Fairness Risk Measures

Robert C. Williamson [1] **Aditya Krishna Menon** [2]

Abstract

Ensuring that classifiers are non-discriminatory or *fair* with respect to a sensitive feature (e.g., race or gender) is a topical problem. Progress in this task requires fixing a definition of fairness, and there have been several proposals in this regard over the past few years. Several of these, however, assume either binary sensitive features (thus precluding categorical or real-valued sensitive groups), or result in non-convex objectives (thus adversely affecting the optimisation landscape). In this paper, we propose a new definition of fairness that generalises some existing proposals, while allowing for generic sensitive features and resulting in a convex objective. The key idea is to enforce that the expected losses (or *risks*) across each subgroup induced by the sensitive feature are commensurate. We show how this relates to the rich literature on *risk measures* from mathematical finance. As a special case, this leads to a new convex fairness-aware objective based on minimising the *conditional value at risk* (CVaR).

1. Introduction

Ensuring that learned classifiers are non-discriminatory or *fair* with respect to some sensitive feature (e.g., race or gender) is a topical problem (Pedreshi et al., 2008; Žliobaitė, 2017; Chouldechova et al., 2018). Progress on this problem requires that one agrees upon some pre-defined notion of fairness; to this end, there have been several definitions of fairness at both the individual (Dwork et al., 2012; Kusner et al., 2017; Speicher et al., 2018) and group level (Calders & Verwer, 2010; Feldman et al., 2015; Hardt et al., 2016; Zafar et al., 2017a; Heidari et al., 2019).

Recently, several works (Zafar et al., 2017a; Dwork et al., 2018; Hashimoto et al., 2018; Alabi et al., 2018; Speicher et al., 2018; Donini et al., 2018; Heidari et al., 2019) have abstracted earlier definitions of fairness by framing the problem in terms of *subgroup losses*. Intuitively, these works posit that a fair predictor incurs similar losses for each sensitive feature subgroup (e.g., men and women). One encourages fairness by minimising specific notions of disparity of subgroup losses. For specific choices of loss, this leads to a convex objective (Zafar et al., 2017c; Donini et al., 2018).

In this paper, we propose a new definition of fairness that follows this theme, but abstracts the notion of subgroup loss disparity. Our resulting framework is applicable for generic base losses, complex sensitive features (e.g., multi-valued), and results in a convex objective when using convex losses. In detail, our contributions are as follows:

(**C1**) building on notions of fairness in terms of subgroup errors (Zafar et al., 2017a; Dwork et al., 2018; Donini et al., 2018), we provide a new definition of fairness (Definition 3) requiring the average losses (or *risks*) for each sensitive feature subgroup have low *deviation*.

(**C2**) we draw a connection (Corollary 12) between our proposed definition of fairness and the rich literature on *risk measures* from mathematical finance (Artzner et al., 1999; Föllmer & Schied, 2011), thus allowing one to leverage tools and analyses from the latter.

(**C3**) we propose a new convex fairness-aware objective (Equation 26) based on minimising the *conditional value at risk* (CVaR) (Rockafellar & Uryasev, 2000), and relate it to existing learning objectives.

In a nutshell, our proposal is to break up the standard risk into risks on each *subgroup* defined by the sensitive feature. We combine these via an aggregator which measures the mean *and deviation* of the subgroup risks. By defining some axioms an aggregator should satisfy, we obtain a connection to risk measures from finance and operations research.

We remark that much of the work in the paper is in setting up the problem to easily exploit a wide body of existing results on risk measures; however, to our knowledge, the application of such tools to fairness is novel. The end result is a simple, powerful framework to learn fair classifiers.

In the sequel, after reviewing existing work (§2), we introduce our new definition of fairness (§3), and relate it to financial risk measures (§4). We detail a special case employing the conditional value at risk (§5), further develop our approach (§6), and confirm its empirical viability (§7).

[1]Australian National University [2]Google Research. Correspondence to: Robert C. Williamson <bob.williamson@anu.edu.au>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

2. Background

We briefly review the fairness-aware learning problem.

2.1. Standard and fairness-aware learning

Given pairs of instances (e.g., job applicants) and target labels (e.g., likelihood of repaying a loan), supervised learning concerns finding a predictor that best estimates the target label for new instances. Formally, suppose there is a *feature set* X, and *label set* Y. A *predictor* is any $f\colon X \to A$ for some *action set* A, where typically $A = Y$. Suppose we are given a class of predictors $\mathcal{F} \subseteq A^X$, and a *loss function* $\ell\colon Y \times A \to \mathbb{R}_{\geq 0}$ measuring the disagreement between a target label and its prediction. The *base goal* of learning is to find an $f^* \in \mathcal{F}$ minimising the *risk* or *expected loss*:[1]

$$f^* = \operatorname*{argmin}_{f \in \mathcal{F}} L(f) \text{ where } L(f) := \mathop{\mathbb{E}}_{\mathsf{X},\mathsf{Y}} \left[\ell(\mathsf{Y}, f(\mathsf{X}))\right], \quad (1)$$

where X, Y are drawn from some distribution over $X \times Y$.

In fairness-aware learning, we augment the base goal by requiring our predictor does not discriminate with respect to some sensitive feature (e.g., race). Formally, suppose there is a *sensitive set* S over which there is a random variable S, and that the feature set X contains S as a subset.[2] A *fairness measure* is some $\Lambda(\cdot)$ for which $\Lambda(\mathsf{Y}, f(\mathsf{X}), \mathsf{S})$ evaluates the level of discrimination of f. The *fairness goal* is to find an f minimising the risk *subject to* Λ being small: for $\epsilon \geq 0$,

$$f^* = \operatorname*{argmin}_{f \in \mathcal{F}} L(f) \text{ such that } \Lambda(\mathsf{Y}, f(\mathsf{X}), \mathsf{S}) \leq \epsilon. \quad (2)$$

2.2. Measures of perfect fairness

To design a fairness measure Λ, it is useful to decide what it means for a predictor to be *perfectly* fair. Most formalisms of perfect (group) fairness are statements of statistical independence. *Demographic parity* (Dwork et al., 2012) requires

$$\mathsf{A} \perp\!\!\!\perp \mathsf{S}, \quad (3)$$

so that knowledge of the predictions $\mathsf{A} := f(\mathsf{X})$ provides no knowledge of the sensitive feature S. For example, when $S = \{\texttt{male}, \texttt{female}\}$, this would mean that the distribution of predictions are identical for both men and women. On the other hand, *equalised odds* (Hardt et al., 2016) requires

$$\mathsf{A} \perp\!\!\!\perp \mathsf{S} \mid \mathsf{Y}, \quad (4)$$

so that *given* knowledge of the true label Y, knowledge of the predictions A provides no knowledge of the sensitive feature S. Continuing the previous example, this requires that the predictions do not discriminate between men and women *beyond* whatever power these have in predicting Y. Similarly, *lack of disparate mistreatment* (Zafar et al., 2017a) constrains the *subgroup error rates* to be identical:

$$(\forall s, s' \in S)\, \mathbb{P}(\mathsf{Y} \neq \mathsf{A} \mid \mathsf{S} = s) = \mathbb{P}(\mathsf{Y} \neq \mathsf{A} \mid \mathsf{S} = s'). \quad (5)$$

There are other extant notions of perfect fairness (Zafar et al., 2017b; Ritov et al., 2017; Heidari et al., 2018; Zhang & Bareinboim, 2018), including those for *individual* rather than group fairness (Dwork et al., 2012; Kusner et al., 2017).

2.3. Measures of approximate fairness

Notions of perfect fairness represent ideal statements about the world. When learning a classifier from a finite training sample, it is infeasible to guarantee perfect fairness on a test sample (Agarwal et al., 2018). In practice, one often instead works instead with measures of *approximate fairness*. The learner may then seek to achieve a *tradeoff* between fairness and accuracy (Menon & Williamson, 2018).

