Algorithms and Analysis for Optimizing Robust Objectives in Fair Machine Learning

Cyrus Cousins

University of Massachusetts Amherst
Columbia Workshop on Fairness in Operations and AI

December 2023

Abstract

The *original position* or *veil of ignorance* argument of John Rawls, perhaps the most famous argument for egalitarianism, states that our concept of fairness, justice, or welfare should be decided from behind a veil of ignorance, and thus must consider everyone impartially (invariant to our identity). This can be posed as a zero-sum game, where a Dæmon constructs a world, and an adversarial Angel then places the Dæmon into the world. This game incentivizes the Dæmon to maximize the minimum utility over all people (i.e., to maximize *egalitarian welfare*). In some sense, this is the most extreme form of *risk aversion* or *robustness*, and we show that by weakening the Angel, milder robust objectives arise, which we argue are effective *robust proxies* for *fair* learning or allocation tasks. In particular, the utilitarian, Gini, and power-mean welfare concepts arise from special cases of the adversarial game, which has philosophical implications for the understanding of each of these concepts. We also motivate a new fairness concept that essentially fuses the nonlinearity of the power-mean with the piecewise nature of the Gini class. Then, exploiting the relationship between fairness and robustness, we show that these robust fairness concepts can all be efficiently optimized under mild conditions via standard maximin optimization techniques. Finally, we show that such methods apply in machine learning contexts, and moreover we show generalization bounds for robust fair machine learning tasks.

Keywords:

Fair Machine Learning *Rawlsian Ethics* *Adversarial Learning* *Convex Optimization* *Robust Fair Learning*

1 Introduction

Fairness and robustness are crucial aspects of machine learning and allocation systems, both of which are generally addressed through modelling, data collection, and objective selection. This work extends ideas and objectives in welfare-centric fair machine learning and optimization introduced by Cousins [2021a,b, 2022, 2023]. We derive robust variants of fair objectives, and explore mathematical and philosophical connections between robustness and fairness. In particular, we consider robust *welfare functions*, which aggregate *utility* across a population, and robust *malfare functions*, which aggregate *disutility*, both serving as fairness metrics and as optimization targets. We then combine these robust objectives with adversarial optimization theory and techniques, which expands on the relationship between fairness, robustness, and uncertainty in machine learning and allocation problems [Mazzetto et al., 2021, Dong and Cousins, 2022, Cousins et al., 2023a].

The core of this paper is the construction of a hierarchy of Rawlsian games, where a Dæmon is tasked with creating a world, and an Angel places them within it. We consider various modifications and restrictions of this basic setup by adjusting the action space of the agents, as well as the payoff function, and show that various game theoretic solution concepts, including adversarial play for constant sum games and Nash equilibria for general sum games, give rise to various welfare and malfare concepts. Of course, this game is a metaphor, but it is strongly motivated by the grounded social planner's problem, wherein a social planner seeks to organize society in a way that is favorable to all, and from these games we derive insight as to how the social planner should behave. The goals of this paper and the purpose of constructing this game are manifold.

1) We provide philosophical insight into a large class of welfare and malfare functions. Section 4.1 draws connections between fairness and robustness, finding that many classical welfare functions can be understood as robust utilitarian welfare in our game. We also show in section 4.2 that some welfare (malfare) concepts arise from a class of concave utility transforms (convex disutility transforms). These derivations complement direct fairness-based understandings of these welfare concepts from cardinal welfare theory. Viewing the prism of fairness from these three angles yields deeper philosophical, mathematical, and algorithmic understanding.
2) We argue that utilitarian and egalitarian welfare or malfare are two ends of a spectrum, and derive a novel class of welfare (malfare) functions, which we term the Gini power-mean class, that falls between these extremes. In particular, both the Gini and power-mean classes also have this property, and our Gini power-mean class *strictly contains* utilitarian, egalitarian, and the entire Gini and power-mean families. Furthermore, our generic robustness analysis allows us to define and motivate robust variants of this class that are still more general.
3) Leveraging the connections between fairness, robustness, and robust fairness, we show in section 6 that for various applications in allocation and machine learning, our objectives can be efficiently optimized. In particular, we consider optimization either through standard maximin optimization techniques, or in some cases by reduction to a simpler maximization problem. We show mild conditions under which the robust optimization problem has convex-concave structure amenable to first-order (gradient descent-ascent) methods. We also show certain special cases in continuous allocation problems that reduce to linear programming or quadratic programming.

It is important to note that robustness and fairness are not the same thing, but they are deeply related. In section 4.1, we derive the Gini social welfare family as the solution to a robust utilitarian optimization problem, but it's worth noting that the Gini family on its own also arises as the unique solution set to a set of cardinal welfare axioms based on fairness that really have nothing to do with robustness. Similarly, egalitarian or "worst case over groups" objectives inherently have some robustness, the form of optimizing them resembles robust objectives, and their derivation via Rawls' original position argument (section 4) resembles robustness against an adversary, they can also be axiomatically derived from cardinal welfare theory, if we start with Gini axioms, then strengthen them to require that transferring utility from any group i to any group j with lower utility is beneficial, even if the transfer is *arbitrarily inefficient*, i.e., for all i, j such that $\boldsymbol{s}_i < \boldsymbol{s}_j$, for any *transfer efficiency* $\gamma > 0$, there exists some *transfer magnitude* α such that

$$\mathrm{M}(\boldsymbol{s}; \boldsymbol{w}) < \mathrm{M}(\boldsymbol{s} + \alpha\gamma\mathbb{1}_i - \alpha\mathbb{1}_j) \ ,$$

i.e., any infinitesimal equitable transfer of utility is beneficial, no matter how inefficient.

Overview of Contributions We describe in section 4 John Rawls' original position argument, followed by several generalizations in which the adversary is weakened, which give rise to various robust fairness concepts, including the utilitarian, Gini, and power-mean welfare concepts, and robust variants thereof. In section 5, we show that our robust proxies of the standard fairness concepts yield probabilistic or adversarial guarantees in terms of their non-robust counterparts. This mathematical motivation complements the philosophical motivation of the previous section. Then, section 6 shows that we can efficiently optimize these fair and robust fair objectives in a variety of allocation and machine learning settings. Finally, in section 7 we analyze the continuity properties of robust fair objectives, and we show generalization bounds for robust fair machine learning tasks.

2 Related Work

In his seminal work, Rawls [1971, 2001] connects fairness, justice, social welfare, and robustness to uncertainty. His *original position* or *veil of ignorance* arguments apply Wald's maximin principle to derive the *egalitarian welfare*, i.e., the principal that we should measure the overall wellbeing of society in terms of its least well-off member, and the social planner should seek to maximize this minimum utility. The Rawlsian school of thought contrasts the earlier prevailing utilitarian theory: *Utilitarian welfare* [Bentham, 1789, Mill, 1863] instead measures overall wellbeing as the *sum* or *average* utility across a population.

However, these are not the only justice criteria of interest. Alternative characterizations of welfare lead to the power-mean class or the Gini class (discussed in section 3, both of which contain the egalitarian and utilitarian welfare as special cases. Indeed, the wellbeing of society overall and of disadvantaged or minority groups is well-studied in welfare economics [Pigou, 1912, Dalton, 1920, Debreu, 1959, Gorman, 1968] and moral philosophy [Parfit, 1997]. Generally speaking, utilitarian and egalitarian welfare stand at two extremes of a spectrum, and *prioritarian* concepts lie somewhere in between [Parfit, 1997, Arneson, 2000]. Utilitarianism is criticized for not

incentivizing *equitable redistribution* of (dis)utility, and egalitarianism is criticized for ignoring all but the *most disadvantaged* groups in society. In contrast, *prioritarianism* encompasses various justice criteria that *prioritize* the wellbeing of the impoverished, without ignoring others, making tradeoffs between them in various ways.

Amadae [2003], Gališanka [2017] discuss the game-theoretic implication of Rawlsian philosophy, and this work considers modifications of a game-theoretic statement of Rawls' original position argument. Nozick [1974] criticizes Rawlsian theory as overly risk-averse, and this work addresses this point by introducing less risk-averse variants of the original position argument. Rawlsian theory is also criticized as unsuitable as a basis for morality [Harsanyi, 1975]; this work makes progress in this direction, by explicitly grounding the application of Wald's [1939, 1945] maximin principle in epistemic uncertainty, and by contrasting Rawlsian welfare with other fairness concepts.

Rawlsian theory has been applied to fair machine learning and algorithmic justice [Ashrafian, 2023], often termed *minimax fair learning* [Diana et al., 2021, Shekhar et al., 2021, Abernethy et al., 2022], and in particular Lokhande et al. [2022], Dong and Cousins [2022] optimize Rawlsian objectives under uncertainty. More general concepts of fair machine learning in terms of other welfare or malfare concepts are also explored in the literature; Thomas et al. [2019] introduces the *Seldonian learner* framework, and Cousins [2021a,b, 2022, 2023] defines *fair-PAC learning*, which both deal with computational and statistical issues arising from optimizing nonlinear fairness objectives.

Other work seeks to optimize welfare-concepts in more specific instances; e.g., in supervised classification [Hu and Chen, 2020, Rolf et al., 2020], in contextual bandits [Metevier et al., 2019], or in reinforcement learning [Siddique et al., 2020, Cousins et al., 2022], generally finding the resulting optimization problems to be tractable. Some authors also seek to apply *fairness constraints* based on welfare [Hu and Chen, 2020, Heidari et al., 2018, Speicher et al., 2018]; Hu and Chen [2020] finds that, in contrast to demographic parity constraints, these are usually at least *convex* (assuming appropriate utility and welfare function choice).

3 Preliminaries

The *Pigou-Dalton transfer principle* [Pigou, 1912, Dalton, 1920] and the *Debreu-Gorman axioms* [Debreu, 1959, Gorman, 1968] lead all welfare functions to concord with sums of *logarithms* or *powers* of utilities, i.e., for g groups and utility vectors $\boldsymbol{s} \in \mathbb{R}^g_{0+}$, for some $p \in \mathbb{R}$, all fairness concepts $\mathrm{M}(\boldsymbol{s})$ define a *partial ordering* over utility vectors that agrees with

$$\mathrm{M}(\boldsymbol{s}) = \mathrm{sgn}(p) \sum_{i=1}^{g} \boldsymbol{s}_i^p \ , \quad \text{or} \quad \mathrm{M}(\boldsymbol{s}) = \sum_{i=1}^{g} \ln(\boldsymbol{s}_i) \ . \tag{1}$$

Weights vectors $\boldsymbol{w} \in \triangle_g$, where $\triangle_g$ denotes the unit probability simplex over g values (excluding 0 values), are essential to this work. Cousins [2021a, 2023] introduces weighted variants of the Debreu-Gorman axioms, as well as *multiplicative linearity* and *unit scale* axioms, which essentially standardize the cardinal values of aggregator functions. Utility and disutility are generically referred to as *sentiment*, and cardinal welfare theory applies equally well to aggregation of disutility (malfare functions). These novel axioms, when combined with the Debreu-Gorman axioms, characterize the *weighted power-mean family* of aggregator functions, defined below.

