Supplementary material for "Fairness risk measures"

A. Justification of fairness risk measure axioms

We now argue why each of the axioms introduced in §4 is natural when $\mathcal{R}$ is used per (14) to ensure fairness across subgroups. We note that apart from (F2), none of these properties can be relaxed without causing problems.

Convexity (F1) is desirable because without it, the risk could be decreased by more fine grained partitioning as we now show. F1 and F2 are equivalent to $\mathcal{R}$ being sub-additive and positive homogeneous (Rockafellar & Uryasev, 2013). Suppose $S = \{0, 1\}$ and the sensitive feature is determinate and thus 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 subadditive and we would have $\mathcal{R}(\mathsf{L}^0 + \mathsf{L}^1) = \mathcal{R}(\mathsf{L}) > \mathcal{R}(\mathsf{L}^0) + \mathcal{R}(\mathsf{L}^1)$. In other words, by splitting into subgroups we could automatically make our risk measure smaller, which is counter to what we wish to achieve. Convexity is also desirable because, combined with F3, it preserves tractability of optimisation.

Positive Homogeneity (F2) is desirable but not essential. We would like our fairness measure to not vary in a manner that changes the optimal f when ℓ varies in a manner that leaves the base problem invariant. For example, if $\ell' = c \cdot \ell$ for some $c > 0$ then obviously $\operatorname{argmin}_{f \in \mathcal{F}} L_\ell(f) = \operatorname{argmin}_{f \in \mathcal{F}} L_{\ell'}(f)$. If $\mathcal{R}$ is positively homogeneous, then $\mathcal{R}(\mathsf{L}') = \mathcal{R}(c \cdot \mathsf{L}) = c \cdot \mathcal{R}(\mathsf{L})$ and thus $\operatorname{argmin}_{f \in \mathcal{F}} \mathcal{R}(\mathsf{L}) = \operatorname{argmin}_{f \in \mathcal{F}} \mathcal{R}(\mathsf{L}')$. Observe this last statement would remain true if $\mathcal{R}$ was k-homogeneous, for any $k \geq 0$. Whether a relaxation of (F2) adds any practical advantage is not yet understood. Positive homogeneity does imply that the units of measurement of $\mathcal{R}(\mathsf{Z})$ are automatically the same as those for Z. The assumption of 1-homogeneity is also beneficial when analysing duality properties of risk measures; see Rockafellar & Uryasev (2013, Section 6).

Monotonicity (F3) is desirable because when combined with convexity (F1) it ensures that if $f \mapsto L(f)$ is convex, then so will be $f \mapsto \mathcal{R}(\mathsf{L}(f))$; see (Rockafellar & Uryasev, 2013, Section 5), and part 3 of Theorem 14. It is also intuitive that one's overall risk not increase if all subgroup risks are decreased. We note that a similar monotonicity assumption, and its implications, were also employed in Dwork et al. (2018, Section 4).

Lower Semicontinuity (F4) is a technical assumption that avoids problems with limits (Rockafellar & Uryasev, 2013).

Translation invariance (F5) is desirable because if we replace ℓ by $\ell + C$ we have not changed the unfairness at all, just the expected risk value.

Aversity (F6) means that deviation from perfect fairness (Definition 1) is penalised; without this property we would not be capturing deviation from ideal fairness.

Law Invariance (F7) means that $\mathcal{R}$ only depends upon Z via its *distribution* $\mathbb{P}_\mathsf{Z}$ through an induced functional $\mathcal{F}_\mathcal{P} \colon \mathcal{P}(S) \to \mathbb{R}$. For a fairness measure this would mean that the identity of each of the values of the sensitive feature do not matter, only the distribution of the risk variable $\mathsf{L}(s)$ as a function of $s \in S$. This is clearly a desirable attribute for a fairness measure.

B. Relationship to fairness methods based on inequality measures

An approach to fair machine learning similar to that based on fairness risk measures proposed in the body of the paper has been developed by Speicher et al. (2018); Heidari et al. (2018; 2019), based instead on the notion of an *inequality* measure. It turns out that this approach is closely related to that proposed in the main body of the paper. In this technical appendix, we explore the relationship between the two methods by working out the relationship between inequality measures and fairness risk measures.

