathbf{0}$, it holds $p > 0 \implies \mathrm{M}_p(\boldsymbol{u};\boldsymbol{w}) > 0$, thus by the contrapositive, $\mathrm{M}_p(\boldsymbol{u};\boldsymbol{w}) = 0 \implies p \leq 0$. To see the converse, first observe that since $\min_i \boldsymbol{u}_i = 0$, it holds $p = 0 \implies \mathrm{M}_p(\boldsymbol{u};\boldsymbol{w}) = 0$. The case of

[6] Note that theorem 3.5 makes use of this result to show item 2, but we use only theorem 3.5 item 1 here, thus there is no cyclic dependency.

$p < 0$ appears to be a bit more subtle, however observe that the power mean is monotonically increasing in p, and thus in this case, we observe the sandwich inequality

$$0 = \mathrm{M}_{-\infty}(\boldsymbol{u}; \boldsymbol{w}) \leq \mathrm{M}_p(\boldsymbol{u}; \boldsymbol{w}) \leq \mathrm{M}_0(\boldsymbol{u}; \boldsymbol{w}) = 0 \ .$$

It thus holds that, for $\boldsymbol{u} \neq \mathbf{0}$, $p \leq 0 \implies \mathrm{M}_p(\boldsymbol{u}; \boldsymbol{w}) = 0$. Both directions of the bijection of the claim have now been shown, thus item 1 is complete.

We now show item 2. From theorem 3.5 item 2, we already have $p \leq 1$ for welfare and $p \geq 1$ for malfare, so in both cases, all we need do is show $p \neq 1$, which we now show via the contrapositive. Now observe that for all ε, it holds that $\mathrm{M}_1(\boldsymbol{u} + \mathbb{1}_i \boldsymbol{w}_j \varepsilon; \boldsymbol{w}) = \mathrm{M}_1(\boldsymbol{u} - \mathbb{1}_j \boldsymbol{w}_i \varepsilon; \boldsymbol{w})$, hence for no choice of ε does the transfer result in a strict improvement to welfare or malfare. We thus conclude $p = 1 \implies \neg$(axiom 13), hence (axiom 13) $\implies p \neq 1$, which concludes item 2.

We now show item 3. Similar logic to item 2 precludes the case of $p = 1$ (indeed, observe that this must be so, as SPDTP implies PDTP, by similar reasoning to that found in the proof of lemma 3.3 item 1), this portion of item 3 is a direct corollary of item 2. Now, observe that the egalitarian cases $p \in \pm\infty$ are also inadmissible, essentially because they are only sensitive to the extreme values of $\boldsymbol{u}$ and thus transfer between any two non-extreme $\boldsymbol{u}_i, \boldsymbol{u}_j$, i.e., transfer between i, j s.t. $\inf_k \boldsymbol{u}_k < \boldsymbol{u}_i < \boldsymbol{u}_j < \sup_k \boldsymbol{u}_k$, has no impact on the egalitarian power means. We thus conclude that under axiom 14 (SPDTP), it holds that $p \neq \pm\infty$. □

We now show lemma 3.9.

Lemma 3.9 (Consequences of Extreme Axioms)**.** Suppose as in lemma 3.7. The following then hold.
1) 0B (axiom 15) $\Leftrightarrow p \leq 0$. 2) FC (axiom 16) $\Leftrightarrow p < 0$.

Proof. We first show item 1. We first show that 0B (axiom 15) $\implies p \leq 0$. This is clear by contrapositive, as for any $p > 0$, it holds $\mathrm{M}_p(\langle 0, 1 \rangle; \langle \frac{1}{2}, \frac{1}{2} \rangle) = \frac{1}{2^{1/p}} > 0$, thus 0B does not hold.

We now show the converse, i.e., $p \leq 0 \implies$ 0B (axiom 15). In particular, we seek to show that $\lim_{\boldsymbol{u}_i \to 0^+} \mathrm{M}_0(\boldsymbol{u}; \boldsymbol{w}) = 0$. We first address the case of $p < 0$. Observe that

$$\begin{aligned}
\lim_{\boldsymbol{u}_i \to 0^+} \mathrm{M}_p(\boldsymbol{u}; \boldsymbol{w}) &= \lim_{\boldsymbol{u}_i \to 0^+} \lim_{\varepsilon \to 0^+} \left(\frac{1}{\sum_{i=1}^{g} \frac{\boldsymbol{w}_i}{(\boldsymbol{u}_i + \varepsilon)^{-p}}} \right)^{-\frac{1}{p}} && \text{DEFINITION 3.4 (POWER MEAN)} \\
&= \left(\frac{1}{\lim_{\boldsymbol{u}_i \to 0^+} \lim_{\varepsilon \to 0^+} \sum_{i=1}^{g} \frac{\boldsymbol{w}_i}{(\boldsymbol{u}_i + \varepsilon)^{-p}}} \right)^{-\frac{1}{p}} && \text{LIMIT LAWS} \\
&= \left(\frac{1}{\infty} \right)^{-\frac{1}{p}} = 0 \ . && \text{LIMIT LAWS}
\end{aligned}$$

We now address the case of $p = 0$; in particular, observe that

$$\begin{aligned}
\lim_{\boldsymbol{u}_i \to 0^+} \mathrm{M}_0(\boldsymbol{u}; \boldsymbol{w}) &= \lim_{\boldsymbol{u}_i \to 0^+} \lim_{\rho \to 0} \lim_{\varepsilon \to 0^+} \mathrm{M}_\rho(\boldsymbol{u} + \varepsilon \mathbf{1}; \boldsymbol{w}) && \text{DEFINITION 3.4 (POWER MEAN)} \\
&= \lim_{\boldsymbol{u}_i \to 0^+} \lim_{\varepsilon \to 0^+} \prod_{i=1}^{g} (\boldsymbol{u}_i + \varepsilon)^{\boldsymbol{w}_i} && \text{GEOMETRIC MEAN LIMIT} \\
&= \lim_{\boldsymbol{u}_i \to 0^+} \lim_{\varepsilon \to 0^+} \exp\left(\sum_{i=1}^{g} \boldsymbol{w}_i \ln(\boldsymbol{u}_i + \varepsilon) \right) && \text{LOGARITHMIC IDENTITIES} \\
&= \exp\left(\lim_{\boldsymbol{u}_i \to 0^+} \lim_{\varepsilon \to 0^+} \sum_{i=1}^{g} \boldsymbol{w}_i \ln(\boldsymbol{u}_i + \varepsilon) \right) && \text{LIMIT LAWS} \\
&= \exp(-\infty) = 0 \ . && \text{LIMIT LAWS}
\end{aligned}$$

We thus have that, for any $p \leq 0$, it holds $\lim_{\boldsymbol{u}_i \to 0^+} \mathrm{M}_0(\boldsymbol{u}; \boldsymbol{w}) = 0$. This completes the converse statement, and thus concludes item 1.