We highlight three popular measures of approximate fairness, using demographic parity (3) as the underlying perfect fairness notion for simplicity. The first is to look at the maximal deviation between subgroup predictions (Calmon et al., 2017), (Alabi et al., 2018, Section 5.2.2):

$$\Lambda_{\text{dev}}(\mathsf{A}, \mathsf{S}) = \sup_{a,s,s'} |\mathbb{P}(\mathsf{A} = a \mid \mathsf{S} = s) - \mathbb{P}(\mathsf{A} = a \mid \mathsf{S} = s')|.$$

This measure is popular for binary S, where it is known as the *mean difference score* (Calders & Verwer, 2010). However, it involves computing $|S|^2$ terms for categorical S, and is infeasible for real-valued S. The former issue can be addressed with a simple variant (Agarwal et al., 2018).

An elegant alternative is to recall that perfect fairness measures assert that certain random variables are independent. One may naturally measure approximate fairness by measuring their *degree* of independence. For example, one might quantify approximate demographic parity (3) via

$$\Lambda_{\text{MI}}(\mathsf{A}, \mathsf{S}) = \text{MI}(\mathsf{A}; \mathsf{S}) = \text{KL}(\mathbb{P}(\mathsf{A}, \mathsf{S}) \,\|\, \mathbb{P}(\mathsf{A}) \cdot \mathbb{P}(\mathsf{S})), \quad (6)$$

where MI denotes the *mutual information*, KL the Kullback-Leibler divergence, $\mathbb{P}(\mathsf{A}, \mathsf{S})$ the joint distribution over predictions and sensitive features, and $\mathbb{P}(\mathsf{A})$, $\mathbb{P}(\mathsf{S})$ the corresponding marginals. Since the MI measures the degree of independence of two random variables, Λ_{MI} is a natural measure of approximate demographic parity (Kamishima et al., 2012; Fukuchi et al., 2013; Calmon et al., 2017; Ghassami et al., 2018). One can replace the KL divergence in (6) with other measures of dissimilarity between distributions, e.g., an f-divergence (Komiyama & Shimao, 2017) or Hilbert-Schmidt criterion (Pérez-Suay et al., 2017).

Conceptually, measures based on (6) have appealing generality: in particular, they can seamlessly handle multi-class,

[1] We do not indicate the implicit dependence of $L(f)$ on the underlying distribution or loss ℓ for brevity.

[2] Omitting S from the feature set does not guarantee fairness, as it is typically correlated with other features (Pedreshi et al., 2008).

multi-label and continuous S. However, they typically result in a non-convex objective (Kamishima et al., 2012). An alternate measure that is similarly general, but convex, is the covariance between the target and sensitive features (Zafar et al., 2017c; Olfat & Aswani, 2018; Donini et al., 2018):

$$\Lambda_{\text{cov}}(\mathsf{A},\mathsf{S}) = \text{Cov}(\mathsf{A},\mathsf{S}) = \mathbb{E}[\mathsf{A}\cdot\mathsf{S}] - \mathbb{E}[\mathsf{A}]\cdot\mathbb{E}[\mathsf{S}]. \quad (7)$$

2.4. Fairness-aware algorithms

Having fixed a notion of perfect or approximate fairness, one may then go about designing a fairness-aware learning algorithm. Broadly, these follow one of three approaches:

(*a*) pre-process the training set to ensure fairness of *any* learned model (Zemel et al., 2013; Johndrow & Lum, 2017; Calmon et al., 2017; Adler et al., 2018; del Barrio et al., 2018; McNamara et al., 2019);

(*b*) post-process model predictions to ensure their fairness (Feldman et al., 2015; Hardt et al., 2016);

(*c*) directly ensure fairness by optimising (2) (Zafar et al., 2016; 2017a; Agarwal et al., 2018; Donini et al., 2018).

This paper focusses on methods of type (c); we defer implications for methods of types (a) and (b) to future work.

2.5. Scope of this paper

In relation to the above (necessarily incomplete) survey, the scope of the present work is in providing:

— a new notion of approximate fairness (Definition 3),
— a new method that optimises for this notion (§5), and
— a new connection between fairness and concepts from mathematical finance (Corollary 12).

In more detail, we consider fairness in terms of *subgroup risk*, following (Zafar et al., 2017a; Donini et al., 2018; Dwork et al., 2018; Alabi et al., 2018). Our new notion of approximate fairness is that these risks exhibit low *deviation*. By connecting this to *risk measures* in mathematical finance, we arrive at a *convex* objective for fairness-aware learning, applicable for generic sensitive features S, and with interesting connections to some existing learning paradigms.

3. Fairness as Subgroup Risk Deviation

We present our new measure of fairness by introducing the notion of *subgroup risks*, and using it to define natural measures of perfect (§3.2) and approximate fairness (§3.3). We also define some recurring notation, summarised in Table 1. The core idea of our proposal is to aggregate the subgroup risks by measuring their mean behaviour *and* deviance (Equations 14 and 15).

3.1. Subgroup risks

Observe that the sensitive feature S partitions the instance space X into subgroups (e.g., men and women). It will be useful to define two induced quantities. The first is the *subgroup risk* for a predictor f, which for any $s \in S$ is

$$L_s(f) := \mathop{\mathbb{E}}_{\mathsf{X},\mathsf{Y}|\mathsf{S}=s} \ell(\mathsf{Y}, f(\mathsf{X})). \quad (8)$$

The second is the random variable $\mathsf{L}(f) := L_{\mathsf{S}}(f)$ summarising all subgroup risks. For $|S| < \infty$, this is simply a discrete random variable taking on $|S|$ possible values, i.e., $\{L_s(f)\}_{s\in S}$, with corresponding probabilities $\mathbb{P}(\mathsf{S} = s)$.

Symbol	Meaning
ℓ, f	Base loss, predictor
$L(f)$	Risk of f on entire population
$L_s(f)$	Risk of f on subgroup with $S = s$
$\mathsf{L}(f)$	Random variable of all subgroup risks
$\mathcal{D}(\mathsf{L}(f))$	Deviation of subgroup risks
$\mathcal{R}(\mathsf{L}(f))$	Aggregation of subgroup risks

Table 1. Glossary of important symbols.

We can now rewrite the original risk $L(f)$ from (1) as an average over these subgroup risks:

$$L(f) = \mathop{\mathbb{E}}_{\mathsf{S}} \mathop{\mathbb{E}}_{\mathsf{X},\mathsf{Y}|\mathsf{S}} [\ell(\mathsf{Y}, f(\mathsf{X}))] = \mathbb{E}[\mathsf{L}(f)]. \quad (9)$$

The base goal of learning (1) is thus expressible as

$$f^* = \mathop{\text{argmin}}_{f\in\mathcal{F}} \mathbb{E}[\mathsf{L}(f)], \quad (10)$$

so that one seeks good average subgroup risk. Equally, we wish to select $f^* \in \mathcal{F}$ based on the expectations of the family of random variables $\{\mathsf{L}(f)\}_{f\in\mathcal{F}}$.

We now introduce our new measure of fairness. Following the discussion in §2, we do so in two steps: we start by settling on a notion of perfect fairness based on the subgroup risks, and then present an approximate version of the same.

3.2. Perfect fairness via subgroup risks

Every measure of fairness in §2.2 specifies that our predictor f behaves similarly across the sub-groups induced by S. We employ a notion of perfect fairness that is faithful to this.

Definition 1. *We say that a predictor $f \in \mathcal{F}$ is* perfectly fair *with respect to ℓ if* all subgroups attain the same average loss*; i.e., $\mathsf{L}(f)$ is a constant random variable, so that*

$$(\forall s, s' \in S) \mathop{\mathbb{E}}_{\mathsf{X},\mathsf{Y}|\mathsf{S}=s} \ell(\mathsf{Y}, f(\mathsf{X})) = \mathop{\mathbb{E}}_{\mathsf{X},\mathsf{Y}|\mathsf{S}=s'} \ell(\mathsf{Y}, f(\mathsf{X})). \quad (11)$$

Abstractly, the idea behind (11) is that the loss ℓ should ideally be chosen to capture all aspects of the problem ignoring fairness; perfect fairness means that regardless of the value of sensitive attribute, the performance does not vary. For a specific choice of ℓ, Definition 1 captures an existing notion of perfect fairness due to Zafar et al. (2017a).

