Revisiting Fair-PAC Learning and the Axioms of Cardinal Welfare

Cyrus Cousins
University of Massachusetts Amherst

Abstract

Cardinal objectives serve as intuitive targets in fair machine learning by summarizing *utility* (welfare) or *disutility* (malfare) $\boldsymbol{u}$ over g groups. Under standard axioms, all welfare and malfare functions are $\boldsymbol{w}$-*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, or $p \geq 1$ for malfare. We show the same under weaker axioms, and also identify stronger axioms that naturally restrict p. It is known that power-mean malfare functions are Lipschitz continuous, and thus *statistically easy* to estimate or learn. We show that *all power means* are locally Hölder continuous, i.e., $\left|\mathrm{M}(\boldsymbol{u}; \boldsymbol{w}) - \mathrm{M}(\boldsymbol{u}'; \boldsymbol{w})\right| \leq \lambda \left\| \boldsymbol{u} - \boldsymbol{u}' \right\|^\alpha$ for some $\lambda, \alpha, \|\cdot\|$. In particular, 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 *optimizing* welfare. This yields a novel concept of *fair-PAC learning*, wherein welfare functions are only *polynomially harder* to optimize than malfare functions, except when $p \approx 0$ or $\min_i \boldsymbol{w}_i \approx 0$, which is *exponentially harder*.

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), 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 and 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 then propose stronger axioms to further specify such functions, and explore the resulting classes of fair learnability.

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 a standard axiom. Subject to these axioms, the 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 the variation in feasible aggregator function concepts reflects natural variation in human morality and social values [Awad et al., 2018, Schneider and Leland, 2021]. We do not reject this claim, however we 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.

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, welfare functions are not always Lipschitz continuous (though they are Hölder continuous), and not uniformly fair-PAC (FPAC) learnable in the sense of Cousins [2021]. However, section 4 shows that under a more lenient definition of FPAC learnability, in which sample complexity (i.e., the sufficient sample size to approximately optimize an objective over some class $\mathcal{H}$) may depend on the welfare function through the *minimum group weight* reciprocal $\frac{1}{\boldsymbol{w}_{\min}}$ and/or $\frac{1}{|p|}$ (which quantifies how close the welfare function is to the challenging $p = 0$ Nash 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 thm. 4.4, which shows that if a bounded utility maximization task is PAC-learnable with polynomial sample complexity, then the *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*. Worst-case sample-complexity bounds are theoretically interesting, as they lend confidence to the proposition that a machine learning system accurately and fairly learns an objective, and sample-efficient analyses are practically significant, as they seek to learn confidently while minimizing the cost of data collection.

All proofs are meticulously derived in appx. A. 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, we can't uniquely characterize fairness with mathematics alone, however analysis does 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.

Prior work in computer science primarily handles malfare. Group-DRO (distributionally robust optimization) methods minimize worst-case (egalitarian) risk [Hu et al., 2018, Oren et al., 2019, Sagawa et al., 2019], 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. 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.

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, and explore their computational and statistical learnability properties, bounding 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.

3 ON CARDINAL FAIRNESS AXIOMS

In this section, we define *aggregator functions*, which summarize overall sentiment (either utility or disutility), expressed as a *sentiment vector* $\boldsymbol{u} \in \mathbb{R}_{0+}^g$ 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 $\Lambda(\boldsymbol{u}; \boldsymbol{w})$ for malfare functions (negatively-connoted sentiment).

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 $\langle \ln(1 + \boldsymbol{u}_1), \ln(1 + \boldsymbol{u}_2), \dots, \ln(1 + \boldsymbol{u}_g) \rangle$. 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.

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 thus the grounded situations that give rise to sentiment vectors. The social planner seeks to select a grounded situation (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. 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) of each group.

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.

Classical econometric theory primarily describes the *unweighted case*, wherein an aggregator function aggregates sentiment over an *unweighted finite discrete* population $\mathcal{G}$, i.e., $\mathcal{G} = \{1, \ldots, g\}$. We generalize to the *weighted discrete* case, wherein an aggregator function operates on a $\boldsymbol{w}$-*weighted discrete* countable population $\mathcal{G}$, i.e., $g \in \mathbb{Z}_+ \cup \{\infty\}$. 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 $\mathrm{M}(\boldsymbol{u}; \boldsymbol{w}) = \mathrm{M}(\pi(\boldsymbol{u}); \pi(\boldsymbol{w}))$.
3) *Weighted Decomposability* (WD): Suppose $\alpha \in (0, 1)$. Then $\mathrm{M}(\boldsymbol{u}; \boldsymbol{w}) = \mathrm{M}(\langle \boldsymbol{u}_1, \boldsymbol{u}_1, \boldsymbol{u}_2, \ldots \rangle; \langle \alpha \boldsymbol{w}_1, (1-\alpha)\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots \rangle)$.
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 $i \doteq \mathrm{argmin}_i \boldsymbol{u}_i$ & $j \doteq \mathrm{argmax}_j \boldsymbol{u}_j$. If $\boldsymbol{u}_i \neq \boldsymbol{u}_j$, then $\exists\, \varepsilon > 0$ 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 $\mathrm{M}(\boldsymbol{u} + \varepsilon \boldsymbol{w}_j \mathbb{1}_i - \varepsilon \boldsymbol{w}_i \mathbb{1}_j; \boldsymbol{w}) \leq \mathrm{M}(\boldsymbol{u}; \boldsymbol{w})$ 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 $\boldsymbol{w}' \in \triangle_{g'}$ s.t. for all $u \in \mathbb{R}_{0+}$: $\sum_{i \in \mathcal{G}} \boldsymbol{w}_i \mathbb{1}_u(\boldsymbol{u}_i) = \sum_{i \in \mathcal{G}'} \boldsymbol{w}'_i \mathbb{1}_u(\boldsymbol{u}'_i)$. Then $\mathrm{M}(\boldsymbol{u}; \boldsymbol{w}) = \mathrm{M}(\boldsymbol{u}'; \boldsymbol{w}')$.
10) *Pigou-Dalton Transfer Principle* (PDTP): 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 & 10 generalize the standard basis of *axioms of cardinal welfare* to *weighted discrete populations*, and together they imply any aggregator function can be expressed as $\mathrm{M}(\boldsymbol{u}; \boldsymbol{w}) \doteq F\left(\sum_{i=1}^g \boldsymbol{w}_i f(\boldsymbol{u}_i)\right)$ for strictly monotonically increasing functions f, F. Axiom 6 is also known in the *constant elasticity of substitution* (CES) literature [Arrow et al., 1961, McFadden, 1963] as *homogeneity of degree one*, and it strengthens the final standard basis axiom, namely *independence of common scale*, which specifies the same up to monotonic transformation. Axiom 7 then characterizes scale, in accordance with *average utilitarianism* [Hurka, 1982], rather than *sum utilitarianism*, to be comparable across populations $\mathcal{G}, \mathcal{G}'$ of different sizes.

On weighted additivity In the unweighted case, it is standard to define symmetry as simply $\mathrm{M}(\boldsymbol{u}) = \mathrm{M}(\pi(\boldsymbol{u}))$ for all permutations π over $\mathcal{G}$. With weights, *weighted symmetry* (axiom 2), i.e., $\mathrm{M}(\boldsymbol{u}; \boldsymbol{w}) = \mathrm{M}(\pi(\boldsymbol{u}); \pi(\boldsymbol{w}))$, only requires *equal treatment* given *equal weights*. *Weighted decomposability* (axiom 3) strengthens this by requiring that a group may be decomposed into multiple groups of equal sentiment value and total weight without impacting 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, it would be perverse not to assume them. We now show that, despite their vastly simpler form, together axioms 2 & 3 are equivalent 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 equitable redistribution and transfer principles The *Pigou-Dalton transfer principle* (PDTP, axiom 10) 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. We now show that, subject to the standard Debreu-Gorman axioms, the WTP and PDTP are equivalent.