We now show item 2. To see this result, first observe that for any $\boldsymbol{u} \succ 0$, say, $\boldsymbol{u} \doteq \mathbf{1}$, it holds that $\lim_{c\to\infty} \mathrm{M}_p(\boldsymbol{u} + c\mathbb{1}_i; \boldsymbol{w}) = \infty$ if and only if $p \geq 0$. In particular, this can be seen by observing that, for the forward direction, that $p < 0 \implies \mathrm{M}_p(\boldsymbol{u}; \boldsymbol{w}) \leq \sqrt[p]{\boldsymbol{w}_j}\boldsymbol{u}_j \leq \infty$ for any $j \in \mathcal{G}$, thus $p \geq 0$, and for the reverse direction,

$$\mathrm{M}_0(\boldsymbol{u} + c\mathbb{1}_i; \boldsymbol{w}) = \lim_{c\to\infty} (1+c)^{\boldsymbol{w}_i} = \infty \ ,$$

and then that $\mathrm{M}_0(\boldsymbol{u}; \boldsymbol{w}) \leq \mathrm{M}_p(\boldsymbol{u}; \boldsymbol{w})$ for any $p \geq 0$, i.e., lemma A.1. Note that the same would hold for all $\boldsymbol{u}$ if we assumed $p > 0$, but not for $\boldsymbol{u}$ s.t. $\boldsymbol{u}_j = 0$ for some $j \neq i$ if $p = 0$, however, we only need the *existence* of a single $\boldsymbol{u}$ for which the statement holds.

From here, logically, we have that

$$\left(\exists \boldsymbol{u} \in \mathbb{R}^g_{0+} \quad \text{s.t.} \quad \lim_{c\to\infty} \mathrm{M}_p(\boldsymbol{u} + c\mathbb{1}_i; \boldsymbol{w}) = \infty\right) \Leftrightarrow (p \geq 0) \ .$$

Now, observe that, by contraposition of the bijection, it holds that

$$(p < 0) \Leftrightarrow \underbrace{\left(\forall \boldsymbol{u} \in \mathbb{R}^g_{0+} : \lim_{c\to\infty} \mathrm{M}_p(\boldsymbol{u} + c\mathbb{1}_i; \boldsymbol{w}) < \infty\right)}_{\text{FC Axiom}} \ ,$$

and observe that the RHS is, by definition, the FC axiom. □

We now show lemma 3.10.

Lemma 3.10 (Power-Mean Differentiation)**.** Suppose $\boldsymbol{u}_{\setminus i} \succ \mathbf{0}$, some weights vector $\boldsymbol{w} \in \triangle_g$, and $p \in \mathbb{R}$. The power mean then differentiates in $\boldsymbol{u}_i$ as follows.
1) If $\boldsymbol{u}_i > 0$, then $\frac{\partial}{\partial \boldsymbol{u}_i} \mathrm{M}_p(\boldsymbol{u}; \boldsymbol{w}) = \frac{\boldsymbol{w}_i \boldsymbol{u}_i^{p-1}}{\mathrm{M}_p^{p-1}(\boldsymbol{u}; \boldsymbol{w})}$.
2) If $p < 0$, then $\lim_{\boldsymbol{u}_i \to 0^+} \frac{\partial}{\partial \boldsymbol{u}_i} \mathrm{M}_p(\boldsymbol{u}; \boldsymbol{w}) = \sqrt[-p]{\frac{1}{\boldsymbol{w}_i}}$.
3) If $p \in [0, 1)$, then $\lim_{\boldsymbol{u}_i \to 0^+} \frac{\partial}{\partial \boldsymbol{u}_i} \mathrm{M}_p(\boldsymbol{u}; \boldsymbol{w}) = \infty$.

Proof. Observe first that item 1 is an elementary application of the chain, power, and summation rules; the only subtlety to this result arises in the remaining cases.

We now show item 2. This case is difficult, as naïve application of item 1 results in an indeterminate $\frac{0}{0}$ form. An experienced practitioner of the calculus of infinitesimals may expect results via L'Hôpital's rule, however in this case, said approach is unwieldy, and a simple limit calculus argument yields the desideratum much more concisely. Observe now that the result follows as

$$\begin{aligned}
\lim_{\boldsymbol{u}_i \to 0^+} \frac{\partial}{\partial \boldsymbol{u}_i} \mathrm{M}_p(\boldsymbol{u}; \boldsymbol{w}) &= \lim_{\boldsymbol{u}_i \to 0^+} \frac{\boldsymbol{w}_i \boldsymbol{u}_i^{p-1}}{\mathrm{M}_p^{p-1}(\boldsymbol{u}; \boldsymbol{w})} && \text{Item 1} \\
&= \lim_{\boldsymbol{u}_i \to 0^+} \boldsymbol{w}_i \mathrm{M}_p^{1-p}\left(j \mapsto \frac{\boldsymbol{u}_j}{\boldsymbol{u}_i}; \boldsymbol{w}\right) && \text{Multiplicative Linearity} \\
&= \boldsymbol{w}_i \left(\sum_{j=1}^g \boldsymbol{w}_j \lim_{\boldsymbol{u}_i \to 0^+} \left(\frac{\boldsymbol{u}_j}{\boldsymbol{u}_i}\right)^p\right)^{\frac{1-p}{p}} && \text{Limit Laws} \\
&= \boldsymbol{w}_i \left(\boldsymbol{w}_i 1^p + \sum_{j=1, j\neq i}^g \boldsymbol{w}_j 0^p\right)^{\frac{1-p}{p}} && \text{Limit Laws} \\
&= \boldsymbol{w}_i^{1+\frac{1-p}{p}} = \boldsymbol{w}_i^{\frac{1}{p}} = \sqrt[-p]{\frac{1}{\boldsymbol{w}_i}} \ . && \text{Algebra}
\end{aligned}$$

We now show item 3. We split this case into two subcases; namely the $p = 0$ and $p > 0$ subcases, essentially because whether $\mathrm{M}_p(\boldsymbol{u}; \boldsymbol{w}) = 0$ in the limit is of material significance to the proof technique.