Definition 3.1 (Weighted Power-Mean Family)**.** Suppose some power parameter $p \in \mathbb{R}$ and weights vector $\boldsymbol{w} \in \triangle_g$. For any $\boldsymbol{s} \in \mathbb{R}^g_{0+}$, we define

$$\mathrm{M}_p(\boldsymbol{s}; \boldsymbol{w}) = \sqrt[p]{\sum_{i=1}^{g} \boldsymbol{w}_i \boldsymbol{s}_i^p} \text{ for } p \neq 0 \ , \quad \mathrm{M}_0(\boldsymbol{s}; \boldsymbol{w}) = \exp\left(\sum_{i=1}^{g} \boldsymbol{w}_i \ln(\boldsymbol{s}_i) \right) \ , \quad \text{or} \quad \mathrm{M}_{\pm\infty}(\boldsymbol{s}; \boldsymbol{w}) = \min_{i \in 1,\ldots,g}^{\max} \boldsymbol{s}_i \ . \tag{2}$$

The taking the limit as $p \to 0$ yields the $p = 0$ case, known as the *Nash social welfare* or *geometric mean*, and the limits as $p \to \pm\infty$ yield the *egalitarian* welfare or malfare.

Power-means with $p \in (-\infty, 1)$ are valid welfare functions, as maximizing them strictly incentivizes equitable redistribution of utility (except around 0). Similarly, power-means with $p \in (1, \infty)$ are valid malfare functions, as minimizing them strictly incentivizes equitable redistribution of disutility. The utilitarian and egalitarian endpoints of these open intervals ($p \in \{-\infty, 1, \infty\}$) are generally also considered valid, as they arise as limiting sequences of valid welfare or malfare functions, they still satisfy most of the same cardinal welfare axioms as the power-mean family, and they at least weakly incentivize equitable redistribution.

Power-means in general require *nonnegative sentiment* to remain well-defined, real-valued, and preserve their curvature, so when working with them we restrict the sentiment vector space to $\boldsymbol{s} \in \mathbb{R}^g_{0+}$, but for other aggregator function classes, we may relax this assumption to $\boldsymbol{s} \in \mathbb{R}^g$.

The Rawlsian Game

g: *Number of inhabitants* in the game world, labeled $1, \ldots, g$.
Θ: Set of *feasible parameters* for worlds the Dæmon can create.
$\boldsymbol{s}(\theta) : \Theta \to \mathbb{R}^g$: Sentiment vector of the inhabitants of some world parameterized by θ.
$\mathcal{S} \doteq \{\boldsymbol{s}(\theta) | \, \theta \in \Theta\} \subseteq \mathbb{R}^g$: Set of *feasible sentiment vectors* the Dæmon can create.
$\mathcal{A}_{\text{Dæ}} \doteq \mathcal{S}$: Dæmon action space.
$\mathcal{A}_{\text{Ang}} \doteq \{1, \ldots, g\}$: Angel action space.
$\mathrm{P}(\boldsymbol{s}, i) \doteq \langle \boldsymbol{s}_i, -\boldsymbol{s}_i \rangle$: Zero-sum payoff function.

Strategic gameplay (Dæmon goes first):

$$\arg \min_{\boldsymbol{s} \in \mathcal{A}_{\text{Dæ}}} \max_{i \in \mathcal{A}_{\text{Ang}}} \mathrm{P}_1(\boldsymbol{s}, i) = \arg \min_{\boldsymbol{s} \in \mathcal{S}} \max_{i \in 1, \ldots, g} \boldsymbol{s}_i \ .$$

Figure 1: Metaphoric depiction and game-theoretic description of the Rawlsian original position game. A weak Dæmon (left) plays against an all-seeing Angel (right).

A slightly different set of axioms yields the Gini class [Weymark, 1981, Gajdos and Weymark, 2005].

Definition 3.2 (Gini Welfare and Malfare)**.** Suppose a *ascending sequence* $\boldsymbol{w}^{\uparrow} \in \triangle_g$ or *descending sequence* of *Gini weights* $\boldsymbol{w}^{\downarrow} \in \triangle_g$, *risk vector* $\boldsymbol{s} \in \mathbb{R}^g$, and let $\boldsymbol{s}^{\downarrow}$ denote $\boldsymbol{s}$ in descending order. The generalized Gini social welfare function (GGSWF) is then

$$\mathrm{M}_{\boldsymbol{w}^{\uparrow}}(\boldsymbol{s}) \doteq \sum_{i=1}^{g} \boldsymbol{w}_i^{\uparrow} \boldsymbol{s}_i^{\downarrow} = \boldsymbol{w}^{\uparrow} \cdot \boldsymbol{s}^{\downarrow} \ , \tag{3}$$

and similarly, the Gini social malfare function is

$$\mathrm{M}_{\boldsymbol{w}^{\downarrow}}(\boldsymbol{s}) \doteq \sum_{i=1}^{g} \boldsymbol{w}_i^{\downarrow} \boldsymbol{s}_i^{\downarrow} = \boldsymbol{w}^{\downarrow} \cdot \boldsymbol{s}^{\downarrow} \ . \tag{4}$$

Notably, while the power-mean family is not closed under convex combination, the Gini family is. Furthermore, restricting to convex combinations of *utilitarian* and *egalitarian* welfare or malfare yields the *utilitarian-maximin* social welfare function (UMSWF) family. Several axiomatizations for the UMSWF class exist in the literature [Deschamps and Gevers, 1978, Bossert and Kamaga, 2020, Schneider and Kim, 2020], each essentially strengthening Gini axioms in some way.

4 A Philosophy of Robust Fair Objectives

Inspired by the original position argument and the veil of ignorance of Rawls [1971, 2001],[1] we pose a series of adversarial games, where a Dæmon is tasked with creating a world, and an Angel then punishes the Dæmon by choosing whom to reincarnate them as in their world. In many ways, this is an unfair game, as the Dæmon is given an impossibly difficult task, and the Angel lazily stands by until it is their turn to inflict maximal suffering upon their opponent, but perhaps it is an allegory for the responsibility of political leaders, and were they to face harsher rebuke from the citizenry, perhaps we would live in a more equitable society. We stress that this metaphor does not represent a conflict between good and evil, but rather a cosmic struggle between the freedom of the people (as represented by the Dæmon) and dictatorial power (as represented by the Angel).

[1] Note that Rawls' original position argument is generally phrased in terms of Wald's maximin principle and robustness to uncertainty, rather than explicitly as a zero-sum game. These characterizations are equivalent, and for our purposes, it is often convenient to characterize uncertainty as the action space of an explicit adversary. In general, this is to simplify intuition and the use of standard tools from game theory; it is not meant to be interpreted as a literal adversary in a literal game.

The Constrained Rawlsian Game

$g, \Theta, \boldsymbol{s}(\theta), \mathcal{S}$: Group count, Dæmon parameter space, sentiment function, sentiment space (as in figure 1).
$\mathcal{W} \subseteq \triangle_g$: Constrained weights space.
$\mathcal{A}_{\text{Dæ}} \doteq \mathcal{S}$: Dæmon action space.
$\mathcal{A}_{\text{Ang}} \doteq \mathcal{W}$: Angel action space
(*inhabitant* i becomes *distribution* $\boldsymbol{w}$).
$\mathrm{P}(\boldsymbol{s}; \boldsymbol{w}) : \mathbb{R}^g \times \triangle_g \to \mathbb{R}^2$: Zero-sum payoff function (expected sentiment).

$$\mathrm{P}_1(\boldsymbol{s}, \boldsymbol{w}) \doteq \langle \boldsymbol{w} \cdot \boldsymbol{s}, -\boldsymbol{w} \cdot \boldsymbol{s} \rangle$$

Strategic gameplay (Dæmon goes first):

$$\arg \underset{\boldsymbol{s} \in \mathcal{A}_{\text{Dæ}}}{\overset{\max}{\min}} \underset{\boldsymbol{w} \in \mathcal{A}_{\text{Ang}}}{\overset{\min}{\max}} \mathrm{P}_1(\boldsymbol{s}, i) = \arg \underset{\boldsymbol{s} \in \mathcal{S}}{\overset{\max}{\min}} \underset{\boldsymbol{w} \in \mathcal{W}}{\overset{\min}{\max}} \boldsymbol{w} \cdot \boldsymbol{s} \ .$$

Figure 2: Metaphoric depiction and game-theoretic description of the modified Rawlsian original position game, with restricted Angel action space. A Dæmon (left) plays against a comparably powerful Angel (right).

Because we wish to treat both utility and disutility (with welfare and malfare), we generically refer to these concepts as *sentiment*, and we adopt neutral notation $\boldsymbol{s}$ to represent sentiment vectors. In general, we use stacked operators, e.g., $\pm, \mp, \overset{\max}{\min}, \overset{\inf}{\sup}$, etc. to represent both cases simultaneously. Unless otherwise noted, the upper operator describes the utility branch and the lower operator is for the disutility branch.

Furthermore, to distinguish between the the utility values of the inhabitants of the game world $\boldsymbol{s}$, and those of the Dæmon and Angel playing the game, we refer to the latter as a payoff function $\mathrm{P}(\boldsymbol{s}; \boldsymbol{w})$, representing either positive payoff in the utility case or a negative payoff in the disutility case. In this adversarial zero-sum game, the Dæmon's payoff is the utility of the person they become, and the Angel's payoff is of course its negation.

The Angel's adversarial response to any Dæmon strategy is obvious: *Select the individual with the lowest utility (or highest disutility).* Playing against this Angel, the Dæmon must confront the question, *"How should we construct a world without knowing our place in it?"* Against an adversarial Angel, a strategic Dæmon must maximize the minimum utility (or minimize the maximum disutility). From this interaction, strategic gameplay results in the solution concept

$$\arg \underset{\boldsymbol{s} \in \mathcal{A}_{\text{Dæ}}}{\overset{\max}{\min}} \underset{i \in \mathcal{A}_{\text{Ang}}}{\overset{\min}{\max}} \mathrm{P}_1(\boldsymbol{s}, i) = \arg \underset{\boldsymbol{s} \in \mathcal{S}}{\overset{\max}{\min}} \underset{i \in 1, \ldots, g}{\overset{\min}{\max}} \boldsymbol{s}_i = \begin{cases} \text{Utility} & \max\limits_{\boldsymbol{s} \in \mathcal{S}} \min\limits_{1 \in 1, \ldots, g} \boldsymbol{s}_i = \operatorname{argmax}\limits_{\boldsymbol{s} \in \mathcal{S}} \mathrm{M}_{-\infty}(\boldsymbol{s}) \\ \text{Disutility} & \min\limits_{\boldsymbol{s} \in \mathcal{S}} \max\limits_{1 \in 1, \ldots, g} \boldsymbol{s}_i = \operatorname{argmin}\limits_{\boldsymbol{s} \in \mathcal{S}} \mathrm{M}_{\infty}(\boldsymbol{s}) \ . \end{cases}$$

The game is illustrated and further described in figure 1.

The power of the Dæmon in this game is directly modeled by the scope of worlds that they are capable of creating, and the Angel's power is directly impacted by the number of people that inhabit the world. We then consider variations of this game where the Angel is weakened in various ways, and show that they give rise to other standard notions of welfare, in the sense that the Dæmon's optimal strategy is to maximize some welfare concept. We also observe that in many cases, the original game is equivalent to one where the Angel must move first, but is allowed to employ a randomized strategy. We show that our modifications to the game can be formulated in several equivalent ways, and should not be taken too literally, as one motivation or another may be more or less suitable depending on the context of applications or philosophical stance of the reader.