Example 2. For the zero-one loss $\ell_{01}(y, f) = [\![y \neq f]\!]$, (11) reduces to the previously introduced (5):

$$(\forall s, s' \in S)\, \mathbb{P}(f(\mathsf{X}) \neq \mathsf{Y} \mid \mathsf{S} = s) = \mathbb{P}(f(\mathsf{X}) \neq \mathsf{Y} \mid \mathsf{S} = s').$$

Definition 1 is not new as a measure of perfect fairness. Indeed, Donini et al. (2018, Appendix H) considered essentially the same notion, with additional conditioning on $\mathsf{Y} = 1$. Several other recent works implicitly define perfect fairness in terms of subgroup risks (Dwork et al., 2018; Hashimoto et al., 2018; Alabi et al., 2018). Further, recent welfare-based notions of fairness (Speicher et al., 2018; Heidari et al., 2019) also posit that fair classifiers have equally distributed *benefit* (i.e., negative losses; see Remark 13).

However, we build on Definition 1 to provide a novel notion of *approximate* fairness, one which has appealing properties and provides a bridge to the tools of financial risk measures.

3.3. Approximate fairness via subgroup deviations

A natural way to design an approximate fairness measure based on (11) is to ensure that the subgroup risks $\mathsf{L}(f)$ are roughly constant. Formally, for some *deviation measure* $\mathcal{D}$ of the non-constancy of a random variable (e.g., the standard deviation), we will require that $\mathcal{D}(\mathsf{L}(f))$ is small.

Definition 3. *Let $\mathcal{D}(\cdot)$ be a measure of deviation of a random variable. For any $\epsilon > 0$, we say that $f \in \mathcal{F}$ is ϵ-*approximately fair with respect to $\mathcal{D}$ and ℓ *if* the average subgroup losses have small deviation*; i.e., $\mathcal{D}(\mathsf{L}(f)) \leq \epsilon$.*

Definition 3 is applicable for generic S (e.g., real-valued). For the case of binary S, it is consistent with existing notions of approximate fairness, as we now illustrate.

Example 4. Suppose $S = \{0, 1\}$, and that we use deviation measure $\mathcal{D}_{\mathrm{SD}}(\cdot) = \sigma(\cdot)$, where σ is the standard deviation of a random variable. Fix $f \in \mathcal{F}$, and for brevity write the subgroup risks as $L_s := L_s(f)$ and $\mathsf{L} := \mathsf{L}(f)$. We have

$$\mathcal{D}_{\mathrm{SD}}(\mathsf{L}) = \sqrt{\mathbb{E}(\mathsf{L}^2) - \mathbb{E}^2(\mathsf{L})} = \frac{1}{2} \cdot |L_0 - L_1|. \quad (12)$$

Recall that the subgroup risks L_s depend on the underlying loss ℓ. Employing the zero-one loss ℓ_{01} in (12) yields $\mathcal{D}_{\mathrm{SD}}(\mathsf{L}) = \frac{1}{2} \cdot |\mathbb{P}(f(\mathsf{X}) \neq \mathsf{Y} \mid \mathsf{S} = 0) - \mathbb{P}(f(\mathsf{X}) \neq \mathsf{Y} \mid \mathsf{S} = 1)|$,

i.e., the mean-difference score (Calders & Verwer, 2010) applied to the lack of disparate mistreatment (5).

3.4. Fairness-aware learning via subgroup aggregation

To achieve approximate fairness according to Definition 3, we may augment the standard expected risk (10) with a penalty term: for suitable $\lambda > 0$, we may find

$$f^* = \operatorname*{argmin}_{f \in \mathcal{F}} L(f) + \lambda \cdot \mathcal{D}(\mathsf{L}(f)), \quad (13)$$

so that we find a predictor that predicts the target label, *but* does so consistently across all subgroups. Observe now that in light of (9), we can succinctly summarise (13) as

$$f^* = \operatorname*{argmin}_{f \in \mathcal{F}} \mathcal{R}_\lambda(\mathsf{L}(f)) \quad (14)$$

$$\mathcal{R}_\lambda(\mathsf{L}) := \mathbb{E}(\mathsf{L}) + \lambda \cdot \mathcal{D}(\mathsf{L}). \quad (15)$$

We make two observations. First, both standard risk minimisation (10) and (14) minimise a function of the subgroup risks $\mathsf{L}(f)$; the only difference is the choice of *subgroup risk aggregator* $\mathcal{R}_\lambda$. In (14), we aim to ensure that the subgroup risks are small, *and* that they are roughly commensurate. Intuitively, the latter ensures that we do not exhibit systematic bias in terms of mispredictions on one of the subgroups.

Second, given a finite sample $\{(x_j, y_j, s_j)\}_{j=1}^m$, one may solve the empirical analogue of (14): we minimise $\mathcal{R}_\lambda(\hat{\mathsf{L}}(f))$, where $\hat{\mathsf{L}}$ comprises empirical subgroup risks, i.e., we employ empirical expectations in (8); see, e.g., (27).

We make (14) concrete with an example.

Example 5. For the setting of Example 4, for deviation measure $\mathcal{D}_{\mathrm{SD}}$ we have the fairness-aware objective (14)

$$\mathcal{R}_{\mathrm{SD},\lambda}(\mathsf{L}) = \mathbb{E}(\mathsf{L}) + \lambda \cdot \mathcal{D}_{\mathrm{SD}}(\mathsf{L}) = \mathbb{E}(\mathsf{L}) + \frac{\lambda}{2} \cdot |L_1 - L_2|, \quad (16)$$

so that we ensure that the average subgroup risk is small, *and* that the two subgroup risks are commensurate.

Remark 6. For binary S, previous methods sharing our notion of perfect fairness (Definition 1) have objectives similar to (16). There is, however, a subtle difference: in (14), we use the *same* loss ℓ to measure the standard risk, and its deviation across subgroups. However, Zafar et al. (2017a); Donini et al. (2018) employ *different* losses for these two terms. Specifically, they employ a linear loss for the deviation, which corresponds to measuring the covariance between A and S per (7). This choice is crucial to ensuring convexity of their objective; we shall see that one can preserve convexity for other ℓ by instead modifying $\mathcal{D}$.

Remark 7. The idea of moving beyond expectations to a general aggregation of the *per-instance* losses has precedent in learning theory (Chapelle et al., 2001; Maurer & Pontil, 2009) and robust optimisation (Duchi et al., 2016; Gotoh et al., 2018). These encourage the loss deviance across *all* samples to be small, i.e., effectively, they treat each instance as its own group. Similar connections will also arise in §5.3.

A natural question at this stage is what constitutes a "sensible" choice of deviation measure $\mathcal{D}$. One may of course proceed with intuitively reasonable choices, such as the standard deviation (Example 4); however, we shall now axiomatise the properties we would like *any* sensible deviation measure to satisfy. This shall lead to an admissible family of *fairness risk measures*.

4. Fairness Risk Measures

The proposal of the previous section was boiled down to a simple recipe in (14): rather than minimise the average of the subgroup risks, we minimise a general functional $\mathcal{R}$ of them, which involves an expectation *and* deviation $\mathcal{D}$. We now axiomatically specify the class of admissible subgroup aggregators $\mathcal{R}$, which will in turn specify the class of admissible deviations $\mathcal{D}$ (Theorem 14).

The technical aspects here are not new; rather, we leverage results in the risk measures literature (particularly Rockafellar & Uryasev (2013)) for a novel application to fairness.

4.1. Fairness risk measures: an axiomatic definition

At this stage, we employ a slight change of terminology: we shall refer to $\mathcal{R}$ as a risk *measure* rather than *aggregator*. The reasoning for this change will become evident in the next section. With this, we define the class of *fairness risk measures* $\mathcal{R}$ as those satisfying seven simple mathematical axioms. In what follows, let $\mathcal{L}^2(S)$ comprise real-valued random variables over S with finite second moment.