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 (10) $\Rightarrow$ WTP (8); &
2) Suppose axioms 1–7. Then WTP (8) $\Rightarrow$ PDTP (10).

Note 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.

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* $\boldsymbol{u} \in \mathbb{R}_{0+}^g$ and *weights vector* $\boldsymbol{w} \in \triangle_g$, is

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

Note that in most cases the limits can safely be ignored. The inner $\lim_{\varepsilon \to 0^+}$ avoids indeterminate forms for $p \leq 0$ 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 *maximum* and *minimum* operators for $p = \pm\infty$, also known as *egalitarian* malfare or welfare, respectively.

We now characterize the class of fair aggregator functions in a result similar to thm. 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 $\boldsymbol{u} \in \mathbb{R}_{0+}^g$ and weights vector $\boldsymbol{w} \in \triangle_g$. Then:
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 aggregator functions.

Axiomatization 3.6 (Strong Axioms). Suppose as in axiomatization 3.1. We now define two strengthened axioms.
11) *Strict Monotonicity at 0* (SM0): $\mathrm{M}(\boldsymbol{u}; \boldsymbol{w}) < \mathrm{M}(\boldsymbol{u} + \varepsilon \mathbb{1}_i; \boldsymbol{w})$.

12) *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.
13) *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 $\mathrm{M}_p(\cdot; \boldsymbol{w})$. Then:
1) Strengthening the SM axiom (i.e., $1 \to 11$) implies $p > 0$.
2) Strengthening the WTP axiom (i.e., $8 \to 12$) implies $p \neq 1$, thus $p < 1$ for welfare and $p > 1$ for malfare.
3) Strict PDTP (i.e., $10 \to 13$) implies $p \neq 1$ and $p \neq \pm\infty$, thus $p \in (-\infty, 1)$ for welfare and $p \in (1, \infty)$ for malfare.

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 *extreme inequality*, where some group has 0 or ∞ utility.

Axiomatization 3.8 (Extreme Axioms). Suppose as in axiomatization 3.1. We now define two additional axioms.
14) *Zero Barrier* (0B): $\lim_{\boldsymbol{u}_i \to 0^+} \mathrm{W}(\boldsymbol{u}; \boldsymbol{w}) = 0$.
15) *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* (14) promotes prioritarianism by ensuring that $\mathrm{W}(\boldsymbol{u}; \boldsymbol{w}) \to 0$ as any $\boldsymbol{u}_i \to 0$, thus disincentivizing *extreme harm* to any group, and bounding the harm caused
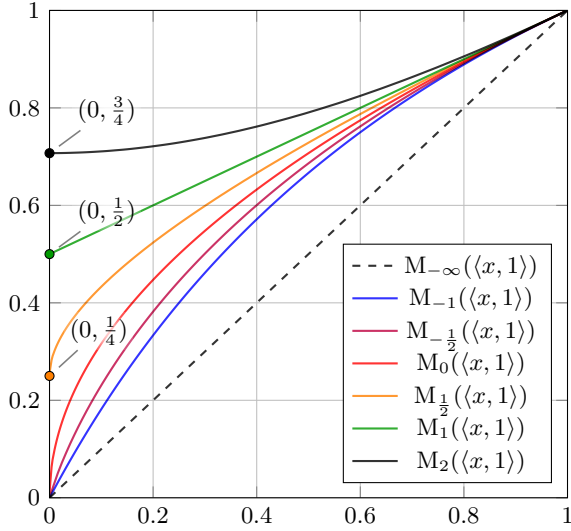
Figure 1: Plots of the unweighted power-mean for various p. Observe that the region around $x = \boldsymbol{u}_1 = 0$, wherein Lipschitz discontinuities may occur, exhibits sharp changes to welfare, as $\mathrm{M}_p(\langle x, 1 \rangle)$ is sensitive to small changes to x.

by the utility monster. Similarly, the *finite ceiling axiom* (15) 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.

Lemma 3.9 (Consequences of Extreme Axioms). Suppose as in lemma 3.7. The following then hold.
1) 0B (axiom 14) $\Leftrightarrow p \leq 0$. 2) FC (axiom 15) $\Leftrightarrow p < 0$.

Observe that, subject to axioms 1–8, FC $\Rightarrow$ 0B $\Rightarrow$ SWTP $\Rightarrow$ 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* take welfare to ∞, in each case concluding the complementary set of permissible p in lemma 3.9.

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 14 and the utilitarian form of 15.

3.4 Continuity Properties of Aggregator Functions

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 $\mathrm{M}_p(\boldsymbol{u}; \boldsymbol{w}) \to 0$, 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 fig. 1 throughout, wherein various power-means are plotted, revealing their pathological behavior for $p \approx 0$.

Lemma 3.10 (Power-Mean Differentiation). Suppose $\boldsymbol{u}_{\backslash i} \succ \boldsymbol{0}$, 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 $\lim_{\boldsymbol{u}_i \to 0^+} \frac{\partial}{\partial \boldsymbol{u}_i} \mathrm{M}_p(\boldsymbol{u}; \boldsymbol{w}) = \sqrt[-p]{\frac{1}{\boldsymbol{w}_i}}$.
3) If $p \in [0, 1)$, then $\lim_{\boldsymbol{u}_i \to 0^+} \frac{\partial}{\partial \boldsymbol{u}_i} \mathrm{M}_p(\boldsymbol{u}; \boldsymbol{w}) = \infty$.

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 $\mathbb{R}_{0+}^g$ if there exist some $\lambda \geq 0, \alpha \in (0, 1]$, s.t. for all $\boldsymbol{u}, \boldsymbol{u}' \in \mathbb{R}_{0+}^g$, it holds that
$$\left| \mathrm{M}(\boldsymbol{u}; \boldsymbol{w}) - \mathrm{M}(\boldsymbol{u}'; \boldsymbol{w}) \right| \leq \lambda \|\boldsymbol{u} - \boldsymbol{u}'\|_{\mathrm{M}}^{\alpha} . \tag{2}$$
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.