4.1 On Mixed Strategies and Weak (Constrained) Adversaries

The power of the Dæmon in this game is directly controlled by the space $\mathcal{S}$ of feasible utility vectors (or some generating parameter space Θ), and allowing the Dæmon to play mixed strategies simply expands their action

space to the convex hull $\mathrm{CH}(\mathcal{S})$ (assuming the payoff function is defined as their *expected* utility). At times, we may require $\mathcal{S}$ to be a convex set, for which mixed Dæmon strategies are sufficient but not necessary. We explicitly assume as such when necessary, thus we make no further mention of mixed strategies on the Dæmon's part.

In our game, the Angel is allowed to condition their action on the Dæmon's *mixture action*, but not on the actual Dæmon action randomly selected from this mixture, thus when $\mathcal{S} \neq \mathrm{CH}(\mathcal{S})$, playing mixed strategies may increase the Dæmon's power in this game. However, because the Angel plays second, mixed strategies do not actually increase the Angel's power. In particular, the expected value of the payoff function given a mixed Dæmon strategy (distribution over (dis)utility vectors) $\mathcal{D} \in \mathcal{A}_{\mathrm{Dæ}}$ and an (independent) mixed Angel strategy (distribution over group indices, i.e., $\boldsymbol{w} \in \triangle_g$) is

$$\mathrm{P}_1(\mathcal{D}, \boldsymbol{w}) = \mathop{\mathbb{E}}_{\boldsymbol{s}\sim\mathcal{D}\perp i\sim\boldsymbol{w}}[\mathrm{P}_1(\boldsymbol{s}, i)] = \mathop{\mathbb{E}}_{\boldsymbol{s}\sim\mathcal{D}\perp i\sim\boldsymbol{w}}[\boldsymbol{s}_i] = \mathop{\mathbb{E}}_{\boldsymbol{s}\sim\mathcal{D}\perp i\sim\boldsymbol{w}}[\mathbb{1}_i \cdot \boldsymbol{s}] = \boldsymbol{w} \cdot \mathop{\mathbb{E}}_{\boldsymbol{s}\sim\mathcal{D}}\big[\boldsymbol{s}\big] \ , \tag{5}$$

where $A \perp B$ denotes *independence of random variables*. Strategic gameplay then results in the objective value (expected Dæmon payoff)

$$\sup_{\mathcal{D}\in\mathcal{A}_{\mathrm{Dæ}}} \inf_{\boldsymbol{w}\in\mathcal{A}_{\mathrm{Ang}}} \mathrm{P}_1(\mathcal{D}, \boldsymbol{w}) = \sup_{\mathcal{D}\in\mathcal{A}_{\mathrm{Dæ}}} \inf_{\boldsymbol{w}\in\mathcal{A}_{\mathrm{Ang}}} \boldsymbol{w} \cdot \mathop{\mathbb{E}}_{\boldsymbol{s}\sim\mathcal{D}}\big[\boldsymbol{s}\big] = \sup_{\boldsymbol{s}\in\mathrm{CH}(\boldsymbol{s})} \inf_{\boldsymbol{w}\in\mathcal{W}} \boldsymbol{w} \cdot \boldsymbol{s} \ . \tag{6}$$

Action Order Interchangeability It is worth noting that, while the order of play matters for specific strategies, under mild conditions, assuming strategic play, the order is interchangeable (and thus the game may be played with simultaneous actions). In particular, if we allow for mixed actions (or at least a convex action set for the Dæmon), we have the following result.

Lemma 4.1 (Maximin Interchangeability)**.** Suppose that both $\mathcal{A}_{\mathrm{Dæ}}$ and $\mathcal{A}_{\mathrm{Ang}}$ are convex sets and $\mathcal{A}_{\mathrm{Ang}}$ is closed (thus also compact). Then

$$\sup_{\mathcal{D}\in\mathcal{A}_{\mathrm{Dæ}}} \inf_{\boldsymbol{w}\in\mathcal{A}_{\mathrm{Ang}}} \mathrm{P}_1(\mathcal{D}, \boldsymbol{w}) = \inf_{\boldsymbol{w}\in\mathcal{A}_{\mathrm{Ang}}} \sup_{\mathcal{D}\in\mathcal{A}_{\mathrm{Dæ}}} \mathrm{P}_1(\mathcal{D}, \boldsymbol{w}) \ . \tag{7}$$

Proof. This result follows from (6), then application of Sion's [1958] minimax theorem. □

On Special Cases We now show that many of the most commonly employed welfare and malfare functions arise as special cases of this constrained-Angel Rawlsian game. This intuitively motivates these aggregator functions from the perspective of robustness, with no robustness representing utilitarianism, establishing a spectrum of various degrees and types of robustness, with egalitarianism at the opposite end of the spectrum (as in the power-mean and Gini classes). Depending on how natural the choice of constrained weight space is, it may lend credence to the use of particular aggregator functions, and this analysis is also relevant to their optimization or analysis, as they can be treated with standard maximinimization tools, discussed further in section 6. We also note that this robustness interpretation merely *complements* (but does not replace or invalidate) existing fairness interpretations of these aggregator functions, and in sections 4.2 and 4.3, we show other categories of fair and fair robust objective that don't seem to arise from just the robustness aspect of this particular game.

For the purposes of this characterization, we consider any fixed Dæmon strategy (or mixture of strategies), as represented by some (expected) (dis)utility vector $\boldsymbol{s}$. The following result characterizes the adversarial (worst case) payoff of the Dæmon against the Angel, i.e., we show that $\min_{\max_{\boldsymbol{w}\in\mathcal{W}}} \mathrm{P}_1(\boldsymbol{s}; \boldsymbol{w}) = \mathrm{M}(\boldsymbol{s})$ for some classical aggregator function $\mathrm{M}(\boldsymbol{s})$, thus the Dæmon's optimal strategy in this game is to optimize these aggregator functions. This generalizes the maximum principle of Rawlsian theory, wherein from behind the veil of ignorance, the social planner chooses to *maximize* the *minimum* utility, i.e., egalitarian welfare.

Theorem 4.2 (Classical Welfare and Malfare Functions as Constrained Angel Solution Concepts)**.** For any $\boldsymbol{s} \in \mathbb{R}^g$, the Angel's best responses under the following special cases of Angel action spaces $\mathcal{A}_{\mathrm{Ang}}$, as represented by weight action spaces $\mathcal{W}$, give rise to standard aggregator functions. Some cases assume fixed "true weights" $\boldsymbol{w}^* \in \triangle_g$ or Gini weights sequence $\boldsymbol{w}^{\downarrow} \in \triangle_g$ or $\boldsymbol{w}^{\uparrow} \in \triangle_g$, and obtain special cases in terms of these parameters.

1) **Egalitarian**: Suppose $\mathcal{W} = \triangle_g$. Then $\min_{\max_{\boldsymbol{w}\in\mathcal{W}}} \mathrm{P}_1(\boldsymbol{s}; \boldsymbol{w}) = \mathrm{M}_{\mp\infty}(\boldsymbol{s})$.

2) **Utilitarian**: Suppose $\mathcal{W} = \{\boldsymbol{w}^*\}$. Then $\min_{\max_{\boldsymbol{w}\in\mathcal{W}}} \mathrm{P}_1(\boldsymbol{s}; \boldsymbol{w}) = \mathrm{M}_1(\boldsymbol{s}; \boldsymbol{w}^*)$.

3) **Weighted Utilitarian-Maximin**: Suppose $\mathcal{W} = \{\boldsymbol{w} \,|\, \boldsymbol{w} \succeq \gamma\boldsymbol{w}^*\} = \gamma\{\boldsymbol{w}^*\} + (1-\gamma)\triangle_g$.
Then $\min_{\max_{\boldsymbol{w}\in\mathcal{W}}} \mathrm{P}_1(\boldsymbol{s}; \boldsymbol{w}) = \gamma\mathrm{M}_1(\boldsymbol{s}; \boldsymbol{w}^*) + (1-\gamma)\mathrm{M}_{\mp\infty}(\boldsymbol{s})$.

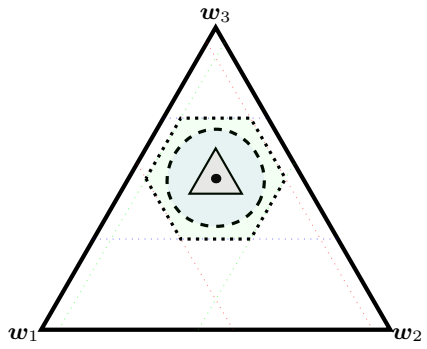

Figure 3: A simplicial plot over $\triangle_3$ of the robustness sets defined by intersection with the $\mathcal{L}_\infty$, $\mathcal{L}_2$, and $\mathcal{L}_1$ norm balls of radius $\frac{1}{5}$ around the point $\boldsymbol{w}^* = \langle \frac{1}{4}, \frac{1}{4}, \frac{1}{2} \rangle$. The boundaries of the $\mathcal{L}_\infty$, $\mathcal{L}_2$, and $\mathcal{L}_1$ balls are plotted in solid, dashed, and dotted lines, respectively. Assuming positive radius r such that each $\boldsymbol{w}^*$-centered norm ball is contained by the unit hypercube, i.e., $\|\boldsymbol{w}^*\|_\infty \leq \|\boldsymbol{w}^*\|_2 \leq \|\boldsymbol{w}^*\|_1 \leq r$, intersection with the unit simplex yields an equilateral triangular, circular, or hexagonal region, respectively, with $g = 3$. In higher dimensions, the regions become simplicial, hyperspherical, or regular-polytopal, respectively.

4) **Generalized Gini**: Suppose $\mathcal{W} = \{\pi(\boldsymbol{w}^\downarrow) | \pi \in \Pi_g\}$, where Π_g is the set of all permutations on g items. Then $\min_{\boldsymbol{w} \in \mathcal{W}} \mathrm{P}_1(\boldsymbol{s}; \boldsymbol{w}) = \mathrm{M}_{\boldsymbol{w}^\uparrow}(\boldsymbol{s})$ or $\max_{\boldsymbol{w} \in \mathcal{W}} \mathrm{P}_1(\boldsymbol{s}; \boldsymbol{w}) = \mathrm{M}_{\boldsymbol{w}^\downarrow}(\boldsymbol{s})$.
Furthermore, for each of the above items, the RHS follows for any Angel action space $\mathcal{W}'$, $\mathcal{W}$ and $\mathcal{W}'$ have the same convex hull, e.g., for item 1, we may use $\{\mathbb{1}_i \,|\, i \in 1, \ldots, g\}$ in place of $\triangle_g$. With this expansion, a sort of converse follows for each result. For each of the above items, if the conclusion holds *for all* $\boldsymbol{s} \in \mathbb{R}^g$, then $\mathcal{W}$ is some such $\mathcal{W}'$. For example, from item 1 we have $\min_{\boldsymbol{w} \in \mathcal{W}} \max \mathrm{P}_1(\boldsymbol{s}; \boldsymbol{w}) = \mathrm{M}_{\mp\infty}(\boldsymbol{s}) \implies \{\mathbb{1}_i \,|\, i \in 1, \ldots, g\} \subseteq \mathcal{W} \subseteq \triangle_g$.