We begin with the $p \in (0,1)$ case. This case is simpler than the case of $p = 0$, as here $\mathrm{M}_p(\boldsymbol{u};\boldsymbol{w}) \neq 0$ in the limit, thus the difficulty of resolving the $\frac{0}{0}$ indeterminate form in the limit vanishes entirely. In fact, in this case we have a finite nonzero denominator, and an infinite numerator. In particular, we may observe the result as

$$\begin{aligned}
\lim_{\boldsymbol{u}_i \to 0^+} \frac{\partial}{\partial \boldsymbol{u}_i} \mathrm{M}_p(\boldsymbol{u};\boldsymbol{w}) &= \lim_{\boldsymbol{u}_i \to 0^+} \frac{\boldsymbol{w}_i \boldsymbol{u}_i^{p-1}}{\mathrm{M}_p^{p-1}(\boldsymbol{u};\boldsymbol{w})} && \text{ITEM 1} \\
&= \boldsymbol{w}_i \frac{\lim_{\boldsymbol{u}_i \to 0^+} \boldsymbol{u}_i^{p-1}}{\lim_{\boldsymbol{u}_i \to 0^+} \mathrm{M}_p^{p-1}(\boldsymbol{u};\boldsymbol{w})} && \text{LIMIT LAWS} \\
&= \underbrace{\boldsymbol{w}_i \mathrm{M}_p^{1-p}(\boldsymbol{u};\boldsymbol{w})}_{\text{POSITIVE FINITE}} \left(\lim_{\boldsymbol{u}_i \to 0^+} \frac{1}{\boldsymbol{u}_i} \right)^{1-p} = \infty \ . && \text{ALGEBRA}
\end{aligned}$$

We now show the case of $p = 0$. Direct proof is more subtle than for $p \in (0,1)$, but can be derived via reasoning akin to that of item 2. However, it is much easier to observe that the power mean exhibits continuity in p, and therefore taking $\lim_{p \to 0}$ via either the $p < 0$ or $p \in (0,1)$ case yields the desideratum. □

We now show lemma 3.12.

Lemma 3.12 (Power-Mean Lipschitz Continuity)**.** Suppose $p \in \mathbb{R}$, sentiment vectors $\boldsymbol{u}, \boldsymbol{u}' \in \mathbb{R}^g_{0+}$, and weights vector $\boldsymbol{w} \in \triangle_g$. The following then hold.
1) Suppose $p \geq 1$. Then $\mathrm{M}_p(\cdot;\boldsymbol{w})$ is $\sqrt[p]{\boldsymbol{w}_{\max}}$-$\|\cdot\|_1$, 1-$\mathrm{M}_p(|\cdot|;\boldsymbol{w})$, and 1-$\|\cdot\|_\infty$ Lipschitz.
2) Suppose $p < 0$. Then $\mathrm{M}_p(\cdot;\boldsymbol{w})$ is $\frac{1}{\sqrt[p]{\boldsymbol{w}_{\min}}}$-$\|\cdot\|_\infty$ Lipschitz.

Proof. Items 1 & 2 follow directly from lemma 3.10, and consideration of the curvature and monotonicity of these functions. Briefly put, observe that malfare functions (i.e., $p \geq 1$ weighted power-means) exhibit monotonically-increasing convexity, thus derivatives increase as $\boldsymbol{u}_i \to \infty$, whereas welfare functions (i.e., $p \leq 1$ weighted power-means) exhibit monotonically-increasing concavity, thus derivatives increase as $\boldsymbol{u}_i \to 0^+$. We now show each result in detail.

We first show item 1. We begin with the $\sqrt[p]{\boldsymbol{w}_{\max}}$-$\|\cdot\|_1$ Lipschitz property. Observe that for any group index $i \in \mathcal{G}$, and any sentiment value $\boldsymbol{u}_i > 0$, it holds that

$$\frac{\partial}{\partial \boldsymbol{u}_i} \mathrm{M}_p(\boldsymbol{u};\boldsymbol{w}) = \frac{\boldsymbol{w}_i \boldsymbol{u}_i^{p-1}}{\mathrm{M}_p^{p-1}(\boldsymbol{u};\boldsymbol{w})} \leq \lim_{\boldsymbol{u}_i \to \infty} \frac{\boldsymbol{w}_i \boldsymbol{u}_i^{p-1}}{\mathrm{M}_p^{p-1}(\boldsymbol{u};\boldsymbol{w})} = \frac{\boldsymbol{w}_i}{\boldsymbol{w}_i^{\frac{p-1}{p}}} = \boldsymbol{w}_i^{1-\frac{p-1}{p}} = \sqrt[p]{\boldsymbol{w}_i} \ .$$

From here, maximizing over group indices yields the $\sqrt[p]{\boldsymbol{w}_{\max}}$-$\|\cdot\|_1$ Lipschitz characterization.

We now show the 1-$\mathrm{M}_p(|\cdot|;\boldsymbol{w})$ Lipschitz property. Observe that $|\mathrm{M}_p(\boldsymbol{u};\boldsymbol{w}) - \mathrm{M}_p(\boldsymbol{u}';\boldsymbol{w})| \leq \mathrm{M}_p(|\boldsymbol{u} - \boldsymbol{u}'|;\boldsymbol{w})$ follows via the subadditivity of $p \geq 1$ power-mean functions, i.e., they are convex and have the unique zero of $\mathrm{M}_p(\mathbf{0};\boldsymbol{w}) = 0$.