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. The following result (visualized in fig. 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) If $p \geq 1$, then $\left| \mathrm{M}_p(\boldsymbol{u}; \boldsymbol{w}) - \mathrm{M}_p(\boldsymbol{u}'; \boldsymbol{w}) \right| \leq \mathrm{M}_p(|\boldsymbol{u} - \boldsymbol{u}'|; \boldsymbol{w})$, and $\mathrm{M}_p(\cdot; \boldsymbol{w})$ is $\sqrt[p]{\boldsymbol{w}_{\max}}$-$\|\cdot\|_1$ and 1-$\|\cdot\|_{\infty}$ Lipschitz.
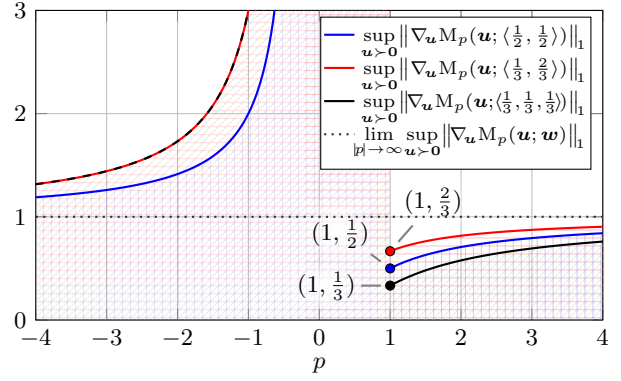

Figure 2: Plots of $\|\cdot\|_1$ Lipschitz constants of weighted power-means, as functions of p. The region below each Lipschitz constant plot is shaded and patterned, to emphasize that *higher values* allow for *sharper rates of change*.
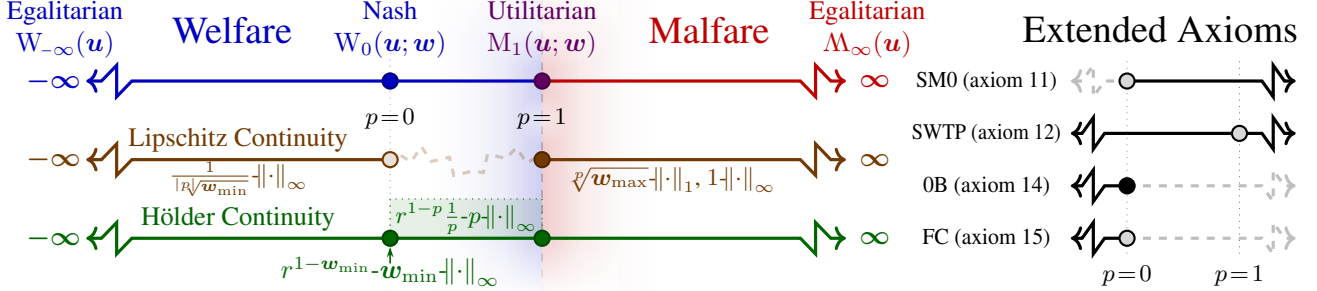
Figure 3: Illustration of power-mean properties and axioms. Solid lines and filled circles denote values of p that concord with an axiom or property, while dashed lines and unfilled circles denote their complement. Basic properties (thm. 3.5) and continuity (lemmata 3.12 & 3.13) are plotted on the left, and extended axioms (lemmata 3.7 & 3.9) are shown on the right.

2) If $p < 0$, then $\mathrm{M}_p(\cdot; \boldsymbol{w})$ is $\frac{1}{\sqrt[|p|]{\boldsymbol{w}_{\min}}}$-$\|\cdot\|_\infty$ Lipschitz.

While $p \in [0, 1)$ power-means are not Lipschitz continuous (cf. lemma 3.10 item 3), we find they are Hölder continuous.

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) *General Welfare Hölder Condition*: Suppose $p \leq 1$. Then $\left| \mathrm{M}_p(\boldsymbol{u} + \varepsilon \mathbb{1}_i; \boldsymbol{w}) - \mathrm{M}_p(\boldsymbol{u}; \boldsymbol{w}) \right| \leq r^{1 - \boldsymbol{w}_i} \varepsilon^{\boldsymbol{w}_i}$, 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 $\left| \mathrm{M}_p(\boldsymbol{u} + \varepsilon \mathbb{1}_i; \boldsymbol{w}) - \mathrm{M}_p(\boldsymbol{u}; \boldsymbol{w}) \right| \leq r^{1-p} \frac{\boldsymbol{w}_i}{p} \varepsilon^p$. This implies $\mathrm{M}_p(\cdot; \boldsymbol{w})$ 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-$\|\cdot\|_{1, \boldsymbol{w}}$ (where $\|\boldsymbol{u}\|_{1, \boldsymbol{w}} \doteq |\boldsymbol{u}| \cdot \boldsymbol{w}$); &
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*, *strategy proofness*, and *statistical learnability*. Due to their commonalities, we briefly treat the first four here, while sections 4 & 5 explore statistical learning in detail. To clarify the relationship of our axioms to these properties, we visualize them in fig. 3.

We assume the *parameters* or *decisions made* by some algorithm are robust to small changes to the objective, and note that Lipschitz or Hölder continuity describe how robust the objective is to small changes to sampled, estimated, or queried utility values. Lipschitz continuity is highly relevant to *differential privacy* [Bassily et al., 2019, 2020, Wang et al., 2022, Patel et al., 2022], as privacy is sensitive to changes to algorithm output caused by *individual-level changes*. For *adversarial robustness* and *strategy 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 colluders, its *impact on welfare* is sublinear, due to the α-power. In particular, under the ε-*truthfulness* assumption [Meir and Rosenschein, 2011, Meir et al., 2012], agents only lie if doing so yields at least ε utility. Assuming

the objective is Hölder continuous, bounded, and strongly concave, the impact of lying is bounded, thus lying is disincentivized. In light of lemmata 3.12 & 3.13, we conclude that while the privacy, stability, and statistical properties of welfare-based algorithms may be strongest under utilitarianism, we can still analyze them for arbitrary power-mean welfare functions.

4 GENERALIZING FAIR-PAC LEARNING

In this section, we seek to estimate or optimize some welfare function, but do not know the utility values, and must instead estimate them via sampling (i.e., from data). In particular, we 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 distribution $\mathcal{D}_i$ over *labeled instance space* $(\mathcal{X} \times \mathcal{Y})$. We assume a hypothesis class $\mathcal{H} \subseteq \mathcal{X} \to \mathcal{Y}'$ mapping 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 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 $(\mathrm{u} \circ h)(x, y) \doteq \mathrm{u}(h(x), y)$.

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

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

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 such bounds 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,

Cousins and Riondato, 2020], or other such tools. The following result can immediately be applied to any model class for which bounds on the supremum deviation are known.

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

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

Consequently, the *empirical welfare maximizer*

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

approximates the *true welfare maximizer*

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

in terms of welfare-optimality, as it holds that

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

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, 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 networks, a vector of per-layer widths [Anthony and Bartlett, 2009] or 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* $m_{\mathcal{H}}(\varepsilon, \delta, r, \boldsymbol{d})$ is some function such that, for utility range r, any distribution $\mathcal{D}$, a sample size of at least $m_{\mathcal{H}}(\varepsilon, \delta, r, \boldsymbol{d})$ ensures that $u \circ \mathcal{H} = \{u \circ h | h \in \mathcal{H}\}$ is *uniformly convergent*, i.e., with probability at least $1 - \delta$, it holds that

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

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

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

We now analyze the sample complexity of welfare.

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

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

is sufficient, i.e., for at least this many samples per group, (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* $u: \mathcal{Y}' \times \mathcal{Y} \to \mathbb{R}_{0+}$, and *welfare class* $\mathcal{W} \subseteq \mathbb{R}_{0+}^g \to \mathbb{R}_{0+}$. $\mathcal{H}$ is *FPAC-learnable* w.r.t. u and $\mathcal{W}$ if there exists an algorithm $\mathcal{A}$ and sample complexity function $m_{\mathcal{W}, \mathcal{H}}$ 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 $W(\cdot; \boldsymbol{w})$ 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., $\hat{h} \leftarrow \mathcal{A}(\mathcal{D}_{1:g}, W, \varepsilon, \delta, \boldsymbol{d})$, such that
1) for each group, $\mathcal{A}(\mathcal{D}_{1:g}, W, \varepsilon, \delta, \boldsymbol{d})$ draws no more than $m_{\mathcal{W}, \mathcal{H}}(\varepsilon, \delta, W, g, \boldsymbol{d})$ samples; &
2) with probability at least $1 - \delta$, $\hat{h}$ obeys

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

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

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 $p \geq 1$ weighted power-means. We also relax the definition to allow sample complexity to depend on the welfare function $W(\cdot; \boldsymbol{w}) \in \mathcal{W}$, but our concept of *uniform FPAC-learnability* strictly generalizes that of Cousins [2021].