Definition 8. *We say* $\mathcal{R}\colon \mathcal{L}^2(S) \to \bar{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$ *is a* fairness risk measure *if, for any* $\mathsf{Z}, \mathsf{Z}' \in \mathcal{L}^2(S)$ *and* $C \in \mathbb{R}$*, it satisfies the following axioms (F1)–(F7):*

F1 Convexity $\mathcal{R}((1-\lambda)\mathsf{Z}+\lambda\mathsf{Z}') \leq (1-\lambda)\mathcal{R}(\mathsf{Z})+\lambda\mathcal{R}(\mathsf{Z}')$, $\forall \lambda \in (0,1)$.

F2 Positive Homogeneity $\mathcal{R}(0) = 0$, $\mathcal{R}(\lambda\mathsf{Z}) = \lambda\mathcal{R}(\mathsf{Z})$ $\forall \lambda > 0$.

F3 Monotonicity $\mathcal{R}(\mathsf{Z}) \leq \mathcal{R}(\mathsf{Z}')$ *if* $\mathsf{Z} \leq \mathsf{Z}'$ *almost surely.*

F4 Lower Semicontinuity $\{\mathsf{Z}\colon \mathcal{R}(\mathsf{Z}) \leq C\}$ *is closed.*

F5 Translation Invariance $\mathcal{R}(\mathsf{Z} + C) = \mathcal{R}(\mathsf{Z}) + C$.

F6 Aversity $\mathcal{R}(\mathsf{Z}) > \mathbb{E}(\mathsf{Z})$ *for any non-constant random variable* Z.

F7 Law Invariance $\mathcal{R}(\mathsf{Z}) = \mathcal{R}(\mathsf{Z}')$ *if* $\mathbb{P}_{\mathsf{Z}} = \mathbb{P}_{\mathsf{Z}'}$.

In Appendix A, we argue why each of these axioms is natural when $\mathcal{R}$ is used per (14) to ensure fairness across subgroups. Here, we highlight the import of two axioms:

Convexity (F1) is desirable because without it, the risk could be decreased by more fine grained partitioning as we now show. Suppose $S = \{0,1\}$ induces a partition (X_0, X_1) of X. Then $\mathsf{L} = \mathsf{L}^0 + \mathsf{L}^1$, where L^i is the restriction of L to X_i, so that e.g. $\mathsf{L}^0_s = [\![s = 0]\!] \cdot \mathbb{P}(\mathsf{S} = 0) \cdot \mathsf{L}_s$. Now if $\mathcal{R}$ were not convex, it would not be sub-additive, and so $\mathcal{R}(\mathsf{L}) = \mathcal{R}(\mathsf{L}^0 + \mathsf{L}^1) > \mathcal{R}(\mathsf{L}^0) + \mathcal{R}(\mathsf{L}^1)$. That is, by splitting into subgroups we could automatically make our risk measure smaller, which is undesirable.

The above primary motivation for convexity has a desirable side benefit: convexity combined with F3 implies that if $f \mapsto L(f)$ is convex, then so is $f \mapsto \mathcal{R}(\mathsf{L}(f))$. Thus, for convex ℓ and $\mathcal{F}$, encouraging fairness does not pose an optimisation burden, in contrast to some existing approaches (Kamishima et al., 2012; Zafar et al., 2016).

Aversity (F6) has a clear justification, as it penalises deviation from perfect fairness (by Definition 1, this corresponds to constant L); this is essential for any fairness measure.

Remark 9. The subgroup risk aggregator $\mathcal{R}_{\mathrm{SD}}$ corresponding to the standard deviation (16) does not satisfy F1, and thus is not a fairness risk measure. This does not necessarily preclude its use; while Appendix A makes a case that these measures are sensible to use, we do *not* claim that these are the *only* legitimate measures. Nonetheless, we now see that a wide class of measures satisfy F1–F7.

4.2. Relation to financial risk measures

In mathematical finance, a *risk measure* (Artzner et al., 1999) is a quantification of the potential loss associated with a position, i.e., a function $\rho\colon \mathcal{L}^2(S) \to \mathbb{R}$ whose input is a random variable, being the possible outcomes for a position. We now show the intimate relationship between fairness risk measures and two classes of risk measures widely studied in finance and operations research (Artzner et al., 1999; Pflug & Romisch, 2007; Krokhmal et al., 2011; Föllmer & Schied, 2011; Rockafellar & Uryasev, 2013). The first class is readily defined in terms of our existing axioms.

Definition 10. *We say* $\mathcal{R}\colon \mathcal{L}^2(S) \to \bar{\mathbb{R}}$ *is a* coherent measure of risk *(Artzner et al., 1999) if it satisfies F1 — F5.*

The second class requires two additional axioms:

F8 Translation Equivariance $\mathcal{R}(\mathsf{Z}) = C$ for any constant random variable Z taking value $C \in \mathbb{R}$.

F9 Positivity under non-constancy $\mathcal{R}(\mathsf{Z}) \geq 0$, with equality if and only if Z is constant.

Equipped with this, we have the following definition.

Definition 11. *We say* $\mathcal{R}\colon \mathcal{L}^2(S) \to \bar{\mathbb{R}}$ *is a* regular measure of risk *(Rockafellar & Uryasev, 2013) if it satisfies F1, F4, F6 and F8. Similarly,* $\mathcal{D}\colon \mathcal{L}^2(S) \to \bar{\mathbb{R}}$ *is a* regular measure of deviation *if satisfies F1, F4 and F9.*

Using $\mathsf{Z} = 0$, (F5 $\wedge$ F6) $\implies$ F8. We thus conclude that:

Corollary 12. *Every fairness risk measure is a coherent and regular measure of risk satisfying law-invariance.*[3]

Remark 13. Our chosen axioms were inspired by *risk* measures. Recently, Speicher et al. (2018); Heidari et al. (2019) proposed axioms inspired by *inequality* measures. The two notions can be related; see Appendix B. In brief, one can start with a risk measure and induce an inequality measure, and vice versa. In both directions, many, but not all, of the desirable attributes of the induced measures are implied by combinations of the attributes of the inducing measure.

[3] Law-invariance is fortunately satisfied by most widely-used measures (Rockafellar et al., 2006; Pflug & Romisch, 2007).

4.3. Practical implications

Connecting fairness and financial risk measures is not merely of conceptual interest. In particular, Corollary 12 lets us construct fairness risk measures $\mathcal{R}$ given a regular measure of deviation $\mathcal{D}$ via $\mathcal{R}(\mathsf{Z}) = \mathbb{E}(\mathsf{Z}) + \mathcal{D}(\mathsf{Z})$. This is a consequence of the following *quadrangle theorem*.

Theorem 14 (Rockafellar & Uryasev (2013))**.** *The relations*

$$\mathcal{R}(\mathsf{Z}) = \mathbb{E}(\mathsf{Z}) + \mathcal{D}(\mathsf{Z}) \text{ and } \mathcal{D}(\mathsf{Z}) = \mathcal{R}(\mathsf{Z}) - \mathbb{E}(\mathsf{Z}) \quad (17)$$

give a one-to-one correspondence between regular measures of risk $\mathcal{R}$ and regular measures of deviation $\mathcal{D}$. Further, $\mathcal{R}$ is positively homogeneous iff $\mathcal{D}$ is positively homogeneous; and monotonic iff $\mathcal{D}(\mathsf{Z}) \leq \sup \mathsf{Z} - \mathbb{E}(\mathsf{Z})$ for all $\mathsf{Z} \in \mathcal{L}^2(S)$.

Remark 15. Using the construction in (17), we arrive at risk aggregators $\mathcal{R}$ that are an expectation plus a deviance $\mathcal{D}$. By contrast, in (13) we applied a scalar λ to the deviance. This is equivalent to using a new deviance $\mathcal{D}_\lambda := \lambda \cdot \mathcal{D}$.

Corollary 12 also allows us to import well-studied financial risk measures for use in a fairness context, as we now show.

5. The CVaR-fairness Risk Measure

We now explicate a special case of our framework based on the conditional value of risk (CVaR), which yields a simple objective (Equation 26) related to the ν-SVM.