Item 1 is rather obvious, as this special case is just the unconstrained game. Item 2 is also unsurprising, as this special case replaces the robust or worst case perspective of the Rawlsian game with an *average case* or expected perspective, which yields *weighted sums* of (dis)utility, i.e., utilitarian malfare or welfare. Of course, uniform individual weights $\boldsymbol{w} = \langle \frac{1}{g}, \frac{1}{g}, \ldots, \frac{1}{g} \rangle$ correspond to *uniformly randomly* selecting among all living individuals, which very much concords with the utilitarian perspective.[2] Despite the mathematical simplicity of these results, it is encouraging to see that the two most popular aggregator functions do arise as special cases of this adversarial game, and in some sense they are the extreme cases, as the Angel action space is *maximal* (complete) in item 1 and *minimal* (singleton) in item 2.

In contrast, item 3 is rather surprising, as we see that a simple lower-bound constraint on weights values produce the classical (weighted) utilitarian maximin social welfare function (UMSWF). The statement of the result gives some intuition: this Angel action space is a convex combination of the egalitarian and utilitarian action spaces, and so too is the weighted UMSWF a convex combination of the egalitarian and utilitarian aggregator functions. Also of note is that the unweighted UMSWF is a more restrictive class then the GGSWF, and is theoretically justified by a rather heavy-handed (strong) set of axioms. This result gives an alternative characterization of UMSWF as a robust variant of utilitarian welfare, where $\gamma \boldsymbol{w}_i^*$ is a lower bound on the weight of each group i (note that WLOG any such set of feasible lower-bounds can be represented for some $\gamma \in [0, 1]$, $\boldsymbol{w}^* \in \triangle_g$). Finally, item 4 is perhaps the most sophisticated result here, as the class of Angel actions has quite a bit more structure (though it is still a bounded polytope). This characterization also provides an alternative characterization of the Gini social welfare as a robust utilitarian objective, where the weights relative population sizes of all groups are known, but the identities of the group associated with each weight is not known.

We now show that expanding the classes of theorem 4.2 (except for egalitarian, which is already maximal) results in novel nontrivial robust aggregator function concepts. Each of these robust aggregators essentially optimizes a classical aggregator function subject to a worst case assumption w.r.t. some type of uncertainty. We illustrate in figure 3 robustness sets defined by the $\mathcal{L}_\infty$, $\mathcal{L}_2$, and $\mathcal{L}_1$ norms around a point $\boldsymbol{w}^*$.

Theorem 4.3 (Robust Welfare and Malfare Functions as Constrained Angel Solution Concepts)**.** Suppose as in theorem 4.3. Suppose also some closed convex *robustness set* $\mathcal{R}$ such that $\mathbf{0} \in \mathcal{R}$ (usually some type of norm-ball). Then for any $\boldsymbol{s} \in \mathbb{R}^g$, the Angel's best responses under the following special cases of Angel action spaces $\mathcal{A}_{\text{Ang}}$, as represented by weight action spaces $\mathcal{W}$, give rise to *robust variants* of standard aggregator functions.
1) **Utilitarian**: Suppose $\mathcal{W} = (\boldsymbol{w}^* + \mathcal{R}) \cap \triangle_g$. Then $\min_{\boldsymbol{w} \in \mathcal{W}} \max \mathrm{P}_1(\boldsymbol{s}; \boldsymbol{w}) = \min_{\boldsymbol{w}' \in \mathcal{W}} \mathrm{M}_1(\boldsymbol{s}; \boldsymbol{w}')$.
2) **Weighted Utilitarian-Maximin**: Suppose $\mathcal{W} = (\{\boldsymbol{w} \,|\, \boldsymbol{w} \succeq \gamma \boldsymbol{w}^*\} + \gamma \mathcal{R}) \cap \triangle_g = (\gamma \boldsymbol{w}^* + \mathcal{R}) \cap \triangle_g + (1 - \gamma) \triangle_g$. Then $\min_{\boldsymbol{w} \in \mathcal{W}} \max \mathrm{P}_1(\boldsymbol{s}; \boldsymbol{w}) = \min_{\boldsymbol{w}' \in (\boldsymbol{w}^* + \mathcal{R}) \cap \triangle_g} \max \gamma \mathrm{M}_1(\boldsymbol{s}; \boldsymbol{w}^*) + (1 - \gamma) \mathrm{M}_{\mp\infty}(\boldsymbol{s})$.

[2] Furthermore, assuming nonuniform *group weights* $\boldsymbol{w}$ correspond to the *population frequencies* of each group, this perspective still replaces the risk-aversion of Wald's maximin principle with a *uniform average* over all individuals.

The Altruistic Dæmon Rawlsian Game

$g, \mathcal{S}, \mathcal{W}$: Group count, sentiment space, weights space (as in figures 1 and 2).
$\mathcal{A}_{\text{Dæ}} \subseteq \mathbb{R}^g, \mathcal{A}_{\text{Ang}} \subseteq \triangle_g$: Action spaces.
$\mathrm{M}(\boldsymbol{s}; \boldsymbol{w})$: Dæmon aggregator function.
$\mathrm{P}(\boldsymbol{s}; \boldsymbol{w}) : \mathbb{R}^g \times \triangle_g \to \mathbb{R}^2$: Zero-sum payoff function representing the Dæmon's aggregate

$$\mathrm{P}(\boldsymbol{s}, \boldsymbol{w}) \doteq \langle \mathrm{M}(\boldsymbol{s}; \boldsymbol{w}), -\mathrm{M}(\boldsymbol{s}; \boldsymbol{w}) \rangle \ .$$

Strategic gameplay (Dæmon goes first):

$$\arg \min_{\boldsymbol{s} \in \mathcal{A}_{\text{Dæ}}}^{\max} \max_{\boldsymbol{w} \in \mathcal{A}_{\text{Ang}}}^{\min} \mathrm{P}_1(\boldsymbol{s}, \boldsymbol{w}) = \arg \min_{\boldsymbol{s} \in \mathcal{S}}^{\max} \max_{\boldsymbol{w} \in \mathcal{W}}^{\min} \mathrm{M}(\boldsymbol{s}; \boldsymbol{w}) \ .$$

Or with Utility Transforms

$T(u) : \mathbb{R} \to \mathbb{R}$: Dæmon sentiment transform.
$\mathrm{P}(\boldsymbol{s}; \boldsymbol{w}) \doteq \langle \boldsymbol{w} \cdot T(\boldsymbol{s}), -\boldsymbol{w} \cdot T(\boldsymbol{s}) \rangle$: Payoff function.
Strategic gameplay (Dæmon goes first):

$$\arg \min_{\boldsymbol{s} \in \mathcal{A}_{\text{Dæ}}}^{\max} \max_{\boldsymbol{w} \in \mathcal{A}_{\text{Ang}}}^{\min} \mathrm{P}_1(\boldsymbol{s}, \boldsymbol{w}) = \arg \min_{\boldsymbol{s} \in \mathcal{S}}^{\max} \max_{\boldsymbol{w} \in \mathcal{W}}^{\min} T^{-1}(\boldsymbol{w} \cdot T(\boldsymbol{s})) \ .$$

Figure 4: Metaphoric depiction and game-theoretic description of the altruistic Dæmon original position game. A social-planner Dæmon (left) plays a zero-sum game against an adversarial Angel (right). Both the aggregator-function and the utility-transform formulations of the game are presented.

3) **Generalized Gini**: Suppose $\mathcal{W} = (\{\pi(\boldsymbol{w}^{\downarrow}) | \pi \in \Pi_g\} + \mathcal{R}) \cap \triangle_g$, where Π_g is the set of all permutations on g items. Then $\min_{\boldsymbol{w} \in \mathcal{W}}^{\max} \mathrm{P}_1(\boldsymbol{s}; \boldsymbol{w}) = \min_{\boldsymbol{w}^{\downarrow\prime} \in (\boldsymbol{w}^{\downarrow} + \mathcal{R}) \cap \triangle_g}^{\max} \mathrm{M}_{\boldsymbol{w}^{\downarrow\prime}}(\boldsymbol{s})$.

Furthermore, in general none of the above are equivalent to *any* egalitarian, utilitarian, weighted utilitarian-maximin, or generalized Gini welfare or malfare function.

4.2 From Egocentric to Altruistic Agents

We now show that under our randomized game, if the Dæmon plays pure strategies and the mixed Angel strategy space is constrained to a compact set, then any power mean welfare function arises as a solution concept when the Dæmon's payoff is a concave utility transform (or convex disutility transform) of their *ex ante* (dis)utility. Alternatively, we can think of this as a Dæmon that is altruistically concerned with the wellbeing of groups of the people in their world, where the Angel is allowed to reweight the sizes of these groups. As a third interpretation, we can think of the game as a metaphysical construct where the Dæmon is not reincarnated once, but lives all lives within their world, and thus wants to ensure a just and equitable society.

In this game, the Dæmon's (dis)utility transform, or their welfare or malfare function, determine the power-mean p, and the weights $\boldsymbol{w}$ and robustness are determined by the Angel's action space $\mathcal{W}$. Finally, we develop a novel class of aggregator functions that combines the power-mean and the Gini classes, and show that it arises as the solution concept for particular parameterizations of this game. These games are depicted and described in figure 4.

Theorem 4.4 (Strategic Gameplay from Nonlinear Objectives)**.** Suppose payoff function $\mathrm{P}(\boldsymbol{s}; \boldsymbol{w}) = \langle \mathrm{M}_p(\boldsymbol{s}; \boldsymbol{w}), -\mathrm{M}_p(\boldsymbol{s}; \boldsymbol{w}) \rangle$. Then strategic gameplay yields

$$\arg \min_{\boldsymbol{s} \in \mathcal{A}_{\text{Dæ}}}^{\max} \max_{\boldsymbol{w} \in \mathcal{A}_{\text{Ang}}}^{\min} \mathrm{P}_1(\boldsymbol{s}, \boldsymbol{w}) = \arg \min_{\boldsymbol{s} \in \mathcal{S}}^{\max} \max_{\boldsymbol{w} \in \mathcal{W}}^{\min} \mathrm{M}_p(\boldsymbol{s}; \boldsymbol{w}) \ .$$

Furthermore, if $\mathcal{S} = \mathrm{CH}(\mathcal{S})$ and $\mathrm{M}(\boldsymbol{s}; \boldsymbol{w})$ exhibits concave curvature in $\boldsymbol{s}$ (or convex for disutility), then a pure Dæmon strategy is always optimal. Furthermore, if the curvature is *strictly concave*, then a pure Dæmon strategy is *strictly optimal* (over all other pure and mixed Dæmon strategies).