Finally, to see the 1-$\|\cdot\|_\infty$ Lipschitz property, observe that $\mathrm{M}_p(|\boldsymbol{u} - \boldsymbol{u}'|;\boldsymbol{w}) \leq \mathrm{M}_\infty(|\boldsymbol{u} - \boldsymbol{u}'|;\boldsymbol{w}) = \|\boldsymbol{u} - \boldsymbol{u}'\|_\infty$, which follows from monotonicity of the power mean in p, i.e., lemma A.1.

We now show item 2. A $\frac{1}{\sqrt[p]{\boldsymbol{w}_{\min}}}$-$\|\cdot\|_1$ Lipschitz continuity guarantee can easily be seen by maximizing derivatives, via lemma 3.10 item 2 (note that this limit maximizes the derivative, since the welfare function is concave and increasing). It may seem surprising that we could get the same Lipschitz constant for $\|\cdot\|_\infty$, however observe that even taking two values of $\boldsymbol{u}_i, \boldsymbol{u}_j$ to 0 simultaneously actually results in *smaller change*, as it is effectively the same as increasing the weight $\boldsymbol{w}_i$ of a single group, thus the same analysis yields an $\|\cdot\|_\infty$ Lipschitz constant. □

We now show lemma 3.13.

Lemma 3.13 (Power-Mean Hölder Continuity)**.** Suppose $\boldsymbol{u} \in [0,r]^g$, group index $i \in \mathcal{G}$, weights vector $\boldsymbol{w} \in \triangle_g$, and assume where appropriate that $\boldsymbol{u}_i + \varepsilon \leq r$. The power mean then obeys the following Hölder continuity criteria.
1) *Generic Welfare Hölder Condition*: Suppose $p \leq 1$. Then $|\mathrm{M}_p(\boldsymbol{u} + \varepsilon \mathbb{1}_i;\boldsymbol{w}) - \mathrm{M}_p(\boldsymbol{u};\boldsymbol{w})| \leq r^{1-\boldsymbol{w}_i} \varepsilon^{\boldsymbol{w}_i}$, and $\mathrm{M}_p(\cdot;\boldsymbol{w})$ is $r^{1-\boldsymbol{w}_{\min}}$-$\boldsymbol{w}_{\min}$-$\|\cdot\|_\infty$ Hölder continuous.
2) *Positive Welfare Hölder Condition*: Suppose $p \in (0,1]$. Then $|\mathrm{M}_p(\boldsymbol{u} + \varepsilon \mathbb{1}_i;\boldsymbol{w}) - \mathrm{M}_p(\boldsymbol{u};\boldsymbol{w})| \leq r^{1-p} \frac{\boldsymbol{w}_i}{p} \varepsilon^p$. Furthermore, $\mathrm{M}_p(\cdot;\boldsymbol{w})$ meets the following Hölder conditions:

A) $r^{1-p}\frac{\boldsymbol{w}_{\max}}{p}$-$p$-$\|\cdot\|_1$;
B) $r^{1-p}\frac{1}{p}$-p-$\mathrm{M}_1(|\cdot|;\boldsymbol{w})$; &
C) $r^{1-p}\frac{1}{p}$-p-$\|\cdot\|_\infty$.

Proof. Due to the complicated and multifaceted nature of this result, we break the proof into several parts. Before showing the first item, we begin with two auxiliary results that will prove useful throughout.

We first note that a generic way to show λ-α-$|\cdot|$ Hölder continuity w.r.t. $\boldsymbol{u}_i$ is to show that

$$\sup_{\boldsymbol{u},\varepsilon} \frac{|\mathrm{M}_p(\boldsymbol{u}+\varepsilon\mathbb{1}_i;\boldsymbol{w}) - \mathrm{M}_p(\boldsymbol{u};\boldsymbol{w})|}{\varepsilon^\alpha} \leq \lambda \ ,$$

and similar techniques can be used to analyze Hölder continuity w.r.t. norms over per-group differences.

We now note that due to its scale-dependence, it is often convenient to show local Hölder continuity, i.e., Hölder continuity over a bounded region. For simplicity, we assume that utility values have range $[0,1]$, and then extend this analysis to a larger region through the multiplicative linearity axiom. The remainder of the proof assumes WLOG this range, and the below analysis is applied to produce the final (range-dependent) result.

Observe that if $f(x)$ exhibits multiplicative linearity (as do all power means, by the multiplicative linearity axiom), and is λ-α Hölder continuous, then for any $r>0$, it holds that $x \mapsto rf(\frac{x}{r})$ is $r^{1-\alpha}\lambda$-α Hölder continuous. To see this, first suppose $f(x)$ exhibits multiplicative linearity and is λ-α Hölder continuous. Then $g(x) \doteq rf(\frac{x}{r})$ obeys

$$\frac{|g(x)-g(y)|}{|x-y|^\alpha} = \frac{\left|rf(\frac{x}{r}) - rf(\frac{y}{r})\right|}{|x-y|^\alpha} = \frac{r\left|f(\frac{x}{r}) - f(\frac{y}{r})\right|}{r^\alpha\left|\frac{x}{r}-\frac{y}{r}\right|^\alpha} = r^{1-\alpha}\frac{\left|f(\frac{x}{r}) - f(\frac{y}{r})\right|}{\left|\frac{x}{r}-\frac{y}{r}\right|^\alpha} \leq r^{1-\alpha}\lambda \ .$$