Theorem 4.4 (Characterizing FPAC Learnability). Suppose axioms 1–8, and a welfare function $W_p(\cdot; \boldsymbol{w})$ that satisfies these axioms. Suppose also that the utility function u has constant range r, and sample complexity function $m_{\mathcal{H}}(\varepsilon, \delta, r, \boldsymbol{d}) \in \text{Poly}(\frac{1}{\varepsilon}, \log \frac{1}{\delta}, r, \boldsymbol{d})$. We then bound the sample complexity $m \doteq m_{W, \mathcal{H}}(\varepsilon, \delta, g, r, \boldsymbol{d})$ of FPAC learning $\mathcal{H}$ w.r.t. welfare objective $W(\cdot; \boldsymbol{w})$ as
1) $m \leq m_{\mathcal{H}}\left(\sqrt[\alpha]{\frac{\varepsilon}{2\lambda}}, \frac{\delta}{g}, r, \boldsymbol{d}\right) \in \text{Poly}\left(\sqrt[\alpha]{\lambda}, \frac{1}{\sqrt[\alpha]{\varepsilon}}, \log \frac{1}{\delta}, \log g, r, \boldsymbol{d}\right)$;
2) $p \in (0, 1] \Rightarrow m \in \text{Poly}\left(\sqrt[p]{r}, \frac{1}{\sqrt[p]{p}}, \frac{1}{\sqrt[p]{\varepsilon}}, \log \frac{1}{\delta}, \log g, \boldsymbol{d}\right)$;
3) $p = 0 \Rightarrow m \in \text{Poly}\left(\sqrt[w_{\min}]{r}, \frac{1}{w_{\min}\sqrt{\varepsilon}}, \log \frac{1}{\delta}, \log g, \boldsymbol{d}\right)$;
4) $p < 0 \Rightarrow m \in \text{Poly}\left(\frac{1}{\varepsilon}, \frac{1}{|p|\sqrt{w_{\min}}}, \log \frac{1}{\delta}, \log g, r, \boldsymbol{d}\right)$; &
5) for *any constant* $c > 0$, if $|p| \geq c$ and the *nonnegligibility condition* $\boldsymbol{w}_{\min} \geq \frac{c}{g}$ holds, then $m \in \text{Poly}\left(\frac{1}{\varepsilon}, g, \log \frac{1}{\delta}, r, \boldsymbol{d}\right)$.

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 11), item 3 follows from either IIA or from 0B (axiom 14) and SM0, and item 4 follows from either Lipschitz continuity and SWTP (axiom 12; to prevent $p = 1$) or from FC (axiom 15) — see fig. 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.

The bounds of thm. 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 class of *all fair malfare functions*, i.e., any $\mathbb{M}_p(\cdot; \boldsymbol{w})$ 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, 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 appropriately axiomatically-motivated conditions, the gap in sample complexity between the two settings is polynomial.

Sample-Efficient Learning and Estimation 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, e.g., under near-equality, power-mean functions are roughly linear, and may thus be easier to estimate or optimize. We now briefly voyage into the world of *progressive sampling* to show that efficient learning algorithms can adapt their sample consumption to the *inherent difficulty* of the task at hand (rather than *worst-case* over a class of related tasks).

Without considering the delicate intricacies of probabilistic reasoning, one might naïvely try to iteratively draw one sample per group, and terminate when welfare 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, Cousins et al., 2020, 2022b], where estimators query *a single distribution*, empirical game theoretic analysis [Viqueira et al., 2020, 2021, Cousins et al., 2023], where estimators query a noisy utility oracle at *strat-*

egy profiles, 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 $\boldsymbol{\Theta}(\log \frac{1}{\delta})$. Then, 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) may 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 can still adaptively consume about as many samples as are actually required for the task at hand.

The question remains, "How long must the 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 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

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

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

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

relative to the 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 negligible.

5 EXPERIMENTS

To demonstrate practical relevance, we present a synthetic experiment on a *welfare maximization 1-armed bandit*, 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 and $\boldsymbol{g}_2$ the minority. Utility samples are $\mathrm{UNIFORM}(\boldsymbol{u}_i - \frac{1}{2}, \boldsymbol{u}_i + \frac{1}{2})$ i.i.d. random variables with expectation $\boldsymbol{u}_i$ for each group i.

(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
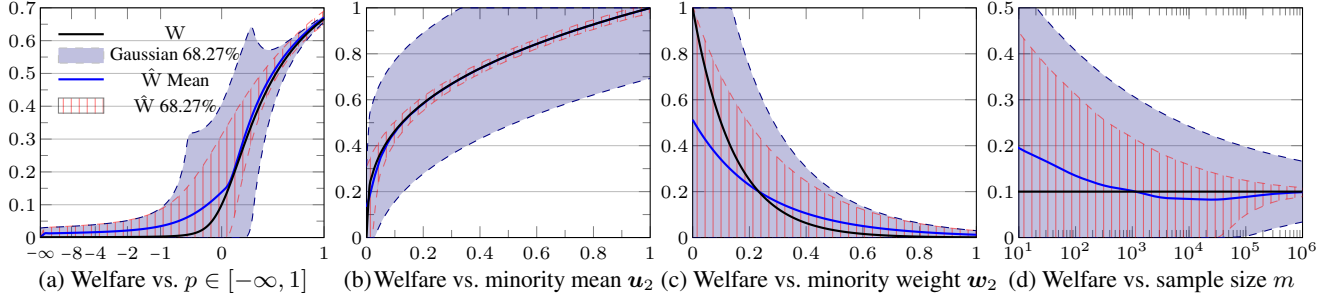
Figure 4: Estimating the Welfare of a 1-Armed Bandit with Uniform Noise. Each plot studies the response of welfare to one parameter, and the remaining parameters are selected from $p = 0$, $\boldsymbol{u} = \langle 0.999, 0.001 \rangle$, $\boldsymbol{w} = \langle \frac{2}{3}, \frac{1}{3} \rangle$, and $m = 100$. 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.

We vary the key parameters of welfare p, minority mean $\boldsymbol{u}_2$, minority 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 fig. 4. Here empirical utilities $\hat{\boldsymbol{u}}_i$ have variance $\frac{1}{12m}$, but similar results are shown for other noise models with more complicated variance structure in appx. B. We present the true welfare and approximate Gaussian $\pm 1\sigma$ (68.27%) confidence intervals on empirical welfare, i.e., $W \pm \lambda(\frac{1}{\sqrt{12m}})^{\alpha}$, 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.

Figure 4a shows the impact of changing p on the welfare. Observe that $W_p(\cdot; \boldsymbol{w})$ 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, $W_p(\cdot; \boldsymbol{w})$ for $p \approx 0$ remains difficult to estimate. 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 fig. 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 $W_0(\hat{\boldsymbol{u}}; \boldsymbol{w})$ 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 show 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 fig. 4d, we study a mitigation to this problem, by observing the impact of sample size m on $W_0(\hat{\boldsymbol{u}}; \boldsymbol{w})$. Note that by thm. 4.2 the sample complexity of $W_0(\cdot; \boldsymbol{w})$ estimation is $\mathbf{O} \frac{1}{w \sqrt[p]{\varepsilon^2}}$, whereas for $W_1(\cdot; \boldsymbol{w})$ it is $\mathbf{O}(\frac{w_2}{\varepsilon})^2$, and this larger sample complexity is manifest as slow convergence rates of all 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 than prior art. We also draw 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 fig. 3. We follow with a discussion of applications in privacy, 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 treats only the continuity and sample-complexity analysis for $p \geq 1$ (malfare), and we show that while $p < 1$ is more challenging, the difficulty of learning actually increases smoothly as $p \to 0$ *from both directions*, yielding intuitive, 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 learning settings.