Theorem 4.5 (Power-Means as Utility Transforms)**.** Suppose some $p \leq 1$ for utility or $p \geq 1$ for disutility, and a (dis)utility transform $T(u) = \operatorname{sgn}(p)u^p$ for $p \neq 0$, or $T(u) = \ln(u)$ for $p = 0$, and take $\mathrm{P}_1(\boldsymbol{s}; \boldsymbol{w}) = \boldsymbol{w} \cdot T(\boldsymbol{s})$. Then

$$\min_{\substack{\max \\ \boldsymbol{w} \in \mathcal{A}_{\mathrm{Ang}}}} \mathrm{P}_1(\boldsymbol{s}, \boldsymbol{w}) = \min_{\substack{\max \\ \boldsymbol{w} \in \mathcal{W}}} \begin{cases} p \neq 0 & \operatorname{sgn}(p)\mathrm{M}_p^p(\boldsymbol{s}; \boldsymbol{w}) \\ p = 0 & \exp\big(\mathrm{M}_0(\boldsymbol{s}; \boldsymbol{w})\big) \end{cases} .$$

Consequently, as both of the above cases are strict monotonic functions of the power-mean $\mathrm{M}_p(\boldsymbol{s}; \boldsymbol{w})$, it holds that

$$\arg\max_{\boldsymbol{s} \in \mathcal{A}_{\mathrm{Dæ}}} \min_{\substack{\max \\ \boldsymbol{w} \in \mathcal{A}_{\mathrm{Ang}}}} \mathrm{P}_1(\boldsymbol{s}, \boldsymbol{w}) = \arg\max_{\boldsymbol{s} \in \mathcal{S}} \min_{\substack{\max \\ \boldsymbol{w} \in \mathcal{W}}} \mathrm{M}_p(\boldsymbol{s}; \boldsymbol{w}) .$$

Furthermore, if $\mathcal{S} = \mathrm{CH}(\mathcal{S})$ and $T(\cdot)$ exhibits concave curvature in $\boldsymbol{s}$ (or convex for disutility), then a pure Dæmon strategy is always optimal. Furthermore, if the curvature is *strictly concave*, then a pure Dæmon strategy is *strictly optimal* (over all other pure and mixed Dæmon strategies).

The $\boldsymbol{w}^\uparrow$-p Gini Power-Mean Class A natural instinct when confronted with the Gini and power-mean classes is to "combine them" into something like the following.

Definition 4.6 (The $\boldsymbol{w}^\uparrow$-p Gini Power-Mean Class)**.** Suppose some $p \leq 1$ and decreasing weights sequence $\boldsymbol{w}^\downarrow \in \triangle_g$ for utility, or some $p \geq 1$ and increasing weights sequence $\boldsymbol{w}^\downarrow \in \triangle_g$ for disutility. Then, letting $\boldsymbol{s}^\uparrow$ denote some $\boldsymbol{s} \in \mathbb{R}^g_{0+}$ in ascending order, we define

$$\mathrm{M}_{\boldsymbol{w}^\downarrow, p}(\boldsymbol{s}) \doteq \mathrm{M}_p(\boldsymbol{s}^\uparrow; \boldsymbol{w}^\downarrow) = \sqrt[p]{\sum_{i=1}^{g} \boldsymbol{w}_i^\downarrow (\boldsymbol{s}_i^\uparrow)^p} \text{ for welfare } (p \leq 1) , \quad \text{or}$$

$$\mathrm{M}_{\boldsymbol{w}^\uparrow, p}(\boldsymbol{s}) \doteq \mathrm{M}_p(\boldsymbol{s}^\uparrow; \boldsymbol{w}^\uparrow) = \sqrt[p]{\sum_{i=1}^{g} \boldsymbol{w}_i^\uparrow (\boldsymbol{s}_i^\uparrow)^p} \text{ for malfare } (p \geq 1) ,$$

i.e., we sort (dis)utilities, assign weights in ascending or descending order, and take a weighted power-mean.

This class clearly generalizes both the unweighted power-mean and Gini families (for $p = 1$ and $\boldsymbol{w}^\uparrow = \frac{1}{g}\mathbf{1}$ or $\boldsymbol{w}^\downarrow = \frac{1}{g}\mathbf{1}$), but now combines the piecewise-differentiable nature and ordinal boundaries of the Gini family with the continuously-differentiable nonlinear nature of the power-mean family.

Unfortunately, there is no known axiomatic characterization of definition 4.6, and although the Gini axioms and power-mean axioms overlap heavily, combining them yields only their *intersection*, i.e., the family consisting only of utilitarian and egalitarian welfare or malfare. However, from either of the above power-mean characterizations (theorems 4.4 and 4.5), with the appropriate constrained Angel (selected as in theorem 4.2 item 4), definition 4.6 arises as a solution concept to our game.

4.3 Coercing Altruistic Play from Egocentric Dæmons

We now show that an altruistic Angel can coerce altruistic play from an egocentric Dæmon. Metaphorically, this game is a bit more abstract than those previously discussed, as the Dæmon still serves as the social planner, but the solution concept we seek optimizes the Angel's aggregator function.

There is an interesting parallel to representative government here, where the populace (Angel) elects leaders (Dæmon) that perform social planning, but the voting process itself creates incentives for the leaders, though we don't see a direct technical connection to our results. Similarly, we wonder whether altruistic behavior on behalf of corporations, such as highly-visible campaigns of corporate "greenwashing" or "rainbow capitalism," may arise from similar interactions with consumers.

Suppose the Angel has a weighted power-mean aggregator function $\mathrm{M}_p(\boldsymbol{s}; \boldsymbol{w}^*)$. We construct a payoff function to model the self-centered Dæmon and altruistic Angel, obtaining

$$\mathrm{P}(\boldsymbol{s}; \boldsymbol{w}) : \mathbb{R}^g \times \triangle_g \to \mathbb{R}^2 \doteq \langle \boldsymbol{w} \cdot \boldsymbol{s}, \mathrm{M}(\boldsymbol{s}; \boldsymbol{w}^*) \rangle . \tag{8}$$

In this game, the Angel's action space does not represent robustness, but is rather used to influence the actions of the Dæmon, so we take $\mathcal{W} = \triangle_g$. Until now, we have considered turn-based games, but to analyze Nash equilibria, we must convert the game to normal form. Here the Dæmon and Angel act simultaneously, but to preserve the

The Altruistic Angel Rawlsian Game

$g, \mathcal{S}, \mathcal{W}$: Group count, sentiment space, weights space (as in figures 1 and 2).
$\mathcal{A}_{\text{Dæ}} \subseteq \mathbb{R}^g, \mathcal{A}_{\text{Ang}} = \triangle_g$: Action spaces.
$\mathrm{M}(\boldsymbol{s}; \boldsymbol{w})$: Angel aggregator function.
$\mathrm{P}(\boldsymbol{s}; \boldsymbol{w}) : \mathbb{R}^g \times \triangle_g \to \mathbb{R}^2$: Payoff function (Dæmon is self-interested, Angel is altruistic)

$$\mathrm{P}(\boldsymbol{s}, \boldsymbol{w}) \doteq \langle \boldsymbol{w} \cdot \boldsymbol{s}, \mathrm{M}(\boldsymbol{s}; \boldsymbol{w}^*) \rangle \ .$$

Angel action *does not impact* Angel payoff: Any strategy is a "best response."
Angel has NFG strategies for which the Dæmon's best response is to select $\arg\max_{\boldsymbol{s} \in \mathcal{S}} \mathrm{M}_p(\boldsymbol{s}; \boldsymbol{w}^*)$.
Neither Dæmon nor Angel has incentive to deviate: This is a Nash equilibrium.
Higher Angel utility is *not possible*, thus this Nash equilibrium is optimal (from Angel's perspective).

Figure 5: Metaphoric depiction and game-theoretic description of the altruistic Angel original position game.A self-interested Dæmon (left) is coerced into altruistic play by a social-planner Angel (right).

original turn-based game dynamics, the Angel's strategy is conditional on the Dæmon's action. In other words, in NFG form, the Angel's strategy space becomes $\mathcal{S} \to \triangle_g$, and an Angel strategy $S_{\text{Ang}}(\cdot) : \mathcal{S} \to \triangle_g$ given any Dæmon action $\boldsymbol{s} \in \mathcal{S}$ is to play $S_{\text{Ang}}(\boldsymbol{s})$.

In this game, the Angel action *does not impact* the Angel's payoff, thus any strategy is a "best response." Consequently, the Dæmon's best response to any Angel strategy is a Nash equilibrium. The Angel may seem powerless here, but we now show that for a particular choice of Angel strategy, we obtain a Nash equilibrium in which the Angel receives the greatest possible payoff.

Theorem 4.7 (Strategic Gameplay in Altruistic Angel Games)**.** Suppose the payoff function of (8) for some $p > 0$. If the Angel adopts the strategy $S_{\text{Ang}}(\boldsymbol{s}) = \boldsymbol{w}_i \propto \boldsymbol{w}_i^* \boldsymbol{s}_i^{p-1}$, then the Dæmon's best response is to select

$$\arg\max_{\boldsymbol{s} \in \mathcal{S}} \mathrm{P}_1\left(\boldsymbol{s}; S_{\text{Ang}}(\boldsymbol{s})\right) = \arg\max_{\boldsymbol{s} \in \mathcal{S}} S_{\text{Ang}}(\boldsymbol{s}) \cdot \boldsymbol{s} = \arg\max_{\boldsymbol{s} \in \mathcal{S}} \boldsymbol{w}^* \cdot \boldsymbol{s}^p = \arg\max_{\boldsymbol{s} \in \mathcal{S}} \mathrm{M}_p(\boldsymbol{s}; \boldsymbol{w}^*) \ .$$

Similarly, for $p = 0$, suppose the Dæmon is limited to utility values at least $\boldsymbol{s}_{\min} > 0$, i.e., $\boldsymbol{s} \succeq \mathbf{1}\boldsymbol{s}_{\min}$. If the Angel adopts the strategy $S_{\text{Ang}}(\boldsymbol{s}) = \boldsymbol{w}_i \propto \boldsymbol{w}_i^* \frac{\ln(\boldsymbol{s}_i / \boldsymbol{s}_{\min})}{\boldsymbol{s}_i / \boldsymbol{s}_{\min}}$, then the Dæmon's best response is to select

$$\operatorname*{argmax}_{\boldsymbol{s} \in \mathcal{S}} \mathrm{P}_1\left(\boldsymbol{s}; S_{\text{Ang}}(\boldsymbol{s})\right) = \operatorname*{argmax}_{\boldsymbol{s} \in \mathcal{S}} S_{\text{Ang}}(\boldsymbol{s}) \cdot \boldsymbol{s} = \operatorname*{argmax}_{\boldsymbol{s} \in \mathcal{S}} \boldsymbol{w}^* \cdot \ln \boldsymbol{s} = \operatorname*{argmax}_{\boldsymbol{s} \in \mathcal{S}} \mathrm{M}_0(\boldsymbol{s}; \boldsymbol{w}^*) \ .$$

Finally, for $p < 0$, again suppose the Dæmon is limited to $\boldsymbol{s} \succeq \mathbf{1}\boldsymbol{s}_{\min}$. If the Angel adopts the strategy $S_{\text{Ang}}(\boldsymbol{s}) = \boldsymbol{w}_i \propto \boldsymbol{w}_i^* \left(\frac{\boldsymbol{s}_{\min}}{\boldsymbol{s}_i} - (\frac{\boldsymbol{s}_{\min}}{\boldsymbol{s}_i})^{1-p} \right)$, then the Dæmon's best response is to select

$$\operatorname*{argmax}_{\boldsymbol{s} \in \mathcal{S}} \mathrm{P}_1\left(\boldsymbol{s}; S_{\text{Ang}}(\boldsymbol{s})\right) = \operatorname*{argmax}_{\boldsymbol{s} \in \mathcal{S}} S_{\text{Ang}}(\boldsymbol{s}) \cdot \boldsymbol{s} = \operatorname*{argmax}_{\boldsymbol{s} \in \mathcal{S}} 1 - \boldsymbol{w}^* \cdot \boldsymbol{s}^p = \operatorname*{argmax}_{\boldsymbol{s} \in \mathcal{S}} \mathrm{M}_p(\boldsymbol{s}; \boldsymbol{w}^*) \ .$$

Furthermore, in each case, this is a Nash equilibrium, and no strategy profile yields higher Angel payoff (or lower payoff for disutility).