We now show item 1. We first consider the case of $p=0$, i.e., we analyze the Nash social welfare $\mathrm{W}_0(\cdot;\boldsymbol{w})$. In this case, observe that the most rapid change to $\mathrm{M}_0(\boldsymbol{u};\boldsymbol{w})$ occurs as some $\boldsymbol{u}_i$ approaches zero, and furthermore, the degree of change is maximized when each remaining $\boldsymbol{u}_j = 1$, i.e., is maximized (this much is clear from concavity). In particular, for each $i \in \mathcal{G}$, taking $\alpha = \boldsymbol{w}_i$, here we have

$$\begin{aligned}
\sup_{\boldsymbol{u},\varepsilon} \frac{|\mathrm{M}_0(\boldsymbol{u}+\varepsilon\mathbb{1}_i;\boldsymbol{w}) - \mathrm{M}_0(\boldsymbol{u};\boldsymbol{w})|}{\varepsilon^{\boldsymbol{w}_i}} &\leq \frac{1}{\varepsilon^{\boldsymbol{w}_i}}\left|\mathrm{M}_0(\mathbf{0}+\varepsilon\mathbb{1}_i+\mathbb{1}_{\mathcal{G}\setminus\{i\}};\boldsymbol{w}) - \mathrm{M}_0(\mathbf{0}+\mathbb{1}_{\mathcal{G}\setminus\{i\}};\boldsymbol{w})\right| && \text{CONCAVITY MONOTONICITY} \\
&= \frac{1}{\varepsilon^{\boldsymbol{w}_i}}\exp\left(\boldsymbol{w}_i\ln(\varepsilon) + (1-\boldsymbol{w}_i)\ln(1)\right) && \text{DEFINITION OF } \mathrm{M}_0(\cdot;\boldsymbol{w}) \\
&= \frac{1}{\varepsilon^{\boldsymbol{w}_i}}\exp\left(\boldsymbol{w}_i\ln(\varepsilon)\right) = 1 \ , && \text{ALGEBRA}
\end{aligned}$$

from which we may conclude $\lambda = 1$. This is enough to bound the Hölder constants for the $\|\cdot\|_1$ norm, however observe that even taking two values of $\boldsymbol{u}_i, \boldsymbol{u}_j$ to 0 simultaneously actually results in *slower growth*, as it is effectively the same as increasing the weight $\boldsymbol{w}_i$, and the Hölder constants are actually higher for smaller weights values $\boldsymbol{w}_i$. We thus conclude that the same bounds hold for the $\|\cdot\|_\infty$ case.

The above completes item 1 for $p=0$, so we now show that the result holds for all $p \leq 1$. In other words, we show that $p=0$ is in some sense the "worst case" for small-scale local deviations. To see this, observe that, for any $\varepsilon > 0$, $i \in \mathcal{G}$, it holds that $\mathrm{M}_p(\boldsymbol{u}+\varepsilon\mathbb{1}_i;\boldsymbol{w}) - \mathrm{M}_p(\boldsymbol{u};\boldsymbol{w})$ is decreasing as $p \to 0$, from both the positive and negative sides. We thus conclude that λ-α-$\|\cdot\|$ Hölder continuity for $\mathrm{M}_0(\boldsymbol{u};\boldsymbol{w})$ implies the same for $\mathrm{M}_p(\boldsymbol{u};\boldsymbol{w})$.

We now show item 2. Assume $p \in (0,1]$. Observe then that

$$\begin{aligned}
\sup_{\boldsymbol{u},\varepsilon} \frac{|\mathrm{M}_p(\boldsymbol{u}+\varepsilon\mathbb{1}_i;\boldsymbol{w}) - \mathrm{M}_p(\boldsymbol{u};\boldsymbol{w})|}{\varepsilon^p} &= \sup_{\varepsilon\in(0,1)} \frac{|\mathrm{M}_p(\langle\varepsilon,1\rangle;\langle\boldsymbol{w}_i,1-\boldsymbol{w}_i\rangle) - \mathrm{M}_p(\langle 0,1\rangle;\langle\boldsymbol{w}_i,1-\boldsymbol{w}_i\rangle)|}{\varepsilon^p} && \text{CONCAVITY MONOTONICITY} \\
&= \sup_{\varepsilon\in(0,1)} \frac{(\boldsymbol{w}_i\varepsilon^p + (1-\boldsymbol{w}_i))^{\frac{1}{p}} - (1-\boldsymbol{w}_i)^{\frac{1}{p}}}{\varepsilon^p} && \text{DEFINITION OF } \mathrm{M}_p(\cdot;\boldsymbol{w}) \\
&\leq \sup_{\varepsilon\in(0,1)} \frac{\frac{1}{p}\boldsymbol{w}_i\varepsilon^p + (1-\boldsymbol{w}_i)^{\frac{1}{p}} - (1-\boldsymbol{w}_i)^{\frac{1}{p}}}{\varepsilon^p} && \text{SEE BELOW} \\
&= \frac{\boldsymbol{w}_i}{p} \ . && \text{ALGEBRA}
\end{aligned}$$

For the step marked SEE BELOW, suppose $a, b \geq 0$ s.t. $a + b \leq 1$. Then for all $c \geq 1$, it holds that $(a+b)^c \leq ca + b^c$. This algebraic manipulation yields the result.

From here, item A follows immediately, and item B follows via a similar argument (i.e., the total weighted deviation, as measured by $\mathrm{M}_p(|\cdot|; \boldsymbol{w})$, plays the role of ε). Finally, item C follows from item B by noting that, for any $\boldsymbol{u} \in \mathbb{R}^g_{0+}$ and weights vector $\boldsymbol{w} \in \triangle_g$, it holds that $\|\boldsymbol{u}\|_{1,\boldsymbol{w}} \leq \|\boldsymbol{u}\|_\infty \leq \|\boldsymbol{u}\|_1$. □

A.2 Analysis of Fair-PAC Learning

We now show theorem 4.1.