5.1. CVaR as a fairness risk measure

We first recall the definition of CVaR. For $\alpha \in (0, 1)$ and random variable Z, let $q_\alpha(\mathsf{Z})$ be the quantile at level α. The *conditional value at risk* is (Rockafellar & Uryasev, 2000)[4]

$$\mathrm{CVaR}_\alpha(\mathsf{Z}) := \mathbb{E}(\mathsf{Z} \mid \mathsf{Z} > q_\alpha(\mathsf{Z})), \quad (18)$$

i.e., it measures the tail behaviour of Z. Now define

$$\mathcal{R}_{\mathrm{CV},\alpha}(\mathsf{Z}) := \mathrm{CVaR}_\alpha(\mathsf{Z}) \quad (19)$$
$$\mathcal{D}_{\mathrm{CV},\alpha}(\mathsf{Z}) := \mathrm{CVaR}_\alpha(\mathsf{Z} - \mathbb{E}(\mathsf{Z})). \quad (20)$$

Intuitively, $\mathcal{D}_{\mathrm{CV},\alpha}$ measures the tail behaviour of $\mathsf{Z}' = \mathsf{Z} - \mathbb{E}(\mathsf{Z})$, i.e., how much Z deviates above its mean.

One has that $\mathcal{R}_{\mathrm{CV},\alpha}$ and $\mathcal{D}_{\mathrm{CV},\alpha}$ are regular, coherent measures of risk and deviation respectively (Rockafellar & Uryasev, 2013). By Theorem 14, one may equally write $\mathcal{D}_{\mathrm{CV},\alpha}(\mathsf{Z}) = \mathrm{CVaR}_\alpha(\mathsf{Z}) - \mathbb{E}(\mathsf{Z})$. Further, $\mathcal{R}_{\mathrm{CV},\alpha}$ is a fairness risk measure with fairness-aware objective (14)

$$\min_{f \in \mathcal{F}} \mathrm{CVaR}_\alpha(\mathsf{L}(f)) = \min_{f \in \mathcal{F}} \mathbb{E}(\mathsf{L}(f)) + \mathcal{D}_{\mathrm{CV},\alpha}(\mathsf{L}(f)). \quad (21)$$

[4]We gloss over the subtleties of defining quantiles when Z has atomic components; see (Rockafellar & Uryasev, 2013).

Here, $\alpha \in (0, 1)$ is a tuning parameter. From (18), increasing α focusses attention to the most extreme values of $\mathsf{L}(f)$, i.e., the largest subgroup risks. Interestingly, the limiting cases of α relate to existing fairness principles.[5] Per Rockafellar (2007, Equation 5.8), as $\alpha \to 1$, (21) becomes

$$\min_{f \in \mathcal{F}} \max_{s \in S} L_s(f), \quad (22)$$

i.e., we seek *all* subgroup risks to be small, in line with the maximin principle (Rawls, 1971). As $\alpha \to 0$, (21) becomes

$$\min_{f \in \mathcal{F}} \mathbb{E}_{\mathsf{S}}(L_{\mathsf{S}}(f)),$$

i.e., we seek the *average* subgroup risks to be small, in line with the impartial observer principle (Harsanyi, 1977) for uniform S (see §6.1). To intuit the effect of generic $\alpha \in (0, 1)$, suppose $n = |S| < \infty$, and S has uniform distribution. Then,

$$\mathrm{CVaR}_\alpha(\mathsf{L}(f)) = \frac{\lambda_\alpha}{k_\alpha} \sum_{i=1}^{k_\alpha} L_{[i]}(f) + (1 - \lambda_\alpha) \cdot L_{[k_\alpha+1]}(f), \quad (23)$$

where $L_{[i]}(f)$ denotes the ith largest subgroup risk, $k_\alpha := \lceil n\alpha \rceil$ and λ_α is a weighting parameter given by Rockafellar & Uryasev (2002, Proposition 8). When k_α is an integer,

$$\mathrm{CVaR}_\alpha(\mathsf{L}(f)) = \frac{1}{k_\alpha} \sum_{i=1}^{k_\alpha} L_{[i]}(f), \quad (24)$$

Minimising (21) seeks that the average of the *largest* subgroup risks is small. This tightens the range of subgroup risks, thus ensuring they are commensurate.

Remark 16. The maximal subgroup risk (22) was also considered in Hashimoto et al. (2018); Mohri et al. (2019), the former motivated by settings where group identity is *unknown*. Objectives that interpolate between maximum and average subgroup risk have been proposed, e.g., Alabi et al. (2018, Section 6.1). These are similar in spirit to (23); note however that (23) allows any $\alpha \in (0, 1)$, and thus can partially account for the $(k_\alpha + 1)$th largest subgroup risk.

5.2. Optimising CVaR-fairness

Using CVaR as an aggregator (or deviance measure) yields intuitive objectives. Further, these are feasible to optimise. Optimisation of quantities based on the CVaR is aided by a *variational representation*: for any $\alpha \in (0, 1)$ and random variable Z, (Rockafellar & Uryasev, 2000, Theorem 1)

$$\mathrm{CVaR}_\alpha(\mathsf{Z}) = \min_{\rho \in \mathbb{R}} \left\{ \rho + \frac{1}{1 - \alpha} \cdot \mathbb{E}[\mathsf{Z} - \rho]_+ \right\}. \quad (25)$$

[5]Thus, CVaR_α serves as an effective alternative to the range of fairness measures considered by Traub et al. (2005), who performed an empirical study of people's attitudes to the tradeoff between utility and fairness, and found that something "in-between" the proposals of Harsanyi and Rawls fit the data best.

Consequently, the CVaR-fairness objective (21) becomes

$$\min_{f\in\mathcal{F},\rho\in\mathbb{R}} \left\{ \rho + \frac{1}{1-\alpha}\cdot \mathbb{E}[\mathsf{L}(f)-\rho]_+ \right\}. \quad (26)$$

This is a convex objective when $f \mapsto \mathsf{L}(f)$ is convex (e.g., using a convex base ℓ and $\mathcal{F}$). Given a finite sample $\{(x_j, y_j, s_j)\}_{j=1}^m$ with $n = |S| < +\infty$, this becomes

$$\min_{f\in\mathcal{F},\rho\in\mathbb{R}} \left\{ \rho + \frac{1}{n\cdot(1-\alpha)} \sum_{s\in S} \left[\frac{1}{m_s} \sum_{j:\, s_j=s} \ell(y_j, f(x_j)) - \rho \right]_+ \right\}, \quad (27)$$

for m_s the number of examples with sensitive feature s. In words, for fixed ρ, we find a predictor $f \in \mathcal{F}$ which minimises the average of "hard" subgroup risks.

5.3. Relation to existing paradigms

Fan et al. (2017) considered the problem of learning a robust binary classifier given a sample $\{(x_j, y_j)\}_{j=1}^m$ and loss $\ell\colon Y \times A \to \mathbb{R}$. To achieve this, it was proposed to minimise the average of the *top-k per-instance* losses for $k \ll m$:

$$\min_{f\in\mathcal{F}} \frac{1}{k} \sum_{i=1}^{k} \ell_{[i]}(f), \quad (28)$$

where $\ell_{[i]}(f)$ is the ith largest element of the per-instance losses $[\,\ell(y_j, f(x_j))\,]_{j=1}^m$. Following (24), this is equal to[6]

$$\min_{f\in\mathcal{F}} \mathrm{CVaR}_{\alpha_k}(\mathsf{L}_{\mathrm{inst}}(f))$$

where $\alpha_k := k/m$, and $\mathsf{L}_{\mathrm{inst}}(f)$ is the discrete random variable of *per-instance losses*, with values $\{\ell(y_j, f(x_j)\}_{j=1}^m$. Consequently, despite its motivating goal being ostensibly different, this objective is a special case of our framework where *each instance belongs to a separate group*.