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 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.

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 networks. *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.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

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, Chloe Wohlgemuth, and Matteo Riondato. BAVarian: Betweenness centrality approximation with variance-aware Rademacher averages. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, December 2022b. ISSN 1556-4681. doi: 10.1145/3577021.

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)*, 2023.

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.

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.

Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkata-subramanian. 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.

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.

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

Reshef Meir, Ariel D Procaccia, and Jeffrey S Rosenschein. Algorithms for strategyproof classification. *Artificial Intelligence*, 186:123–156, 2012.

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

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.

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 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1005–1014. ACM, 2015.

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.

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.

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 $\boldsymbol{w}' \in \triangle_{g'}$ s.t. 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) $\Rightarrow$ 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) $\boldsymbol{u}_1^\star < \boldsymbol{u}_2^\star < \boldsymbol{u}_3^\star < \dots$;
(2) for all group indices $i \in \mathcal{G}^\star$, there exists some $j \in \mathcal{G}$ such that $\boldsymbol{u}_i^\star = \boldsymbol{u}_j$; &
(3) for all group indices $i \in \mathcal{G}^\star$, it holds $\boldsymbol{w}_k^\star = \sum_{j \in \mathcal{G}} \boldsymbol{w}_j \mathbb{1}_u(\boldsymbol{u}_j)$ 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

$$\overbrace{\sum_{i \in \mathcal{G}} \boldsymbol{w}_i \mathbb{1}_u(\boldsymbol{u}_i) = \underbrace{\sum_{i \in \mathcal{G}^\star} \boldsymbol{w}_i^\star \mathbb{1}_u(\boldsymbol{u}_i^\star)}_{\text{\scriptsize UNIQUE MINIMAL REDUCTION for } \boldsymbol{u}', \boldsymbol{w}'} = \sum_{i \in \mathcal{G}'} \boldsymbol{w}_i' \mathbb{1}_u(\boldsymbol{u}_i')}^{\text{\scriptsize UNIQUE MINIMAL REDUCTION for } \boldsymbol{u}, \boldsymbol{w}} \ .$$

We may thus conclude WA. $\qquad\square$

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 (10) $\Rightarrow$ WTP (8); &
2) Suppose axioms 1–7. Then WTP (8) $\Rightarrow$ PDTP (10).

Proof. We first show item 1. Observe that the PDTP (10) $\Rightarrow$ 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,[1] 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 10 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 $\Lambda(\boldsymbol{u} + \varepsilon\boldsymbol{w}_j\mathbb{1}_i - \varepsilon\boldsymbol{w}_i\mathbb{1}_j; \boldsymbol{w})$ obeys $\Lambda(\boldsymbol{u} + \varepsilon\boldsymbol{w}_j\mathbb{1}_i - \varepsilon\boldsymbol{w}_i\mathbb{1}_j; \boldsymbol{w}) < \Lambda(\boldsymbol{u}; \boldsymbol{w})$ 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. $\qquad\square$

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 $\boldsymbol{w} \in \triangle_g$. Then:
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 10.

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 10 (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. $\qquad\square$

We now show lemma 3.7.

Lemma 3.7 (Consequences of Strong Axioms). Suppose power-mean aggregator function $\mathrm{M}_p(\cdot; \boldsymbol{w})$. Then:
1) Strengthening the SM axiom (i.e., $1 \to 11$) implies $p > 0$.
2) Strengthening the WTP axiom (i.e., $8 \to 12$) implies $p \neq 1$, thus $p < 1$ for welfare and $p > 1$ for malfare.
3) Strict PDTP (i.e., $10 \to 13$) 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 $\boldsymbol{u} \neq \boldsymbol{0}$ and $\min_{i \in \mathcal{G}} \boldsymbol{u}_i = 0$, then $\big(\mathrm{M}_p(\boldsymbol{u}; \boldsymbol{w}) = 0\big) \Leftrightarrow (p \leq 0)$." In particular, here $\mathrm{M}(\boldsymbol{u}; \boldsymbol{w}) = \mathrm{M}(\boldsymbol{0}; \boldsymbol{w}) = 0$ for $\boldsymbol{u} \neq \boldsymbol{0}$ would violate axiom 11, thus by contraposition, axiom 11 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 $\boldsymbol{u} \neq \boldsymbol{0}$, it holds $p > 0 \Rightarrow \mathrm{M}_p(\boldsymbol{u}; \boldsymbol{w}) > 0$, thus by the contrapositive, $\mathrm{M}_p(\boldsymbol{u}; \boldsymbol{w}) = 0 \Rightarrow p \leq 0$. To see the converse, first observe that since $\min_i \boldsymbol{u}_i = 0$, it holds $p = 0 \Rightarrow \mathrm{M}_p(\boldsymbol{u}; \boldsymbol{w}) = 0$. The case of $p < 0$

---

[1] 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.

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 $\boldsymbol{u} \neq \boldsymbol{0}$, $p \leq 0 \Rightarrow \mathrm{M}_p(\boldsymbol{u}; \boldsymbol{w}) = 0$. 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 $p = 1 \Rightarrow \neg(\text{axiom 12})$, hence $(\text{axiom 12}) \Rightarrow p \neq 1$, 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 13 (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 14) $\Leftrightarrow p \leq 0$. 2) FC (axiom 15) $\Leftrightarrow p < 0$.

Proof. We first show item 1. We first show that 0B (axiom 14) $\Rightarrow p \leq 0$. 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., $p \leq 0 \Rightarrow$ 0B (axiom 14). 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

$$\lim_{\boldsymbol{u}_i \to 0^+} \mathrm{M}_p(\boldsymbol{u}; \boldsymbol{w}) = \lim_{\boldsymbol{u}_i \to 0^+} \lim_{\varepsilon \to 0^+} \left(\frac{1}{\sum_{i=1}^g \frac{\boldsymbol{w}_i}{(\boldsymbol{u}_i + \varepsilon)^{-p}}} \right)^{-\frac{1}{p}} \qquad \text{DEFINITION 3.4 (POWER MEAN)}$$

$$= \left(\frac{1}{\lim_{\boldsymbol{u}_i \to 0^+} \lim_{\varepsilon \to 0^+} \sum_{i=1}^g \frac{\boldsymbol{w}_i}{(\boldsymbol{u}_i + \varepsilon)^{-p}}} \right)^{-\frac{1}{p}} \qquad \text{LIMIT LAWS}$$

$$= \left(\frac{1}{\infty} \right)^{-\frac{1}{p}} = 0 \ . \qquad \text{LIMIT LAWS}$$