Theorem 4.1 (Hölder Continuity and Welfare Optimality)**.** Suppose $\mathrm{W}(\cdot; \boldsymbol{w})$ is λ-α-$\|\cdot\|_{\mathrm{W}}$ Hölder continuous w.r.t. some norm $\|\cdot\|_{\mathrm{W}}$, and additive error bounds $\boldsymbol{\varepsilon}$ that obey (3). Then

$$\sup_{h \in \mathcal{H}} \left| \mathrm{W}\left(i \mapsto \mathop{\mathbb{E}}_{\mathcal{D}_i}[\mathrm{u} \circ h]; \boldsymbol{w}\right) - \mathrm{W}\left(i \mapsto \mathop{\hat{\mathbb{E}}}_{\hat{\mathcal{D}}_i}[\mathrm{u} \circ h]; \boldsymbol{w}\right) \right| \leq \lambda \|\boldsymbol{\varepsilon}\|^\alpha_{\mathrm{W}}.$$

Consequently, the *empirical welfare maximizer*

$$\hat{h} \doteq \sup_{h \in \mathcal{H}} \mathrm{W}\left(i \mapsto \mathop{\hat{\mathbb{E}}}_{\hat{\mathcal{D}}_i}[\mathrm{u} \circ h]; \boldsymbol{w}\right)$$

approximates the *true welfare maximizer*

$$h^\star \doteq \sup_{h^\star \in \mathcal{H}} \mathrm{W}\left(i \mapsto \mathop{\mathbb{E}}_{\mathcal{D}_i}[\mathrm{u} \circ h^\star]; \boldsymbol{w}\right) \ ,$$

in terms of welfare-optimality, as it holds that

$$\mathrm{W}\left(i \mapsto \mathop{\mathbb{E}}_{\mathcal{D}_i}[\mathrm{u} \circ \hat{h}]; \boldsymbol{w}\right) \geq \mathrm{W}\left(i \mapsto \mathop{\mathbb{E}}_{\mathcal{D}_i}[\mathrm{u} \circ h^\star]; \boldsymbol{w}\right) - 2\lambda \|\boldsymbol{\varepsilon}\|^\alpha_{\mathrm{W}} \ .$$

Proof. The first portion of the result follows directly from the assumption, and the definition of Hölder continuity (definition 3.11).

The next applies a standard technique in learning theory, wherein the first bound is applied twice: once for $h^\star$ and once more for $\hat{h}$, alongside the fact that, by definition $\hat{h}$ realizes the supremum over the empirical welfare. In particular, we have

$$\begin{aligned}
\mathrm{W}\left(i \mapsto \mathop{\mathbb{E}}_{\mathcal{D}_i}[\mathrm{u} \circ \hat{h}]; \boldsymbol{w}\right) &\geq \mathrm{W}\left(i \mapsto \mathop{\hat{\mathbb{E}}}_{\hat{\mathcal{D}}_i}[\mathrm{u} \circ \hat{h}]; \boldsymbol{w}\right) - \lambda \|\boldsymbol{\varepsilon}\|^\alpha_{\mathrm{W}} && \text{FIRST PORTION ON } \hat{h} \\
&\geq \mathrm{W}\left(i \mapsto \mathop{\hat{\mathbb{E}}}_{\hat{\mathcal{D}}_i}[\mathrm{u} \circ h^\star]; \boldsymbol{w}\right) - \lambda \|\boldsymbol{\varepsilon}\|^\alpha_{\mathrm{W}} && \mathrm{W}\left(i \mapsto \mathop{\hat{\mathbb{E}}}_{\hat{\mathcal{D}}_i}[\mathrm{u} \circ \hat{h}]; \boldsymbol{w}\right) \geq \mathrm{W}\left(i \mapsto \mathop{\hat{\mathbb{E}}}_{\hat{\mathcal{D}}_i}[\mathrm{u} \circ h^\star]; \boldsymbol{w}\right) \\
&\geq \mathrm{W}\left(i \mapsto \mathop{\mathbb{E}}_{\mathcal{D}_i}[\mathrm{u} \circ h^\star]; \boldsymbol{w}\right) - 2\lambda \|\boldsymbol{\varepsilon}\|^\alpha_{\mathrm{W}} \ . && \text{FIRST PORTION ON } h^\star
\end{aligned}$$

□

We now show theorem 4.2.

Theorem 4.2 (Welfare Sample Complexity)**.** Suppose sample complexity function $\mathrm{m}_{\mathcal{H}}(\varepsilon, \delta, r, \boldsymbol{d})$ for hypothesis class $\mathcal{H}$, and some welfare function $\mathrm{W}(\cdot; \boldsymbol{w})$ that is λ-α-$\|\cdot\|_\infty$ Hölder continuous. Then the sample complexity function

$$\mathrm{m}_{\mathrm{W},\mathcal{H}}(\varepsilon, \delta, g, r, \boldsymbol{d}) \leq \mathrm{m}_{\mathcal{H}}\left(\sqrt[\alpha]{\tfrac{\varepsilon}{\lambda}}, \tfrac{\delta}{g}, r, \boldsymbol{d}\right)$$

is sufficient, i.e., for at least this many samples from each of the g groups, (5) holds. Moreover, for this sample size, with probability at least $1 - \delta$, the empirical welfare maximizer is 2ε-optimal.

Proof. This result essentially follows from theorem 4.1 and the definitions of sample complexity and Hölder continuity. By definition, a sample of size at least $\mathrm{m}_{\mathcal{H}}\left(\sqrt[\alpha]{\tfrac{\varepsilon}{\lambda}}, \delta, r, \boldsymbol{d}\right)$ ensures a probability $1 - \delta$ bound on the supremum deviation *for a single group*, and thus by union bound, a sample of size at least $\mathrm{m}_{\mathcal{H}}\left(\sqrt[\alpha]{\tfrac{\varepsilon}{\lambda}}, \tfrac{\delta}{g}, r, \boldsymbol{d}\right)$ ensures a probability $1 - \delta$ on the $\|\cdot\|_\infty$ norm of per-group supremum deviations *over all groups*, i.e., it shall hold with the above probability that $\|\boldsymbol{\varepsilon}\|_\infty \leq \sqrt[\alpha]{\tfrac{\varepsilon}{\lambda}}$. Then, applying theorem 4.1 yields

$$\sup_{h \in \mathcal{H}_{\boldsymbol{d}}} \left| \mathrm{W}\left(i \mapsto \mathop{\mathbb{E}}_{\mathcal{D}_i}[\mathrm{u} \circ h]; \boldsymbol{w}\right) - \mathrm{W}\left(i \mapsto \mathop{\hat{\mathbb{E}}}_{\hat{\mathcal{D}}_i}[\mathrm{u} \circ h]; \boldsymbol{w}\right) \right| \leq \lambda \|\boldsymbol{\varepsilon}\|^\alpha_\infty \leq \lambda \left(\sqrt[\alpha]{\frac{\varepsilon}{\lambda}}\right)^\alpha = \varepsilon \ .$$

As we have ε-estimated $\mathrm{W}(\cdot; \boldsymbol{w})$ with this sample, we may conclude that $\mathrm{m}_{\mathrm{W},\mathcal{H}}(\varepsilon, \delta, g, r, \boldsymbol{d}) \leq \mathrm{m}_{\mathcal{H}}\left(\sqrt[\alpha]{\tfrac{\varepsilon}{\lambda}}, \delta, r, \boldsymbol{d}\right)$. Finally, the statement about approximate optimality of the empirical welfare maximizer follows from the second portion of theorem 4.1. □

We now show theorem 4.4.