CVaR also arises in the ν-SVM (Schölkopf et al., 2000), which alternately parametrises the SVM with $\nu \in (0, 1)$, and whose objective is expressible as (Gotoh & Takeda, 2005; Takeda & Sugiyama, 2008; Tsyurmasto et al., 2014)

$$\min_{f\in\mathcal{H}} \frac{1}{2}\|f\|_{\mathcal{H}}^2 + \nu\cdot \mathrm{CVaR}_{1-\nu}(\mathsf{M}(f)),$$

where $\mathsf{M}(f)$ is the random variable of per-instance margins, taking values $[\,-y_j\cdot f(x_j)\,]_{j=1}^m$. This is a special case of our framework where each instance is a separate group, and one employs the "linear" loss $\ell(y, f) = -y\cdot f$: while the ν-SVM down-weights any *instance* with low *margin error*, we down-weight any *subgroup* with low *average loss*.

[6]The connection to CVaR was not explicitly noted in Fan et al. (2017). However, they employed the variational representation (25) as derived in a different context by Ogryczak & Tamir (2003).

6. Extensions and Discussion

We briefly observe some extensions of our formulation.

6.1. Sensitive feature weighting

In forming our fairness-aware objective (14), we employed the standard risk $L(f)$, which is a weighted sum of the subgroup risks (Equation 9). The default weighting is the underlying sensitive feature distribution. However, one could easily apply different a weighting ν_S to privilege certain groups over others. For $|S| < \infty$, we could define (c.f. (9))

$$L(f; \nu_S) := \mathop{\mathbb{E}}_{\mathsf{S}\sim\nu_S}[L_{\mathsf{S}}(f)] = \sum_{s\in S} \nu_S(s)\cdot L_i(f). \quad (29)$$

For example, when $S = \{0, 1\}$, if one felt that individuals with $s = 0$ were more important to treat well, one could simply put a large mass on 1, e.g. $\nu_S(0) = 0.9$ and $\nu_S(1) = 0.1$. The effects of imposing $\mathsf{S} \sim \nu_S$ will similarly be reflected in one's deviation measure $\mathcal{D}(L_{\mathsf{S}}(f))$.

To treat both groups equally in terms of risk, one could alternately choose ν_S to be uniform. This forms the basis for Harsanyi's principle of justice (Harsanyi, 1977), and would be analogous to the use of the balanced error in classification (Brodersen et al., 2010; Menon et al., 2013).

6.2. Non-binary sensitive features

Our examples thus far have focussed on binary S. However, the risk measures underpinning our framework seamlessly handle generic S. We make this concrete with two examples. The first is where $S = \mathbb{R}_{\geq 0}$ (as is appropriate for a person's income, e.g.). Then, for $\alpha \in (0, 1)$ and measure ν_S over S per (29), the CVaR-fairness objective (26) is:

$$\min_{f\in\mathcal{F},\rho\in\mathbb{R}} \left\{ \rho + \frac{1}{1-\alpha}\cdot \int_S [L_s(f)-\rho]_+ \,\nu_S(\mathrm{d}s) \right\}. \quad (30)$$

On a finite sample $\{(x_j, y_j, s_j)\}_{j=1}^m$ with all s_j's distinct for simplicity, taking the empirical measure $\hat{\nu}_S$ gives

$$\min_{f\in\mathcal{F},\rho\in\mathbb{R}} \left\{ \rho + \frac{1}{1-\alpha}\cdot \frac{1}{m} \sum_{j=1}^{m} [\ell(y_j, f(x_j)) - \rho]_+ \right\}, \quad (31)$$

so that each instance is considered as belonging to the same group. Interestingly, this is equivalent to the top-k objective (28) for $k = m\alpha$. However, one may consider other natural alternatives; e.g., one may construct a non-parametric estimate of ν_S from the given sample, and use this in (30).

The case of multiple sensitive features $\{S_1, \ldots, S_l\}$ can be similarly handled: all one needs to do is define a suitably structured S, and a valid measure ν_S over S. As an example, one can set $S := S_1 \times \cdots \times S_l$ and define ν_S as the product of measures ν_{S_i} on each individual sensitive feature.

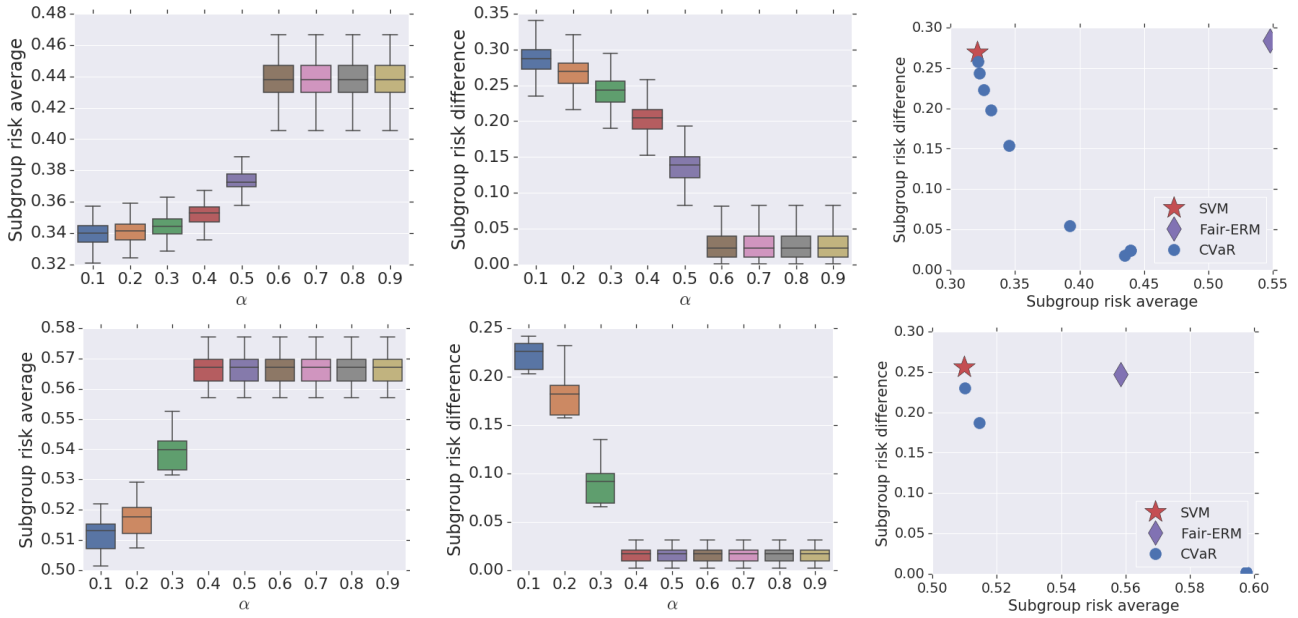

Figure 1. Results on `synth` (top) and `adult` (bottom) datasets. The left and middle panel show that as α is varied, CVaR-based optimisation results in a decrease in predictive accuracy and fairness violation, as measured by the average and absolute difference of the subgroup risks. The right panel overlays the performance of CVaR-based optimisation for various α with that of two baselines.

7. Empirical Illustration

We demonstrate that the CVaR-fairness minimiser (27) empirically yields reasonable fairness-accuracy tradeoffs. We present results on a synthetic two-dimensional dataset (`synth`) from Donini et al. (2018), where there is a single binary sensitive feature S, and the UCI `adult` dataset with gender as the binary S. We use square-hinge $\ell_{\text{sh}}(y, f) = [1 - yf]_+^2$ as our base loss, and regularised linear scorers as our $\mathcal{F}$. We use the validation procedure of Donini et al. (2018) to tune the regularisation strength, using balanced error as the base measure.

We assess CVaR-based optimisation (27) as α is tuned in $\{0.1, 0.2, \ldots, 0.9\}$. For each α, we compute the optimal empirical predictor's subgroup risks L_0, L_1 per (8). We then compute their average to assess the classifier's ability to predict the target Y, and their absolute difference to assess the classifier's ability to treat the subgroups equally. This is repeated over 100 random 80—20% train-test splits.

The left and middle panels of Figure 1 evince that α allows one to tune between predictive accuracy and fairness (in the sense of Definition 3): as α is increased, the subgroup risk's absolute difference decreases, while their average increases. This is as expected: for $\alpha \to 1^-$, the CVaR method explicitly minimises the maximal subgroup risk. The right panel of Figure 1 compares (on one train-test split) the fairness-accuracy tradeoff against a standard SVM, and the fair-ERM approach of Donini et al. (2018). For suitable α, CVaR achieves relatively favourable tradeoffs.