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

$$\lim_{\boldsymbol{u}_i \to 0^+} \mathrm{M}_0(\boldsymbol{u}; \boldsymbol{w}) = \lim_{\boldsymbol{u}_i \to 0^+} \lim_{\rho \to 0} \lim_{\varepsilon \to 0^+} \mathrm{M}_\rho(\boldsymbol{u} + \varepsilon \boldsymbol{1}; \boldsymbol{w}) \qquad \text{DEFINITION 3.4 (POWER MEAN)}$$

$$= \lim_{\boldsymbol{u}_i \to 0^+} \lim_{\varepsilon \to 0^+} \prod_{i=1}^g (\boldsymbol{u}_i + \varepsilon)^{\boldsymbol{w}_i} \qquad \text{GEOMETRIC MEAN LIMIT}$$

$$= \lim_{\boldsymbol{u}_i \to 0^+} \lim_{\varepsilon \to 0^+} \exp\left(\sum_{i=1}^g \boldsymbol{w}_i \ln (\boldsymbol{u}_i + \varepsilon) \right) \qquad \text{LOGARITHMIC IDENTITIES}$$

$$= \exp\left(\lim_{\boldsymbol{u}_i \to 0^+} \lim_{\varepsilon \to 0^+} \sum_{i=1}^g \boldsymbol{w}_i \ln (\boldsymbol{u}_i + \varepsilon) \right) \qquad \text{LIMIT LAWS}$$

$$= \exp(-\infty) = 0 \ . \qquad \text{LIMIT LAWS}$$

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, $\boldsymbol{u} \doteq \boldsymbol{1}$, it holds that $\lim_{c \to \infty} M_p(\boldsymbol{u} + c\mathbb{1}_i; \boldsymbol{w}) = \infty$ if and only if $p \geq 0$. In particular, this can be seen by observing that, for the forward direction, that $p < 0 \Rightarrow M_p(\boldsymbol{u}; \boldsymbol{w}) \leq \sqrt[p]{\boldsymbol{w}_j}\boldsymbol{u}_j \leq \infty$ for any $j \in \mathcal{G}$, thus $p \geq 0$, and for the reverse direction,

$$M_0(\boldsymbol{u} + c\mathbb{1}_i; \boldsymbol{w}) = \lim_{c \to \infty}(1+c)^{\boldsymbol{w}_i} = \infty \ ,$$

and then that $M_0(\boldsymbol{u}; \boldsymbol{w}) \leq M_p(\boldsymbol{u}; \boldsymbol{w})$ 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

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

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

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

and observe that the RHS is, by definition, the FC axiom. $\square$

We now show lemma 3.10.

Lemma 3.10 (Power-Mean Differentiation). Suppose $\boldsymbol{u}_{\backslash i} \succ 0$, 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 $\frac{\partial}{\partial \boldsymbol{u}_i} M_p(\boldsymbol{u}; \boldsymbol{w}) = \frac{\boldsymbol{w}_i \boldsymbol{u}_i^{p-1}}{M_p^{p-1}(\boldsymbol{u}; \boldsymbol{w})}$.
2) If $p < 0$, then $\lim_{\boldsymbol{u}_i \to 0^+} \frac{\partial}{\partial \boldsymbol{u}_i} M_p(\boldsymbol{u}; \boldsymbol{w}) = \sqrt[-p]{\frac{1}{\boldsymbol{w}_i}}$.
3) If $p \in [0, 1)$, then $\lim_{\boldsymbol{u}_i \to 0^+} \frac{\partial}{\partial \boldsymbol{u}_i} M_p(\boldsymbol{u}; \boldsymbol{w}) = \infty$.

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

$$\lim_{\boldsymbol{u}_i \to 0^+} \frac{\partial}{\partial \boldsymbol{u}_i} M_p(\boldsymbol{u}; \boldsymbol{w}) = \lim_{\boldsymbol{u}_i \to 0^+} \frac{\boldsymbol{w}_i \boldsymbol{u}_i^{p-1}}{M_p^{p-1}(\boldsymbol{u}; \boldsymbol{w})} \qquad \text{ITEM 1}$$

$$= \lim_{\boldsymbol{u}_i \to 0^+} \boldsymbol{w}_i M_p^{1-p}\left(j \mapsto \frac{\boldsymbol{u}_j}{\boldsymbol{u}_i}; \boldsymbol{w}\right) \qquad \text{MULTIPLICATIVE LINEARITY}$$

$$= \boldsymbol{w}_i \left(\sum_{j=1}^g \boldsymbol{w}_j \lim_{\boldsymbol{u}_i \to 0^+}\left(\frac{\boldsymbol{u}_j}{\boldsymbol{u}_i}\right)^p\right)^{\frac{1-p}{p}} \qquad \text{LIMIT LAWS}$$

$$= \boldsymbol{w}_i \left(\boldsymbol{w}_i 1^p + \sum_{j=1, j \neq i}^g \boldsymbol{w}_j 0^p\right)^{\frac{1-p}{p}} \qquad \text{LIMIT LAWS}$$

$$= \boldsymbol{w}_i^{1+\frac{1-p}{p}} = \boldsymbol{w}_i^{\frac{1}{p}} = \sqrt[-p]{\frac{1}{\boldsymbol{w}_i}} \ . \qquad \text{ALGEBRA}$$

We now show item 3. We split this case into two subcases; namely the $p = 0$ and $p > 0$ subcases, essentially because whether $M_p(\boldsymbol{u}; \boldsymbol{w}) = 0$ 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

$$
\lim_{\boldsymbol{u}_i \to 0^+} \frac{\partial}{\partial \boldsymbol{u}_i} \mathrm{M}_p(\boldsymbol{u}; \boldsymbol{w}) = \lim_{\boldsymbol{u}_i \to 0^+} \frac{\boldsymbol{w}_i \boldsymbol{u}_i^{p-1}}{\mathrm{M}_p^{p-1}(\boldsymbol{u}; \boldsymbol{w})} \qquad\qquad \text{ITEM 1}
$$

$$
= \boldsymbol{w}_i \frac{\lim_{\boldsymbol{u}_i \to 0^+} \boldsymbol{u}_i^{p-1}}{\lim_{\boldsymbol{u}_i \to 0^+} \mathrm{M}_p^{p-1}(\boldsymbol{u}; \boldsymbol{w})} \qquad\qquad \text{LIMIT LAWS}
$$

$$
= \underbrace{\boldsymbol{w}_i \mathrm{M}_p^{1-p}(\boldsymbol{u}; \boldsymbol{w})}_{\text{POSITIVE FINITE}} \left(\lim_{\boldsymbol{u}_i \to 0^+} \frac{1}{\boldsymbol{u}_i} \right)^{1-p} = \infty \; . \qquad\qquad \text{ALGEBRA}
$$

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. $\qquad\square$

We now show lemma 3.12.

Lemma 3.12 (Power-Mean Lipschitz Continuity). Suppose $p \in \mathbb{R}$, sentiment vectors $\boldsymbol{u}, \boldsymbol{u}' \in \mathbb{R}_{0+}^g$, and weights vector $\boldsymbol{w} \in \triangle_g$. The following then hold.
1) If $p \geq 1$, then $\left| \mathrm{M}_p(\boldsymbol{u}; \boldsymbol{w}) - \mathrm{M}_p(\boldsymbol{u}'; \boldsymbol{w}) \right| \leq \mathrm{M}_p(\left| \boldsymbol{u} - \boldsymbol{u}' \right|; \boldsymbol{w})$, and $\mathrm{M}_p(\cdot; \boldsymbol{w})$ is $\sqrt[p]{\boldsymbol{w}_{\max}} \|\cdot\|_1$ and $1 \|\cdot\|_\infty$ Lipschitz.
2) If $p < 0$, 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 exhibit monotonically increasing convexity, thus derivatives increase as $\boldsymbol{u}_i \to \infty$, whereas welfare functions 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. Observe that $\left| \mathrm{M}_p(\boldsymbol{u}; \boldsymbol{w}) - \mathrm{M}_p(\boldsymbol{u}'; \boldsymbol{w}) \right| \leq \mathrm{M}_p(\left| \boldsymbol{u} - \boldsymbol{u}' \right|; \boldsymbol{w})$ follows via the subadditivity of $p \geq 1$ power-mean functions, i.e., they are convex and have the unique zero of $\mathrm{M}_p(\boldsymbol{0}; \boldsymbol{w}) = 0$.