Theorem 4.4 (Characterizing FPAC Learnability)**.** Suppose some weighted power-mean welfare function $\mathrm{W}_p(\cdot;\boldsymbol{w})$, utility function u with range r, and hypothesis class $\mathcal{H}$ with sample complexity function $\mathrm{m}_{\mathcal{H}}(\varepsilon,\delta,r,\boldsymbol{d}) \in \mathrm{Poly}(\frac{1}{\varepsilon},\log\frac{1}{\delta},r,\boldsymbol{d})$. We then bound the sample complexity $m \doteq \mathrm{m}_{\mathcal{W},\mathcal{H}}(\varepsilon,\delta,\mathrm{W},g,\boldsymbol{d})$ of FPAC learning $\mathcal{H}$ w.r.t. welfare class $\mathcal{W} \doteq \{\mathrm{W}(\cdot;\boldsymbol{w})\}$ as

1) $m \le \mathrm{m}_{\mathcal{H}}\left(\sqrt[\alpha]{\frac{\varepsilon}{2\lambda}},\frac{\delta}{g},r,\boldsymbol{d}\right) \in \mathrm{Poly}\left(\sqrt[\alpha]{\lambda},\frac{1}{\sqrt[\alpha]{\varepsilon}},\log\frac{1}{\delta},\log g,r,\boldsymbol{d}\right)$;

2) $p \in (0,1] \implies m \in \mathrm{Poly}\left(\sqrt[p]{r},\frac{1}{\sqrt[p]{p}},\frac{1}{\sqrt[p]{\varepsilon}},\log\frac{1}{\delta},\log g,\boldsymbol{d}\right)$;

3) $p = 0 \implies m \in \mathrm{Poly}\left(\sqrt[\boldsymbol{w}_{\min}]{r},\frac{1}{\boldsymbol{w}_{\min}\sqrt[\boldsymbol{w}_{\min}]{\varepsilon}},\log\frac{1}{\delta},\log g,\boldsymbol{d}\right)$;

4) $p < 0 \implies m \in \mathrm{Poly}\left(\frac{1}{\varepsilon},\frac{1}{|p|\sqrt[|p|]{\boldsymbol{w}_{\min}}},\log\frac{1}{\delta},\log g,r,\boldsymbol{d}\right)$; &

5) for any $c \in (0,1)$, if $|p| \ge c$ and group weights obey the *nonnegligibility condition* $\boldsymbol{w}_{\min} \ge \frac{c}{g}$, then $m \in \mathrm{Poly}^{\frac{1}{c}}\left(\frac{1}{c},\frac{1}{\varepsilon},g,\log\frac{1}{\delta},r,\boldsymbol{d}\right)$.

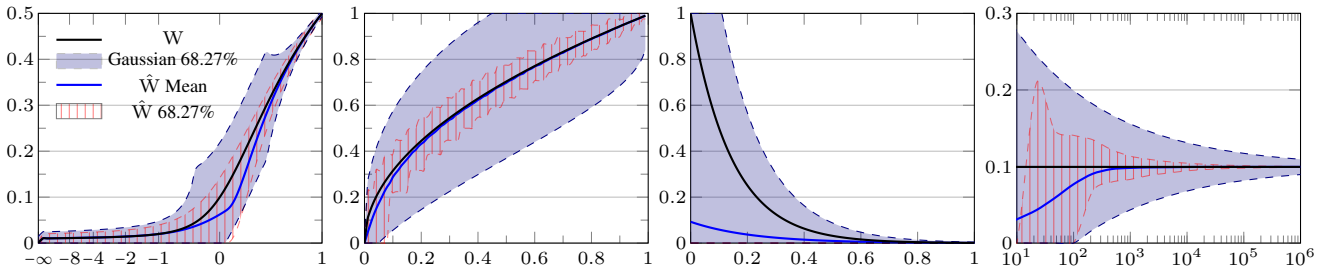
Proof. In each case, the FPAC learning algorithm $\mathcal{A}(\mathcal{D}_{1:g},\mathrm{W},\varepsilon,\delta,\boldsymbol{d})$ is simply empirical welfare maximization on a sufficiently large sample, thus we need only bound the size of such a sufficient sample. Each item of this result is essentially a direct consequence of theorem 4.2, with lemmata 3.12 & 3.13 to bound Lipschitz and Hölder constants. It thus suffices to bound the constants λ for λ-$\|\cdot\|_\infty$ Lipschitz continuity, or λ and α for λ-α-$\|\cdot\|_\infty$ Hölder continuity, for each of the classes under consideration.

In particular, item 1 follows from theorem 4.2 applied to any $\mathrm{W}(\cdot;\boldsymbol{w})$ in the class under consideration. Then, items 2 & 3 follow from item 1, using lemma 3.13 items 2C & 1, respectively, to bound λ and α, and item 4 follows similarly, except using lemma 3.12 item 2 to bound the Lipschitz constant λ (thus $\alpha = 1$).