This result tells us that power-mean fairness concepts can arise even from a straightforward linear-utility egocentric Dæmon. This is surprising, as the results of section 4.2 need imbue the Dæmon with a payoff function that already closely the power-mean in some way, but theorem 4.7 shows that we can instead modify the Angel's payoff in a sequential Rawlsian game. While no longer a simple zero-sum normal form game, we still obtain a Nash equilibrium in which the power-mean arises as the Dæmon's robust solution concept.

5 Mathematical Properties of Robust Fair Objectives

We now argue that the objectives of section 4 also arise naturally as robust proxies for unknown information about the relative weights of groups (Angel actions). In particular, we show theoretical guarantees for the optimization of said robust proxies. Section 5.1 leads with a utilitarian perspective, and section 5.2 generalizes this analysis to a broader class of prioritarian objectives.

We consider two philosophical perspectives on the nature of uncertainty. First, we assume there exists some ground truth weights $\boldsymbol{w}^*$, but due to epistemic uncertainty about these weights, we only have knowledge of some *feasible set* of weights $\mathcal{W}$ in which the true weights $\boldsymbol{w}^*$ are known to be contained. In the second setting, we do not assume that there exists a single ground truth set of weights, and instead argue that our guarantees hold for any weights vector $\boldsymbol{w} \in \mathcal{W}$. The second model feels rather abstract, but is actually quite useful in machine learning contexts: A model may be trained and then deployed in multiple regions with varying demographics, or demographics may change over time in a single region. Our robust objectives then yield model guarantees that hold so long as demographic shift does not take group frequencies outside of the feasible weight space $\mathcal{W}$.

5.1 A Utilitarian Perspective

Note that, by nature, if $\boldsymbol{w}^* \in \mathcal{W}$, it holds that

$$\inf_{\boldsymbol{w}\in\mathcal{W}} \boldsymbol{w}\cdot\boldsymbol{s} \leq \boldsymbol{w}^*\cdot\boldsymbol{s} \leq \sup_{\boldsymbol{w}\in\mathcal{W}} \boldsymbol{w}\cdot\boldsymbol{s} \ . \tag{9}$$

Moreover, this is in some sense the optimal such lower-bound, which holds over adversarial choice of $\boldsymbol{w}^*$.

Consequently, optimizing $\arg\max_{\theta\in\Theta} \min_{\boldsymbol{w}\in\mathcal{W}} \boldsymbol{w}\cdot\boldsymbol{s}(\theta)$ is a safe proxy for optimizing $\arg\max_{\theta\in\Theta} \boldsymbol{w}^*\cdot\boldsymbol{s}(\theta)$, and the gap between the robust proxy objective and the true objective value can be bounded as

$$\left|\max_{\theta\in\Theta} \boldsymbol{w}^*\cdot\boldsymbol{s}(\theta) - \max_{\theta\in\Theta}\min_{\boldsymbol{w}\in\mathcal{W}} \boldsymbol{w}\cdot\boldsymbol{s}(\theta)\right| \leq \left|\max_{\theta\in\Theta}\max_{\boldsymbol{w}\in\mathcal{W}} \boldsymbol{w}\cdot\boldsymbol{s}(\theta) - \max_{\theta\in\Theta}\min_{\boldsymbol{w}\in\mathcal{W}} \boldsymbol{w}\cdot\boldsymbol{s}(\theta)\right| \leq \mathrm{Range}(\boldsymbol{s})\,\mathrm{Diam}_1(\mathcal{W}) \ , \tag{10}$$

where $\mathrm{Range}(\boldsymbol{s})$ is the sentiment range, and $\mathrm{Diam}_1(\mathcal{W})$ is the $\mathcal{L}_1$ diameter of the feasible weights space. Thus while the Rawlsian game gives us an elegant theoretical model of robust fair objectives, in a practical sense they are also relevant as objectives for operating fairly under adversarial uncertainty, and the magnitude of uncertainty, as measured by $\mathrm{Diam}_1(\mathcal{W})$, characterizes the cost of operating under uncertainty via (10).

5.2 A Generalized Prioritarian Perspective

The analysis of section 5.1 considers *robustness* in the sense of adversarial uncertainty over the weights $\boldsymbol{w}$, but not *fairness*, in the sense of nonlinear aggregator functions that incentivize equitable redistribution of (dis)utility. We now show nonlinear variants of the above results for objectives that can be decomposed as

$$\min_{\boldsymbol{w}\in\mathcal{W}} \mathrm{M}(\boldsymbol{s};\boldsymbol{w}) \ .$$

Note that here $\mathcal{W}$ is not necessarily the Angel's action space, but rather it is only the robustness parameters that are not incorporated into the objective function itself (see theorem 4.3). In terms of usage, we may generally assume that a fair objective $\mathrm{M}(\cdot;\boldsymbol{w})$ is known (selected) in advance, and epistemic uncertainty about weights is also known and provided as $\mathcal{W}$. We thus have

$$\inf_{\boldsymbol{w}\in\mathcal{W}} \mathrm{M}(\boldsymbol{s};\boldsymbol{w}) \leq \mathrm{M}(\boldsymbol{s};\boldsymbol{w}^*) \leq \sup_{\boldsymbol{w}\in\mathcal{W}} \mathrm{M}(\boldsymbol{s};\boldsymbol{w}) \ . \tag{11}$$

6 Adversarial Optimization of Robust Fair Objectives

This work centers the motivation for and properties of robust fair objectives, but we now briefly discuss their optimization. We then discuss modeling and applications in fair allocation and machine learning.

In this section, we assume an objective of the form

$$\arg\max_{\boldsymbol{s}\in\mathcal{S}} \min_{\boldsymbol{w}\in\mathcal{W}} \mathrm{M}_p(\boldsymbol{s};\boldsymbol{w}) \ ,$$

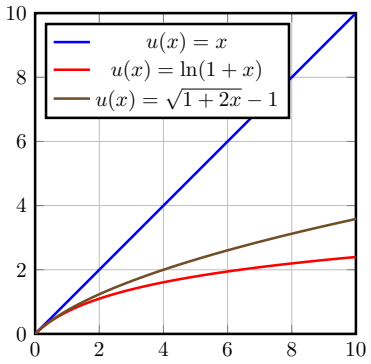

Figure 6: Plots of various nonlinear utility transform functions referenced in the text, as a compared to linear utility $u(x) = x$. The logarithmic transform $u(x) = \ln(1+x)$ and square root transform $u(x) = \sqrt{1+2x} - 1$ on utility values are shown. Note that both are smooth strictly-increasing strictly concave utility transforms that are tangent to the linear utility $u(x) = x$ at $x = 0$, thus they obey $0 \leq u(x) \leq x$, $\lim_{x\to\infty} u(x) = \infty$, $\lim_{x\to\infty} \frac{u(x)}{x} = 0$, and $\lim_{x\to 0^+} \frac{u(x)}{x} = 1$, i.e., they lie strictly below linear utility, and behave asymptotically as sublinear but superconstant (unbounded).

where either $\mathrm{M}(\boldsymbol{s}; \boldsymbol{w})$ exhibits concavity in $\boldsymbol{s}$ and convexity and $\boldsymbol{w}$ for outer maximization, or convexity in $\boldsymbol{s}$ and concavity in $\boldsymbol{w}$ for outer minimization. It is well known from convex optimization theory that we can efficiently maximize concave functions or minimize convex functions. Moreover, inner maximization preserves outer concavity and inner minimization preserves outer convexity. Thus the cards seem to be in our favor, and adversarial optimization exhibiting this concave-convex maximization convex-concave minimization structure is generally tractable [Nemirovski, 2004, Lin et al., 2020], e.g., via a variety of gradient ascent-descent methods.

Lemma 6.1 (Power-Mean Curvature)**.** Power-mean welfare and malfare functions exhibit the following curvature.
1) For any $p \geq 1$, $\mathrm{M}_p(\cdot; \boldsymbol{w})$ is convex, (strictly, but never strongly, for $p > 1$), and $\mathrm{M}_p(\boldsymbol{s}; \cdot)$ is concave (non-strictly).
2) For any $p \leq 1$, $\mathrm{M}_p(\boldsymbol{s}; \cdot)$ is concave, (non-strictly), and $\mathrm{M}_p(\boldsymbol{s}; \cdot)$ is convex (non-strictly).

In other words, the power-mean $\mathrm{M}_p(\boldsymbol{s}; \boldsymbol{w})$ exhibits opposite curvature in $\boldsymbol{s}$ and $\boldsymbol{w}$. We will first consider a few trivial cases, where lemma 6.1 suffices, as it is easy to convert the space of feasible allocation to the space of feasible utility values. In such settings, it is straightforward to apply standard maximin-optimization algorithms.

In general, it is not always so easy to convert between the parameter space space of feasible allocations Θ and the space of feasible utilities. In these more general settings, we need to consider the optimization problem directly as a function of the parameters space Θ. We thus require another technical lemma.

Lemma 6.2 (Power-Mean Composition Curvature)**.** Suppose some per-group (dis)utility function $\boldsymbol{s} : \Theta \to \mathbb{R}^g_{0+}$. Compositions $\mathrm{M}_p(\boldsymbol{s}(\theta); \boldsymbol{w}) : (\Theta \times \triangle_g) \to \mathbb{R}_{0+}$ of power-means with $\boldsymbol{s}$ exhibit the following curvature.
1) For any $p \geq 1$, if $\boldsymbol{s} : \Theta \to \mathbb{R}^g$ is convex, then $\mathrm{M}_p(\boldsymbol{s}(\theta); \boldsymbol{w}) : (\Theta \times \triangle_g) \to \mathbb{R}_{0+}$ is convex in θ and concave in $\boldsymbol{w}$.
2) For any $p \leq 1$, if $\boldsymbol{s} : \Theta \to \mathbb{R}^g$ is concave, then $\mathrm{M}_p(\boldsymbol{s}(\theta); \boldsymbol{w}) : (\Theta \times \triangle_g) \to \mathbb{R}_{0+}$ is concave in θ and convex in $\boldsymbol{w}$.