From here, the $\|\cdot\|_1$ Lipschitz constant can be derived via maximizing derivatives, i.e., applying lemma 3.10 item 1. 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, and the $1 \|\cdot\|_\infty$ Lipschitz characterization follows directly from $\left| \mathrm{M}_p(\boldsymbol{u}; \boldsymbol{w}) - \mathrm{M}_p(\boldsymbol{u}'; \boldsymbol{w}) \right| \leq \left\| \boldsymbol{u} - \boldsymbol{u}' \right\|_\infty$.

Finally, to see the $\|\cdot\|_\infty$ Lipschitz constant, observe that $\mathrm{M}_p(\left| \boldsymbol{u} - \boldsymbol{u}' \right|; \boldsymbol{w}) \leq \mathrm{M}_\infty(\left| \boldsymbol{u} - \boldsymbol{u}' \right|; \boldsymbol{w}) = \left\| \boldsymbol{u} - \boldsymbol{u}' \right\|_\infty$, 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. $\qquad\square$

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) *General Welfare Hölder Condition*: Suppose $p \leq 1$. Then $\left| \mathrm{M}_p(\boldsymbol{u} + \varepsilon \mathbb{1}_i; \boldsymbol{w}) - \mathrm{M}_p(\boldsymbol{u}; \boldsymbol{w}) \right| \leq r^{1 - \boldsymbol{w}_i} \varepsilon^{\boldsymbol{w}_i}$, and $\mathrm{M}_p(\cdot; \boldsymbol{w})$ is $r^{1 - \boldsymbol{w}_{\min}} \cdot \boldsymbol{w}_{\min} \|\cdot\|_\infty$ Hölder continuous.

2) *Positive Welfare Hölder Condition*: Suppose $p \in (0, 1]$. Then $\left| M_p(\boldsymbol{u} + \varepsilon \mathbb{1}_i; \boldsymbol{w}) - M_p(\boldsymbol{u}; \boldsymbol{w}) \right| \leq r^{1-p} \frac{\boldsymbol{w}_i}{p} \varepsilon^p$. This implies $M_p(\cdot; \boldsymbol{w})$ 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-$\|\cdot\|_{1,\boldsymbol{w}}$ (where $\|\boldsymbol{u}\|_{1,\boldsymbol{w}} \doteq |\boldsymbol{u}| \cdot \boldsymbol{w}$); &
C) $r^{1-p} \frac{1}{p}$-p-$\|\cdot\|_\infty$.

Proof. Due to the complex 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

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

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 $x \mapsto r f(\frac{x}{r})$ 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 $g(x) \doteq r f(\frac{x}{r})$ obeys

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

We now show item 1. We first consider the case of $p = 0$, i.e., we analyze the Nash social welfare $W_0(\cdot; \boldsymbol{w})$. In this case, observe that the most rapid change to $M_0(\boldsymbol{u}; \boldsymbol{w})$ 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

$$\sup_{\boldsymbol{u}, \varepsilon} \frac{\left| M_0(\boldsymbol{u} + \varepsilon \mathbb{1}_i; \boldsymbol{w}) - M_0(\boldsymbol{u}; \boldsymbol{w}) \right|}{\varepsilon^{\boldsymbol{w}_i}} \leq \frac{1}{\varepsilon^{\boldsymbol{w}_i}} \left| M_0(\boldsymbol{0} + \varepsilon \mathbb{1}_i + \mathbb{1}_{\mathcal{G} \setminus \{i\}}; \boldsymbol{w}) - M_0(\boldsymbol{0} + \mathbb{1}_{\mathcal{G} \setminus \{i\}}; \boldsymbol{w}) \right| \qquad \begin{matrix} \text{\footnotesize CONCAVITY} \\ \text{\footnotesize MONOTONICITY} \end{matrix}$$

$$= \frac{1}{\varepsilon^{\boldsymbol{w}_i}} \exp \left(\boldsymbol{w}_i \ln(\varepsilon) + (1 - \boldsymbol{w}_i) \ln(1) \right) \qquad \text{\footnotesize DEFINITION OF } M_0(\cdot; \boldsymbol{w})$$

$$= \frac{1}{\varepsilon^{\boldsymbol{w}_i}} \exp \left(\boldsymbol{w}_i \ln(\varepsilon) \right) = 1 \ , \qquad \text{\footnotesize ALGEBRA}$$

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 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 $M_p(\boldsymbol{u} + \varepsilon \mathbb{1}_i; \boldsymbol{w}) - M_p(\boldsymbol{u}; \boldsymbol{w})$ is decreasing as $p \to 0$, from both the positive and negative sides. We thus conclude that λ-α-$|\cdot|$ Hölder continuity for $M_0(\boldsymbol{u}; \boldsymbol{w})$ implies the same for $M_p(\boldsymbol{u}; \boldsymbol{w})$.

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

$$\sup_{\boldsymbol{u}, \varepsilon} \frac{\left| M_p(\boldsymbol{u} + \varepsilon \mathbb{1}_i; \boldsymbol{w}) - M_p(\boldsymbol{u}; \boldsymbol{w}) \right|}{\varepsilon^p} = \sup_{\varepsilon \in (0,1)} \frac{\left| M_p(\langle \varepsilon, 1 \rangle; \langle \boldsymbol{w}_i, 1 - \boldsymbol{w}_i \rangle) - M_p(\langle 0, 1 \rangle; \langle \boldsymbol{w}_i, 1 - \boldsymbol{w}_i \rangle) \right|}{\varepsilon^p} \qquad \begin{array}{l} \text{CONCAVITY} \\ \text{MONOTONICITY} \end{array}$$

$$= \sup_{\varepsilon \in (0,1)} \frac{\left(\boldsymbol{w}_i \varepsilon^p + (1 - \boldsymbol{w}_i) \right)^{\frac{1}{p}} - (1 - \boldsymbol{w}_i)^{\frac{1}{p}}}{\varepsilon^p} \qquad \text{DEFINITION OF } M_p(\cdot; \boldsymbol{w})$$

$$\leq \sup_{\varepsilon \in (0,1)} \frac{\frac{1}{p} \boldsymbol{w}_i \varepsilon^p + (1 - \boldsymbol{w}_i)^{\frac{1}{p}} - (1 - \boldsymbol{w}_i)^{\frac{1}{p}}}{\varepsilon^p} \qquad \text{SEE BELOW}$$

$$= \frac{\boldsymbol{w}_i}{p} \ . \qquad \text{ALGEBRA}$$

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 linearity (considering the weighted norm). Finally, item C follows from item B by noting that, for any $\boldsymbol{u} \in \mathbb{R}_{0+}^g$ and weights vector $\boldsymbol{w} \in \triangle_g$, it holds that $\|\boldsymbol{u}\|_{1, \boldsymbol{w}} \leq \|\boldsymbol{u}\|_\infty \leq \|\boldsymbol{u}\|_1$. $\qquad \square$

A.2 Analysis of Fair-PAC Learning

We now show theorem 4.1.