We make two qualifying remarks on the scope of the above results. First, the subgroup risks used to measure fairness and accuracy employ the *surrogate* loss used for training. Often, one may be interested in assessing the subgroup risks employing the 0-1 loss instead. Second, they do not focus attention on performance when $\mathsf{Y} = 1$. By contrast, fair-ERM is designed to control the equality of opportunity measure (4), which performs such conditioning. In Appendix C, we present additional plots for both these cases, and for a setting with real-valued S. While more extensive experiments are apposite, the above indicates the potential in further studying fairness risk measures.

8. Conclusion and Future Work

We proposed a new definition of fairness that generalises some existing proposals, while allowing for generic sensitive features and resulting in a convex objective. The key idea is to enforce that the expected losses (or *risks*) across each subgroup induced by the sensitive feature are commensurate. We showed how this relates to the rich literature on *risk measures* from mathematical finance. As a special case, this leads to a new convex fairness-aware objective based on minimising the *conditional value at risk* (CVaR).

Our relating of fairness and risk measures motivates study of risk measures beyond CVaR, e.g., spectral measures (Acerbi, 2002), optimised certainty equivalents (Ben-Tal & Teboulle, 2007), & entropic value at risk (Ahmadi-Javid, 2012).

Acknowledgments

Thanks to Katie Steele and Young Lee for useful discussions, and also to the reviewers of FAT*19 who provided comments on an earlier version. This work was supported by the ARC and DATA61. Some of the work was completed while AKM was with the Australian National University.

References

Acerbi, C. Spectral measures of risk: A coherent representation of subjective risk aversion. *Journal of Banking & Finance*, 26(7):1505–1518, 2002.

Adler, P., Falk, C., Friedler, S. A., Nix, T., Rybeck, G., Scheidegger, C., Smith, B., and Venkatasubramanian, S. Auditing black-box models for indirect influence. *Knowledge and Information Systems*, 54(1):95–122, January 2018.

Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J., and Wallach, H. A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 60–69, 2018.

Ahmadi-Javid, A. Entropic value-at-risk: A new coherent risk measure. *Journal of Optimization Theory and Applications*, 155(3):1105–1123, Dec 2012.

Alabi, D., Immorlica, N., and Kalai, A. T. Unleashing linear optimizers for group-fair learning and optimization. In *Proceedings of the 31st Conference On Learning Theory*, pp. 2043–2066, 06–09 Jul 2018.

Artzner, P., Delbaen, F., Eber, J.-M., and Heath, D. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999.

Atkinson, A. B. On the measurement of inequality. *Journal of Economic Theory*, 2:244–263, 1970.

Atkinson, A. B. Economics as a moral science. *Economica*, 76:791–804, 2009.

Ben-Tal, A. and Teboulle, M. An old-new concept of convex risk measures: the optimized certainty equivalent. *Mathematical Finance*, 17(3):449–476, 2007.

Bennett, C. J. and Zitikis, R. Ignorance, lotteries, and measures of economic inequality. *Journal of Economic Inequality*, 13:309–216, 2015.

Boulding, K. E. Economics as a moral science. *The American Economic Review*, 59(1):1–12, 1969.

Brodersen, K. H., Ong, C. S., Stephan, K. E., and Buhmann, J. M. The balanced accuracy and its posterior distribution. In *Proceedings of the 20th International Conference on Pattern Recognition*, pp. 3121–3124, 2010.

Calders, T. and Verwer, S. Three Naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.

Calmon, F. d. P., Wei, D., Vinzamuri, B., Ramamurthy, K. N., and Varshney, K. R. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pp. 3995–4004, 2017.

Chakravarty, S. R. Measuring inequality: The axiomatic approach. In Silber, J. (ed.), *Handbook of Income Inequality Measurement*, pp. 163–186. Springer, 1999.

Chapelle, O., Weston, J., Bottou, L., and Vapnik, V. Vicinal risk minimization. In *Advances in Neural Information Processing Systems 13*, pp. 416–422. MIT Press, 2001.

Chouldechova, A., Prado, D. B., Fialko, O., and Vaithianathan, R. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*, pp. 134–148, 2018.

Cochran, K. P. Economics as a moral science. *Review of Social Economy*, 32(2):186–195, 1974.

Cowell, F. A. Measurement of inequality. In *Handbook of Income Distribution*, pp. 87–166. Elsevier, 2000.

Cowell, F. A. *Measuring Inequality (3rd Edition)*. Oxford University Press, 2011.

Dana, R.-A. A representation result for concave Schur concave functions. *Mathematical Finance*, 15(4):613–634, October 2005.

del Barrio, E., Gamboa, F., Gordaliza, P., and Loubes, J.-M. Obtaining fairness using optimal transport theory. *arXiv e-prints*, art. arXiv:1806.03195, June 2018.

Deutsch, J. and Silber, J. Inequality decomposition by population subgroups and the analysis of interdistributional inequality. In Silber, J. (ed.), *Handbook of Income Inequality Measurement*, chapter 13, pp. 363–403. Springer Science & Business Media, 1999.

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J. S., and Pontil, M. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems 31*, pp. 2796–2806. 2018.

Duchi, J., Glynn, P., and Namkoong, H. Statistics of Robust Optimization: A Generalized Empirical Likelihood Approach. *arXiv e-prints*, art. arXiv:1610.03425, October 2016.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Innovations in Theoretical Computer Science Conference*, pp. 214–226, 2012.

Dwork, C., Immorlica, N., Kalai, A. T., and Leiserson, M. Decoupled classifiers for group-fair and efficient machine learning. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pp. 119–133, 2018.

Fan, Y., Lyu, S., Ying, Y., and Hu, B. Learning with average top-k loss. In *Advances in Neural Information Processing Systems*, pp. 497–505, 2017.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and removing disparate impact. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 259–268, 2015.

Föllmer, H. and Schied, A. *Stochastic finance: an introduction in discrete time*. Walter de Gruyter, 2011.

Foster, J. E. Inequality measurement. In Young, H. P. (ed.), *Fair Allocation*, pp. 31–68. American Mathematical Society, 1985.

Fukuchi, K., Sakuma, J., and Kamishima, T. Prediction with model-based neutrality. In *European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 499–514, 2013.

Gajdos, T. and Weymark, J. A. Introduction to inequality and risk. *Journal of Economic Theory*, 147:1313–1330, 2012.

Ghassami, A., Khodadadian, S., and Kiyavash, N. Fairness in supervised learning: An information theoretic approach. *CoRR*, abs/1801.04378, 2018. URL `http://arxiv.org/abs/1801.04378`.

Gotoh, J.-y. and Takeda, A. A linear classification model based on conditional geometric score. *Pacific Journal of Optimization*, 1:277–296, 2005.

Gotoh, J.-y., Kim, M. J., and Lim, A. E. Robust empirical optimization is almost the same as mean-variance optimization. *Operations Research Letters*, 46(4):448 – 452, 2018.

Grechuk, B. and Zabarankin, M. Schur convex functionals: Fatou property and representation. *Mathematical Finance: An International Journal of Mathematics, Statistics and Financial Economics*, 22(2):411–418, 2012.

Greselin, F. and Zitikis, R. Measuring economic inequality and risk: a unifying approach based on personal gambles, societal preferences and references. Technical report, arXiv:1508.00127v1, 2015.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NIPS)*, December 2016.

Harsanyi, J. C. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *The Journal of Political Economy*, 63(4):309–321, 1955.

Harsanyi, J. C. Ethics in terms of hypothetical imperatives. *Mind*, 67(267):305–316, July 1958.

Harsanyi, J. C. 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.

Harsanyi, J. C. *Rational Behaviour and Bargaining Equilibrium in Games and Social Situations*. Cambridge University Press, 1977.

Harsanyi, J. C. Bayesian decision theory and utilitarian ethics. *The American Economic Review*, 68(2):223–228, 1978.