6.1 Simple Applications in Fair Allocation Problems

We assume here that g agents are being allocated k divisible goods. Each good i has *capacity* $\boldsymbol{c}_i$, thus allocations are *matrices* $\theta \in \mathbb{R}^{g \times k}_{0+}$, where each column (good allocation) i sum is bounded by $\boldsymbol{c}_i$. We let Θ denote the set of *feasible allocations*, and $\boldsymbol{s}(\theta) \in \mathbb{R}^g_{0+}$ is the *utility vector* given some feasible allocation $\theta \in \Theta$. We thus have

$$\theta = \begin{array}{c} \\ \text{Agent 1} \\ \\ \text{Agent } g \end{array} \begin{array}{c} \overset{\text{Good 1}}{} \quad \cdots \quad \overset{\text{Good } k}{} \\ \begin{pmatrix} \theta_{1,1} & \cdots & \theta_{1,k} \\ \vdots & \ddots & \vdots \\ \theta_{g,1} & \cdots & \theta_{g,k} \end{pmatrix} \end{array}.$$

We now study the optimization problems that result from applying robust fair objectives to such tasks.

Example 6.3 (Single Good with Linear Utility)**.** Suppose we have $k = 1$ goods with capacity c, and linear utility

$\boldsymbol{p}_i$ for agent i in their share, i.e., $\boldsymbol{s}_i(\theta) = \theta_{i,1}\boldsymbol{p}_i$. Then

$$\begin{aligned}\max_{\theta\in\Theta}\min_{\boldsymbol{w}\in\mathcal{W}} \mathrm{M}(\boldsymbol{s}(\theta);\boldsymbol{w}) &= \max_{\substack{\theta\in\mathbb{R}^g_{0+}:\\ \sum_{i=1}^g \theta_{i,1}\le c}} \min_{\boldsymbol{w}\in\mathcal{W}} \mathrm{M}(i\mapsto \boldsymbol{p}_i\theta_{i,1};\boldsymbol{w}) \\ &= \max_{\substack{\boldsymbol{s}\in\mathbb{R}^g_{0+}:\\ \sum_{i=1}^g \frac{1}{\boldsymbol{p}_i}\theta_{i,1}\le c}} \min_{\boldsymbol{w}\in\mathcal{W}} \mathrm{M}(\boldsymbol{s};\boldsymbol{w}) \ .\end{aligned}$$

We may approximately solve the convex-concave optimization problem of the RHS via standard maximin programming techniques, and given a maximal $\boldsymbol{s}$ in the RHS objective, we can trivially identify some $\theta \in \Theta$ that gives rise to it in the LHS via linear programming, or in closed form by inversion to get $\theta_{i,1} = \frac{\boldsymbol{s}_i(\theta)}{\boldsymbol{p}_i}$.

Example 6.4 (Single Good with Nonlinear Utility)**.** Suppose *nonlinear utility* $\boldsymbol{s}_i(\theta) = (\sqrt{1+2\theta_{i,1}} - 1)\boldsymbol{p}_i$ (see figure 6). With this very special choice, we can similarly pose

$$\begin{aligned}\max_{\theta\in\Theta}\min_{\boldsymbol{w}\in\mathcal{W}} \mathrm{M}(\boldsymbol{s}(\theta);\boldsymbol{w}) &= \max_{\substack{\theta\in\mathbb{R}^g_{0+}:\\ \sum_{i=1}^g \theta_{i,1}\le c}} \min_{\boldsymbol{w}\in\mathcal{W}} \mathrm{M}\left(i\mapsto (\sqrt{1+2\theta_{i,1}} - 1)\boldsymbol{p}_i;\boldsymbol{w}\right) \\ &= \max_{\substack{\boldsymbol{s}\in\mathbb{R}^g_{0+}:\\ \frac{1}{2}\sum_{i=1}^g (\frac{\boldsymbol{s}_i}{\boldsymbol{p}_i}+1)^2 - 1\le c}} \min_{\boldsymbol{w}\in\mathcal{W}} \mathrm{M}(\boldsymbol{s};\boldsymbol{w}) \ .\end{aligned}$$

Again we may approximately solve the convex-concave optimization problem of the RHS via standard maximin programming techniques. The feasible set of utility values in the RHS is the nonnegative portion of an axis-aligned ellipsoid, and converting some optimal $\boldsymbol{s}$ to the θ that gives rise to it is trivial via convex quadratic programming, or via inversion through the quadratic formula to get $\theta_{i,1} = \frac{1}{2}(\frac{\boldsymbol{s}_i}{\boldsymbol{p}_i}+1)^2 - \frac{1}{2} = \frac{\boldsymbol{s}_i}{\boldsymbol{p}_i} + \frac{\boldsymbol{s}_i^2}{2\boldsymbol{p}_i^2}$.

Example 6.5 (Multiple Goods with Linear Utility)**.** Now suppose k goods with *linear utility* $\boldsymbol{s}_i(\theta) = \theta_i \cdot \boldsymbol{P}_i = \sum_{j=1}^k \theta_{i,j}\boldsymbol{P}_{i,j}$. Suppose also arbitrary linear equality and inequality constraints on Θ, with $\Theta \neq \emptyset$. Via similar techniques, we can convert the space of feasible $\theta \in \Theta$ to some $\boldsymbol{s} \in \mathcal{S}$, where both Θ and $\mathcal{S}$ are polytopes. We can then optimize over $\boldsymbol{s} \in \mathcal{S}$, and finally select some $\theta \in \Theta$ that gives rise to the optimal $\boldsymbol{s}$ via linear programming.

In each of these examples, we optimize a robust fair objective over utility values, and then invert to obtain an allocation $\theta \in \Theta$ that produces utility values that optimize the robust objective. In many applications, the feasible space of (dis)utility vectors $\mathcal{S}$ is not directly known. Instead, we now assume that we have some *parameter* $\theta \in \Theta$, which yields a utility vector $\boldsymbol{s}(\theta)$. Robust fair objectives of $\boldsymbol{s}(\theta)$ can then be optimized. This is directly relevant to many fair ML applications, but we note now that fair allocation of divisible goods (or chores) with nonlinear utility can also be handled in this way, so long as utilities are *concave* in θ and disutilities are *convex* in θ.

For example, suppose store (agent) i will sell up to $\boldsymbol{C}_{i,j}$ units of items j for $\boldsymbol{P}_{i,j}$\$/unit of profit, and the *utility* derived by the store is given by the logarithmic utility transform $\ln(1 + \theta_i \cdot \boldsymbol{P}_i)$ (see figure 6). Then, for g stores and k items, the utility vector of the allocation $\theta \in \mathbb{R}^{g\times k}_{0+}$ can be expressed as

$$\boldsymbol{s}_i(\theta) = \ln\left(1 + \sum_{i=1}^k \boldsymbol{P}_{i,j}\min(\boldsymbol{C}_{i,j}, \theta_{i,j})\right) \ .$$

The final task is then to compute

$$\operatorname*{argmax}_{\theta\in\Theta}\min_{\boldsymbol{w}\in\mathcal{W}} \mathrm{M}\big(\boldsymbol{s}(\theta);\boldsymbol{w}\big) \ ,$$

which by lemma 6.1 is tractable for power-mean welfare objectives.

This idea immediately extends to ML settings, where the (empirical) loss or utility derived by each agent or group is also a complicated function of some parameter $\theta \in \Theta$. The optimization of such objectives is straightforward via standard first-order methods, e.g., subgradient descent for Lipschitz continuous malfare functions as proposed by Cousins [2021a]. Section 7 shows that robust power-mean malfare functions are indeed Lipschitz continuous, and then derives generalization bounds for fair learning with such robust objectives.

7 Statistical Generalization Bounds for Robust Fair Objectives

We now show that our robust objectives preserve the Lipschitz or Hölder continuity of their underlying non-robust counterparts. In particular, robust power-mean malfare functions are Lipschitz continuous, and robust power-mean welfare functions are almost always Lipschitz or Hölder continuous. We then translate these results into generalization bounds for fair learning with such robust objectives.

It is known [Cousins, 2022] that for any monotonic aggregator function $\mathrm{M}(\cdot)$, if the gap between true and empirical risk or utility values $\boldsymbol{s}_i$ and $\hat{\boldsymbol{s}}_i$ of each group i is no more than $\boldsymbol{\varepsilon}_i$, i.e., if we have $|\boldsymbol{s}_i - \hat{\boldsymbol{s}}_i| \leq \boldsymbol{\varepsilon}_i$, then we may bound the generalization error as

$$\mathrm{M}(\hat{\boldsymbol{s}} - \boldsymbol{\varepsilon}) \leq \mathrm{M}(\boldsymbol{s}) \leq \mathrm{M}(\hat{\boldsymbol{s}} - \boldsymbol{\varepsilon}) \ . \tag{12}$$

Bounding the estimation error $\boldsymbol{\varepsilon}_i$ for each group is a nontrivial matter, but standard techniques in statistical learning theory suffice. Cousins [2023] uses Rademacher averages and other statistical methods to bound the supremum deviation over each group, thus deriving values for $\boldsymbol{\varepsilon}$, and Cousins et al. [2024] show that sharper generalization bounds can be achieved by explicitly considering the effect of fair training on generalization error, and in particular that from the perspective of each group i, the fair model is effectively learned over a class that is biased towards strong performance on the remaining groups.

From (12), Lipschitz or Hölder continuity properties of the aggregator function yield loose but algebraically convenient bounds, as well as bounds on the sample complexity of PAC-learning such objectives. In particular, a function $\mathrm{M}(\cdot)$ is λ-α-$\|\cdot\|_{\mathrm{M}}$ Hölder continuous w.r.t. some norm $\|\cdot\|_{\mathrm{M}}$ if for all $\boldsymbol{s}, \boldsymbol{s}'$ in its domain, it holds

$$|\mathrm{M}(\boldsymbol{s}) - \mathrm{M}(\boldsymbol{s}')| \leq \lambda \|\boldsymbol{s} - \boldsymbol{s}'\|_{\mathrm{M}}^{\alpha} \ . \tag{13}$$

Moreover, if $\alpha = 1$, then $\mathrm{M}(\cdot)$ is λ-$\|\cdot\|_{\mathrm{M}}$ Lipschitz continuous. In concert with equation (12), under Hölder continuity, we may thus bound generalization error as

$$|\mathrm{M}(\hat{\boldsymbol{s}}) - \mathrm{M}(\boldsymbol{s})| \leq \lambda \|\boldsymbol{\varepsilon}\|_{\mathrm{M}}^{\alpha} \ , \tag{14}$$

and plugging in a specific expression for per-group generalization error $\boldsymbol{\varepsilon}$ allows us to solve for sample complexity (sufficient sample size) bounds. Using these results, we need only show that fair robust objectives exhibit similar Lipschitz and Hölder continuity properties to their non-robust counterparts.

7.1 Lipschitz and Hölder Continuity of Robust Fair Objectives

To streamline the analysis of our robust fair objectives, we introduce the notation

$$\mathrm{M}(\boldsymbol{s}; \mathcal{W}) \doteq \sup_{\boldsymbol{w} \in \mathcal{W}} \inf \mathrm{M}(\boldsymbol{s}; \boldsymbol{w}) \ .$$

We now show that these objectives have similar continuity properties to their non-robust counterparts, in line with lemmata 3.12 and 3.13 of Cousins [2023].