Finally, item 5 is slightly more involved, but again essentially reduces to item 1. In particular, observe that $|p| \ge c$ means we need not consider $p \approx 0$, and since c is constant, any exponential dependence on c remains polynomial in the remaining variables. Along with the *nonnegligibility condition* $\boldsymbol{w}_{\min} \ge \frac{c}{g}$, this allows us to control the dependence of the Lipschitz constant λ for $p \le -c$ as $\lambda \le \frac{1}{|p|\sqrt[|p|]{\boldsymbol{w}_{\min}}} \le (\frac{g}{c})^{\frac{1}{c}}$, thus $\lambda \in \mathrm{Poly}^{\frac{1}{c}}(g,\frac{1}{c})$ for $p \le -c$. Similarly, for $p \ge c$, note that for welfare functions we need only consider $p \le 1$, and observe that for $\alpha = c$, we have $\lambda = \frac{r^{1-c}}{c}$ by lemma 3.13 item 2C, which yields only $\mathrm{Poly}^{\frac{1}{c}}(\frac{1}{c},\frac{1}{\varepsilon})$ sample complexity terms. In either case, the desideratum is shown. □

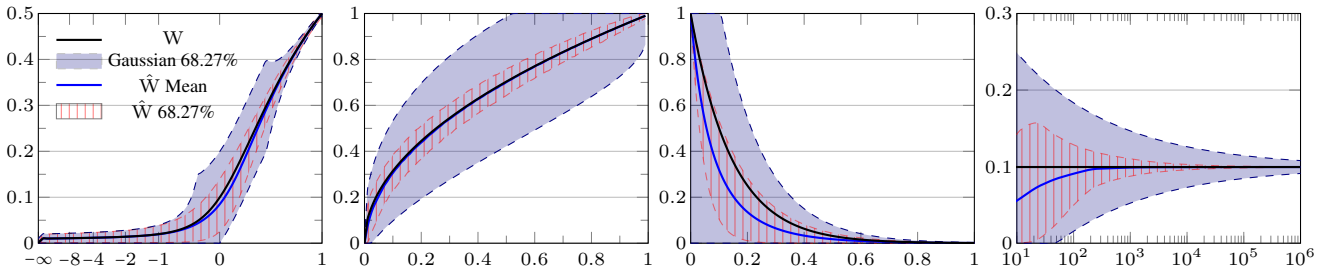
B Supplementary Experiments

We now present two additional one-armed bandit experiments using beta and Bernoulli noise models. Here utility samples are range $[0,1]$ i.i.d. random variables with mean $\boldsymbol{u}_i$ for each group i. For the Bernoulli model, we use $\textsc{Bernoulli}(\boldsymbol{u}_i)$ random variables, and for the beta model, we use $\textsc{Beta}(\boldsymbol{u}_i, 1-\boldsymbol{u}_i)$, which acts as continuous approximation of a $\textsc{Bernoulli}(\boldsymbol{u}_i)$ coin, avoiding issues of discreteness, with exactly half the variance. The main difference here is that the variance of each estimator is now dependent on $\boldsymbol{u}_i$, being either $\frac{\boldsymbol{u}_i(1-\boldsymbol{u}_i)}{\boldsymbol{m}_i} \le \frac{1}{4\boldsymbol{m}_i}$ in the Bernoulli case, or $\frac{\boldsymbol{u}_i(1-\boldsymbol{u}_i)}{2\boldsymbol{m}_i} \le \frac{1}{8\boldsymbol{m}_i}$ in the beta case, as opposed to $\frac{1}{12\boldsymbol{m}_i}$ in the uniform case. For all values of $\boldsymbol{u}_i$ sufficiently far from $\frac{1}{2}$, these variance values are much smaller than under the uniform noise model, so we use only $m = 50$ samples unless otherwise noted.



(a) Welfare vs. $p \in [-\infty, 1]$ (b) Welfare vs. minority mean $\boldsymbol{u}_2$ (c) Welfare vs. minority weight $\boldsymbol{w}_2$ (d) Welfare vs. sample size m

Figure A1: Estimating the Welfare of a 1-Armed Bandit under Bernoulli Noise. Each plot studies the response of welfare to one parameter, and the remaining parameters are selected from $p = 0$, $\boldsymbol{u} = \langle 0.99, 0.01\rangle$, $\boldsymbol{w} = \langle\frac{1}{2},\frac{1}{2}\rangle$, and $m = 50$. All axes are linear, except A1a, which plots $p \in [-\infty, 1]$ by transforming $x = \frac{1}{\pi}\arctan(1-p)$, and A1d, which is logarithmic in x.



(a) Welfare vs. $p \in [-\infty, 1]$ (b) Welfare vs. minority mean $\boldsymbol{u}_2$ (c) Welfare vs. minority weight $\boldsymbol{w}_2$ (d) Welfare vs. sample size m

Figure A2: Estimating the Welfare of a 1-Armed Bandit under Beta Noise. Each plot studies the response of welfare to one parameter, and the remaining parameters are selected from $p = 0$, $\boldsymbol{u} = \langle 0.99, 0.01 \rangle$, $\boldsymbol{w} = \langle \frac{1}{2}, \frac{1}{2} \rangle$, and $m = 50$. All axes are linear, except A2a, which plots $p \in [-\infty, 1]$ by transforming $x = \frac{1}{\pi} \arctan(1 - p)$, and A2d, which is logarithmic in x.

The remainder of the experimental setup is identical to that under the uniform noise model, as described in section 5. In particular, we vary the parameters p, minority utility $\boldsymbol{u}_2$, minority group weight $\boldsymbol{w}_2$, and sample size m in order to study performance around the particularly challenging $p \approx 0$ and $\boldsymbol{w}_{\min} \approx 0$ domains, and present the results in figures A1 & A2.

The beta and Bernoulli experiments are largely similar to the uniform noise experiment of section 5. In figures A1b & A2b which adjust the minority group utility $\boldsymbol{u}_2$, we observe that as the minority utility tends to 0, the empirical confidence intervals remain surprisingly wide, especially when considering that the variance of this coin is extremely small (also tending to 0). In contrast, as the coin bias tends to 1, confidence intervals get much smaller, as here again variance goes to 0, and here the welfare function is not sensitive to small changes. Note also that, generally speaking, the lower variances under the beta noise model (figure A2) result in tighter confidence bounds than the Bernoulli noise model (figure A1).