Hashimoto, T. B., Srivastava, M., Namkoong, H., and Liang, P. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, 2018.

Heidari, H., Ferrari, C., Gummadi, K. P., and Krause, A. Fairness behind a veil of ignorance: A welfare analysis for automated decision making. In *Advances in Neural Information Processing Systems 31*, pp. 1273–1283, 2018.

Heidari, H., Loi, M., Gummadi, K. P., and Krause, A. A moral framework for understanding of fair ML through economic models of equality of opportunity. In *ACM Conference on Fairness, Accountability, and Transparency*, January 2019.

Hodgson, B. *Economics as moral science*. Springer Science & Business Media, 2001.

Johndrow, J. E. and Lum, K. An algorithm for removing sensitive information: application to race-independent recidivism prediction. *arXiv e-prints*, art. arXiv:1703.04957, March 2017.

Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. Fairness-aware classifier with prejudice remover regularizer. In *European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 35–50, 2012.

Kampelmann, S. Inequality measures as conventions: new interpretations of a classic operationalization problem. *Socio-Economic Review*, 7:669–694, 2009.

Kolm, S.-C. Unequal inequalities I. *Journal of Economic Theory*, 12:416–442, 1976a.

Kolm, S.-C. Unequal inequalities II. *Journal of Economic Theory*, 13:82–111, 1976b.

Komiyama, J. and Shimao, H. Two-stage Algorithm for Fairness-aware Machine Learning. *ArXiv e-prints*, October 2017.

Krokhmal, P., Zabarankin, M., and Uryasev, S. Modeling and optimization of risk. *Surveys in Operations Research and Management Science*, 16:49–66, 2011.

Kusner, M. J., Loftus, J. R., Russell, C., and Silva, R. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pp. 4069–4079, 2017.

Le Cam, L. *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, 1986.

Loewenstein, G. That which makes life worthwhile. In Krueger, A. B. (ed.), *Measuring the Subjective Well-Being of Nations: National Accounts of Time Use and Well-Being*, chapter 2. University of Chicago Press, 2009.

Marshall, A. W., Olkin, I., and Arnold, B. C. *Inequalities: Theory of Majorization and its Applications (2nd Edition)*. Springer, 2011.

Maurer, A. and Pontil, M. Empirical bernstein bounds and sample-variance penalization. In *COLT*, 2009.

Mazur, B. When is one thing equal to some other thing? In Gold, B. and Simons, R. A. (eds.), *Proof and Other Dilemmas: Mathematics and Philosophy*, chapter 11, pp. 221–241. The Mathematical Association of America, 2008.

McNamara, D., Ong, C. S., and Williamson, R. C. Costs and benefits of fair representation learning. In *AAAI Conference on Artificial Intelligence, Ethics and Society*, 2019.

Menon, A. K. and Williamson, R. C. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*, pp. 107–118, 2018.

Menon, A. K., Narasimhan, H., Agarwal, S., and Chawla, S. On the statistical consistency of algorithms for binary classification under class imbalance. In *International Conference on Machine Learning (ICML)*, pp. 603–611, 2013.

Mirowski, P. *More heat than light: Economics as social physics: Physics as nature's economics*. Cambridge University Press, 1989.

Mirowski, P. Do economists suffer from physics envy? *Finnish Economic Papers*, 5(1):61–68, 1992.

Mohri, M., Sivek, G., and Suresh, A. T. Agnostic federated learning. In *International Conference on Machine Learning*, 2019.

Núñez-Velázquez, J. J. Inequality measures, Lorenz curves and generating functions. In Pleguezuelo, R. H., Céspedes, J. C., and Velasco, J. M. H. (eds.), *Distribution Models Theory*, pp. 189–219. World Scientific, 2006.

Ogryczak, W. and Tamir, A. Minimizing the sum of the k largest functions in linear time. *Information Processing Letters*, 85(3):117 – 122, 2003.

Olfat, M. and Aswani, A. Spectral algorithms for computing fair support vector machines. In *International Conference on Artificial Intelligence and Statistics*, pp. 1933–1942, 2018.

Pedreshi, D., Ruggieri, S., and Turini, F. Discrimination-aware data mining. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 560–568, 2008.

Pérez-Suay, A., Laparra, V., Mateo-García, G., Muñoz-Marí, J., Gómez-Chova, L., and Camps-Valls, G. Fair kernel learning. In *Machine Learning and Knowledge Discovery in Databases*, pp. 339–355, 2017.

Pflug, G. C. and Romisch, W. *Modeling, measuring and managing risk*. World Scientific, 2007.

Rawls, J. *A Theory of Justice*. Harvard University Press, 1971.

Ritov, Y., Sun, Y., and Zhao, R. On conditional parity as a notion of non-discrimination in machine learning. *ArXiv e-prints*, June 2017.

Rockafellar, R. and Uryasev, S. Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, 26(7):1443–1471, 2002.

Rockafellar, R. T. Coherent approaches to risk in optimization under uncertainty. In *OR Tools and Applications: Glimpses of Future Technologies*, chapter 3, pp. 38–61. 2007.

Rockafellar, R. T. and Uryasev, S. Optimization of conditional value-at-risk. *Journal of Risk*, 2:21–41, 2000.

Rockafellar, R. T. and Uryasev, S. The fundamental risk quadrangle in risk management, optimization and statistical estimation. *Surveys in Operations Research and Management Science*, 18(1-2):33–53, 2013.

Rockafellar, R. T., Uryasev, S., and Zabarankin, M. Generalized deviations in risk analysis. *Finance and Stochastics*, 10(1):51–74, 2006.

Rona, P. and Zsolnai, L. (eds.). *Economics as moral science*. Springer, 2017.

Schölkopf, B., Smola, A. J., Williamson, R. C., and Bartlett, P. L. New support vector algorithms. *Neural Computation*, 12(5):1207–1245, May 2000.

Sen, A. *On Ethics and Economics*. Blackwell, 1987.

Sen, A. *Inequality Reexamined*. Oxford University Press, 1992.

Sen, A. *On Economic Inequality (Enlarged Edition)*. Oxford University Press, 1997.

Shiller, R. J. and Shiller, V. M. Economists as worldly philosophers. *American Economic Review*, 101(3):171–75, 2011.

Shorrocks, A. F. The class of additively decomposable inequality measures. *Econometrica*, 48(3):613–625, April 1980.

Shorrocks, A. F. Inequality decomposition by population subgroups. *Econometrica*, 52(6):1369–1385, 1984.

Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K. P., Singla, A., Weller, A., and Zafar, M. B. 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*, pp. 2239–2248, 2018.

Takeda, A. and Sugiyama, M. *nu*-support vector machine as conditional value-at-risk minimization. In *Proceedings of the Twenty-Fifth International Conference on Machine Learning*, pp. 1056–1063, 2008.

Traub, S., Siedl, C., Schmidt, U., and Levati, M. V. Friedman, Harsanyi, Rawls, Boulding — or somebody else? an experimental investigation of distributive justice. *Social Choice and Welfare*, 24:283–309, 2005.

Tsyurmasto, P., Zabarankin, M., and Uryasev, S. Value-at-risk support vector machine: stability to outliers. *Journal of Combinatorial Optimization*, 28(1):218–232, Jul 2014.

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., and Gummadi, K. Learning fair classifiers. *arXiv preprint arXiv:1507.05259*, 2016.

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., and Gummadi, K. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *International World Wide Web Conference*, 2017a.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., Gummadi, K., and Weller, A. From parity to preference-based notions of fairness in classification. In *Advances in Neural Information Processing Systems 30*, pp. 229–239, 2017b.

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., and Gummadi, K. P. Fairness constraints: Mechanisms for fair classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pp. 962–970, 2017c.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. Learning fair representations. In *International Conference on Machine Learning (ICML)*, 2013.

Zhang, J. and Bareinboim, E. Equality of opportunity in classification: A causal approach. In *Advances in Neural Information Processing Systems 31*, pp. 3675–3685, 2018.

Žliobaitė, I. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, 31 (4):1060–1089, Jul 2017.