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

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

Consequently, the *empirical welfare maximizer*

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

approximates the *true welfare maximizer*

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

in terms of welfare-optimality, as it holds that

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

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

$$W\left(i \mapsto \mathbb{E}_{\mathcal{D}_i} [u \circ \hat{h}]; \boldsymbol{w} \right) \geq W\left(i \mapsto \hat{\mathbb{E}}_{\hat{\mathcal{D}}_i} [u \circ \hat{h}]; \boldsymbol{w} \right) - \lambda \|\boldsymbol{\varepsilon}\|_W^\alpha \qquad \text{FIRST PORTION ON } \hat{h}$$

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

$$\geq W\left(i \mapsto \mathbb{E}_{\mathcal{D}_i} [u \circ h^\star]; \boldsymbol{w} \right) - 2\lambda \|\boldsymbol{\varepsilon}\|_W^\alpha \ . \qquad \text{FIRST PORTION ON } h^\star$$

We now show theorem 4.2. $\qquad \square$

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

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

is sufficient, i.e., for at least this many samples per group, (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 $\mathrm{m}_{\mathcal{H}}\left(\sqrt[\alpha]{\frac{\varepsilon}{\lambda}}, \delta, r, \boldsymbol{d}\right)$ 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 $\mathrm{m}_{\mathcal{H}}\left(\sqrt[\alpha]{\frac{\varepsilon}{\lambda}}, \frac{\delta}{g}, r, \boldsymbol{d}\right)$ 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 $\|\boldsymbol{\varepsilon}\|_\infty \leq \sqrt[\alpha]{\frac{\varepsilon}{\lambda}}$. Then, applying theorem 4.1 yields

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

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

We now show theorem 4.4.

Theorem 4.4 (Characterizing FPAC Learnability). Suppose axioms 1–8, and a welfare function $\mathrm{W}_p(\cdot; \boldsymbol{w})$ that satisfies these axioms. Suppose also that the utility function u has constant range r, and 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 $m \doteq \mathrm{m}_{\mathrm{W}, \mathcal{H}}(\varepsilon, \delta, g, r, \boldsymbol{d})$ of FPAC learning $\mathcal{H}$ w.r.t. welfare objective $\mathrm{W}(\cdot; \boldsymbol{w})$ 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) $p = 0 \Rightarrow m \in \mathrm{Poly}\left(\sqrt[\boldsymbol{w}_{\min}]{r}, \frac{1}{\boldsymbol{w}_{\min}\sqrt{\varepsilon}}, \log\frac{1}{\delta}, \log g, \boldsymbol{d}\right)$;

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

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

Proof. In each case, the 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 $|p| \geq c$ 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* $\boldsymbol{w}_{\min} \geq \frac{c}{g}$, this allows us to control the dependence of the Lipschitz constant λ for $p \leq -c$ as $\lambda \leq \frac{1}{|p|\sqrt{\boldsymbol{w}_{\min}}} \leq \left(\frac{g}{c}\right)^{\frac{1}{c}} \in \mathrm{Poly}(g)$ for $p \leq -c$. Similarly, for $p \geq c$, note that for welfare functions we need only consider $c \leq p \leq 1$, and observe that for $\alpha = c$, we have $\lambda = \frac{r^{1-c}}{c}$ by lemma 3.13 item 2C, i.e., $\frac{1}{\alpha}$ and λ are bounded above by constants, and thus again yield no exponential dependencies. $\square$

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 expectation $\boldsymbol{u}_i$ for each group i. For the Bernoulli model, we use $\mathrm{BERNOULLI}(\boldsymbol{u}_i)$ random variables, and for the beta model, we use $\mathrm{BETA}(\boldsymbol{u}_i, 1 - \boldsymbol{u}_i)$, which acts as continuous approximation of a $\mathrm{BERNOULLI}(\boldsymbol{u}_i)$ 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 $\frac{\boldsymbol{u}_i(1-\boldsymbol{u}_i)}{\boldsymbol{m}_i} \leq \frac{1}{4\boldsymbol{m}_i}$ in the Bernoulli case, or $\frac{\boldsymbol{u}_i(1-\boldsymbol{u}_i)}{2\boldsymbol{m}_i} \leq \frac{1}{8\boldsymbol{m}_i}$ 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
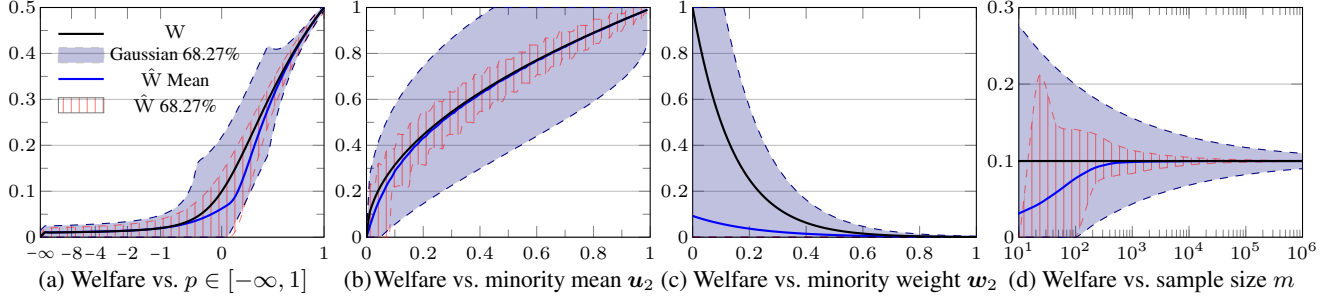
Figure A1: Estimating the Welfare of a 1-Armed Bandit under Bernoulli Noise. Each plot studies the response of welfare to one parameter, and the remaining parameters are selected from $p = 0$, $\boldsymbol{u} = \langle 0.99, 0.01 \rangle$, $\boldsymbol{w} = \langle \frac{1}{2}, \frac{1}{2} \rangle$, 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
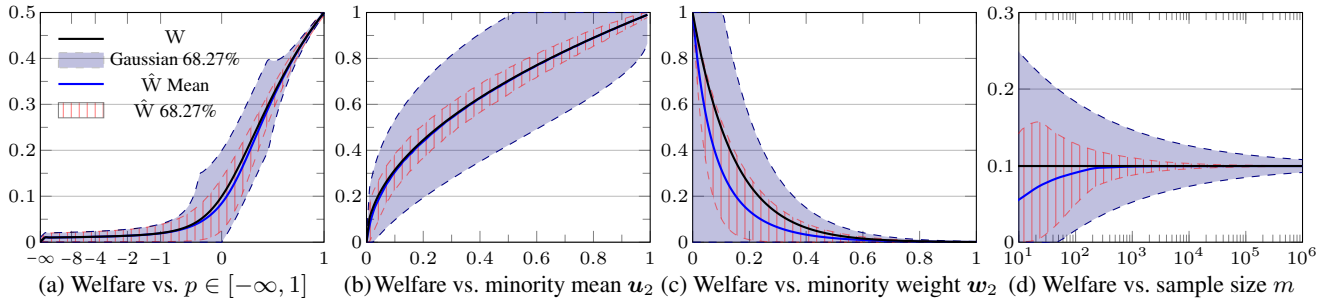
Figure A2: Estimating the Welfare of a 1-Armed Bandit under Beta Noise. Each plot studies the response of welfare to one parameter, and the remaining parameters are selected from $p = 0$, $\boldsymbol{u} = \langle 0.99, 0.01 \rangle$, $\boldsymbol{w} = \langle \frac{1}{2}, \frac{1}{2} \rangle$, and $m = 50$. 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).