Lemma 7.1 (Hölder Continuity of Robust Fair Objectives)**.** Suppose an λ-α-$\|\cdot\|_{\mathrm{M}}$ Hölder continuous weighted aggregator function $\mathrm{M}(\boldsymbol{s}; \boldsymbol{w})$ over feasible weights space $\mathcal{W} \subseteq \triangle_g$. Then the robust aggregator function $\mathrm{M}(\boldsymbol{s}; \mathcal{W}) = \max_{\min \boldsymbol{w} \in \mathcal{W}} \mathrm{M}(\boldsymbol{s}; \boldsymbol{w})$ is λ-α-$\|\cdot\|_{\mathrm{M}}$ Hölder continuous, i.e., for all $\boldsymbol{s}, \boldsymbol{s}'$, it holds

$$|\mathrm{M}(\boldsymbol{s}; \mathcal{W}) - \mathrm{M}(\boldsymbol{s}'; \mathcal{W})| \leq \lambda \|\boldsymbol{s} - \boldsymbol{s}'\|_{\mathrm{M}}^{\alpha} \ .$$

Corollary 7.2 (Hölder Continuity of Robust Power-Means)**.** Suppose a robust power-mean operator $\mathrm{M}_p(\boldsymbol{s}; \mathcal{W})$, sentiment range r, and arbitrary $\boldsymbol{s}, \boldsymbol{s}' \in [0, r]^g$. Let $\boldsymbol{w}_{\min} \doteq \inf_{\boldsymbol{w} \in \mathcal{W}} \min_{i \in 1, \ldots, g} \boldsymbol{w}_i$, and $\boldsymbol{w}_{\max} \doteq \sup_{\boldsymbol{w} \in \mathcal{W}} \max_{i \in 1, \ldots, g} \boldsymbol{w}_i$. Then $\mathrm{M}_p(\boldsymbol{s}; \mathcal{W})$ exhibits the following Lipschitz and Hölder continuity properties.
1) For all $p \geq 1$: $\mathrm{M}_p(\boldsymbol{s}; \mathcal{W})$ is 1 Lipschitz continuous w.r.t. itself, i.e.,

$$|\mathrm{M}_p(\boldsymbol{s}; \mathcal{W}) - \mathrm{M}_p(\boldsymbol{s}'; \mathcal{W})| \leq \mathrm{M}_p(|\boldsymbol{s} - \boldsymbol{s}'|; \mathcal{W}) \leq \|\boldsymbol{s} - \boldsymbol{s}'\|_{\infty} \ .$$

Thus $\mathrm{M}_1(\boldsymbol{s}; \mathcal{W})$ is $\boldsymbol{w}_{\max}$-$\|\cdot\|_1$ Lipschitz continuous, and for $p = \infty$, it is 1-$\|\cdot\|_{\infty}$ Lispchitz continuous, and both of these constants are optimal.
2) For all $p < 0$, if $\boldsymbol{w}_{\min} > 0$, then $\mathrm{M}_p(\boldsymbol{s}; \mathcal{W})$ is $\frac{1}{\sqrt[|p|]{\boldsymbol{w}_{\min}}}$-$\|\cdot\|_{\infty}$ Lipschitz continuous.

3) For all $p \in (0,1)$, $\mathrm{M}_p(\boldsymbol{s};\mathcal{W})$ is $r^{1-p}\frac{1}{p}$-p-$\|\cdot\|_\infty$ Hölder continuous.
4) For all $p \leq 1$, if $\boldsymbol{w}_{\min} > 0$, then $\mathrm{M}_p(\boldsymbol{s};\mathcal{W})$ is $r^{1-\boldsymbol{w}_{\min}}$-$\boldsymbol{w}_{\min}$-$\|\cdot\|_\infty$ Hölder continuous.

Theorem 7.3 (A Codex of Sample Complexity)**.** Suppose that $\mathrm{M}(\boldsymbol{s})$ is λ-α-$\|\cdot\|_{\mathrm{M}}$ Hölder continuous, and for any *sample size* $m \geq m_0$ and *failure probability* $\delta \in (0,1)$, the empirical and true per-group risk values $\hat{\boldsymbol{s}}$ and $\boldsymbol{s}$ of the empirical malfare minimizer on m samples obey $\mathbb{P}\left(|\hat{\boldsymbol{s}}_i - \boldsymbol{s}_i| > \sqrt{\frac{\boldsymbol{v}_i \ln \frac{t}{\delta}}{m}}\right) < \delta$. Then the *sample complexity* of empirical malfare minimization (or welfare maximization) obeys

$$m^*(\varepsilon,\delta) \leq \max\left(m_0, \left\lceil \left(\frac{\lambda}{\varepsilon}\right)^{2/\alpha} \|i \mapsto \sqrt{\boldsymbol{v}_i}\|_{\mathrm{M}}^2 \ln\frac{tg}{\delta} \right\rceil\right) \ ,$$

i.e., for any sample size $m \geq m^*(\varepsilon,\delta)$, it holds that

$$\mathbb{P}\left(\mathrm{M}(\boldsymbol{s}) \leq \mathrm{M}(\boldsymbol{s}^*) + \varepsilon\right) \geq 1-\delta \ \text{ for malfare } , \quad \text{or } \ \mathbb{P}\left(\mathrm{M}(\boldsymbol{s}) \geq \mathrm{M}(\boldsymbol{s}^*) - \varepsilon\right) \geq 1-\delta \ \text{ for welfare } .$$

where $\boldsymbol{s}$ is the true risk (utility) vector of the learned model, and $\mathrm{M}(\boldsymbol{s}^*)$ is the infimum of true malfare (welfare) over the hypothesis class.

From theorem 7.3, it immediately follows that if each group's generalization error behaves as $\sqrt{\frac{v \ln \frac{t}{\delta}}{m}}$ (i.e., variance proxy v, tail count t; such bounds arise frequently via Rademacher averages and other statistical techniques [Cousins, 2022, Cousins et al., 2024]), the sample complexity of optimizing any robust power-mean malfare function is bounded by

$$m^*(\varepsilon,\delta) \leq \left\lceil \frac{v\lambda^2 \ln \frac{tg}{\delta}}{\varepsilon^2} \right\rceil \ .$$

8 Conclusion

We find that fair robust learning and optimization tasks can be expressed as maximin or minimax optimization problems, and efficiently solved via standard convex optimization methodology. In continuous optimization settings, as arise in machine learning and fair allocation with divisible goods, standard techniques from the adversarial optimization literature are appropriate. Furthermore, the inner maximization is often representable in closed form, and is thus amenable to general optimization settings, e.g., in sequential allocation [Viswanathan and Zick, 2023, Cousins et al., 2023b,c] tasks.

We show that the concept of robust fair objectives that arise from our constrained Rawlsian games preserve the Lipschitz and/or Hölder continuity properties of their underlying welfare or malfare concepts. This has implications to their optimization via convex optimization convergence rates, and is also directly relevant to the generalization error and sample complexity of fair learning these objectives.

We stress that while uncertainty about the world, in particular about the weights $\boldsymbol{w}$ (or relative sizes) of each group, does give rise to many standard fairness concepts, these fairness concepts are also inherently motivated through the lens of fairness itself. In other words, we help needy groups because we feel that it is the right thing to do, not because we labor under the delusion that we may someday *literally become them*. It is a deep philosophical question of human nature as to whether our capacity for empathy inherently predisposes us to think in this way, though such philosophical quandaries are well beyond the scope of this work.

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.

Sonja Michelle Amadae. *Rationalizing capitalist democracy: The cold war origins of rational choice liberalism*. University of Chicago Press, 2003.

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

Hutan Ashrafian. Engineering a social contract: Rawlsian distributive justice through algorithmic game theory and artificial intelligence. *AI and Ethics*, 3(4):1447–1454, 2023.

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

Walter Bossert and Kohei Kamaga. An axiomatization of the mixed utilitarian-maximin social welfare orderings. *Economic Theory*, 69(2):451–473, 2020.

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

Cyrus Cousins. *Bounds and Applications of Concentration of Measure in Fair Machine Learning and Data Science*. PhD thesis, Brown University, 2021b.

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. Revisiting fair-PAC learning and the axioms of cardinal welfare. In *Artificial Intelligence and Statistics (AISTATS)*, 2023.

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)*, 2022.

Cyrus Cousins, Justin Payan, and Yair Zick. Into the unknown: Assigning reviewers to papers with uncertain affinities. In *Proceedings of the 16th International Symposium on Algorithmic Game Theory*, 2023a.

Cyrus Cousins, Vignesh Viswanathan, and Yair Zick. Dividing good and better items among agents with submodular valuations. In *International Conference on Web and Internet Economics*. Springer, 2023b.

Cyrus Cousins, Vignesh Viswanathan, and Yair Zick. The good, the bad and the submodular: Fairly allocating mixed manna under order-neutral submodular preferences. In *International Conference on Web and Internet Economics*. Springer, 2023c.

Cyrus Cousins, Indra Elizabeth Kumar, and Suresh Venkatasubramanian. To pool or not to pool: Analyzing the regularizing effects of group-fair training on shared models. In *Artificial Intelligence and Statistics (AISTATS)*, 2024.

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.

Robert Deschamps and Louis Gevers. Leximin and utilitarian rules: A joint characterization. *Journal of Economic Theory*, 17(2):143–163, 1978.

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

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

Thibault Gajdos and John A Weymark. Multidimensional generalized Gini indices. *Economic Theory*, 26(3): 471–496, 2005.

Andrius Gališanka. Just society as a fair game: John Rawls and game theory in the 1950s. *Journal of the History of Ideas*, 78(2):299–308, 2017.

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

John C Harsanyi. Can the maximin principle serve as a basis for morality? A critique of John Rawls's theory. *American Political Science Review*, 69(2):594–606, 1975.

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.

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.

Tianyi Lin, Chi Jin, and Michael I Jordan. Near-optimal algorithms for minimax optimization. In *Conference on Learning Theory*, pages 2738–2779. PMLR, 2020.

Vishnu Suresh Lokhande, Kihyuk Sohn, Jinsung Yoon, Madeleine Udell, Chen-Yu Lee, and Tomas Pfister. Towards

group robustness in the presence of partial group labels. In *ICML 2022: Workshop on Spurious Correlations, Invariance and Stability*, 2022.

Alessio Mazzetto, Cyrus Cousins, Dylan Sam, Stephen H. Bach, and Eli Upfal. Adversarial multiclass learning under weak supervision with performance guarantees. In *International Conference on Machine Learning (ICML)*, pages 7534–7543. PMLR, 2021.

Blossom Metevier, Stephen Giguere, Sarah Brockman, Ari Kobren, Yuriy Brun, Emma Brunskill, and Philip S Thomas. Offline contextual bandits with high probability fairness guarantees. *Advances in neural information processing systems*, 32, 2019.

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

Arkadi Nemirovski. Prox-method with rate of convergence $O(\frac{1}{t})$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1): 229–251, 2004.

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

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

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.

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

Mark Schneider and Byung-Cheol Kim. The utilitarian-maximin social welfare function and anomalies in social choice. *Southern Economic Journal*, 87(2):629–646, 2020.

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.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(4):171–176, 1958.

Till Speicher, Hoda Heidari, Nina Grgić-Hlača, Krishna P Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2239–2248, 2018.

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.

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

Abraham Wald. Contributions to the theory of statistical estimation and testing hypotheses. *The Annals of Mathematical Statistics*, 10(4):299–326, 1939.

Abraham Wald. Statistical decision functions which minimize the maximum risk. *Annals of Mathematics*, pages 265–280, 1945.

John A Weymark. Generalized Gini inequality indices. *Mathematical Social Sciences*, 1(4):409–430, 1981.