The appendix is independent of the main paper in the sense that the paper can be understood entirely without looking at this appendix. It is included for completeness, and because we believe the results may be of independent interest.

The literature on inequality measures is quite formal (axiomatic) and consequently this appendix is too. The conclusion we draw (§B.5) is essentially that the requirements on a fairness risk measure are more stringent than those usually imposed upon inequality measures: every fairness risk measure induces a "nice" inequality measure, but it is not the case that every nice inequality measure induces a fairness risk measure. The additional constraints on fairness risk measures (convexity, continuity and monotonicity) are exactly what one wants for the sake of solving fair machine learning problems, since they guarantee computationally easy optimization problems.

B.1. Inequality Measures

We will show that the two approaches to solving fair ML problems, based respectively on inequality measures and risk measures, are intimately related by demonstrating that every fairness risk measure induces an inequality measure compatible with the Lorenz ordering (defined below). A weaker converse result holds which demonstrates that the requirements on a fairness risk measure are more stringent than those traditionally imposed upon an inequality measure. This result may be of interest in its own right since the literatures on risk measures and inequality measures appear to have not been explicitly connected before in this manner. There are a four specific exceptions of which we are aware:

- Bennett & Zitikis (2015) showed that some existing inequality measures can be derived from a model of choice under risk inspired by Harsanyi, especially his observation (Harsanyi, 1975) that Rawls' maximin principle corresponds "to an expected utility evaluation based on a lottery that assigns probability 1 to the event that one assumes the identity of the worst-off individual in society."
- Greselin & Zitikis (2015) unified a range of inequality measures and risk measures via a similar device — by choosing what they call "societal references" (e.g. a statistic of a population distribution such as a measure of centrality, or a tail measure) combined with distributions of personal gambles (that determine an individual's position on a population-based function). They make explicit connections with aspects of coherent risk measures.
- Gajdos & Weymark (2012) made a connection between random variables being "less risky" (meaning derivable from comparator variables by adding zero mean noise), and measures of inequality. However, they did not draw connections to the notion of *risk measures*.
- It has also been shown that the mathematical notion underlying the Pigou-Dalton transfer principle for inequality measures (Schur convexity) is known to reduce to law invariance for coherent risk measures on atomless probability spaces (Dana, 2005; Grechuk & Zabarankin, 2012), but these authors did not proceed to develop the more detailed connections we develop below.

None of these works have systematically related the *axioms* for inequality measures to the axioms for risk as we do in this appendix, nor do they propose the formulaic correspondence between risk measures and inequality measures that we do (see (33)–(36)), which however is implicit in the work of Kolm (1976a;b).

Inequality measures[7] have been investigated by "welfare" economists, interested in such issues as the distribution of wealth or income. Their perspective is intrinsically moral, rather than the more traditional view of economics as aspirational physics (Mirowski, 1989; 1992), and such an engagement with the moral dimensions has been forcefully advocated recently (Shiller & Shiller, 2011). Formal measures of economic inequality were made famous by Tony Atkinson's 1970 widely cited paper (Atkinson, 1970). Late in his career Atkinson (2009)[8] bemoaned the decline in explicit discussion of welfare economics from its heyday in the 1960s. Atkinson stressed the need for *plurality* — that there is no single criteria that captures welfare:

[7] Sometimes called "measures of inequality" or "inequality indices".

[8] The title of Atkinson's paper is "economics as a moral science". There are at least four other works with that exact title, whose

> Many of the ambiguities and disagreements [in economics] stem not from differences of view about how the economy works but about the criteria to be applied when making judgments. ... People can legitimately reach different conclusions because they apply different theories of justice. (Atkinson, 2009, page 803).

The rest of the appendix is organised as follows. In Section B.2 we introduce the main axioms used in studying inequality measures; in §B.3 we present a number of apparently new results formally relating the axioms from the previous subsection with those for fairness measures stated in the main body of the paper; in §B.4 we pull the various lemmas together and state our main result; and finally in §B.5 we make some brief general conclusions regarding the consequences for the use of either risk or inequality measures to solve the problems of fair machine learning.

B.2. Axiomatizing Inequality Measures

The idea of an inequality measure $\mathcal{I}$ is to measure the degree of inequality of a population, say, in terms of incomes; that is some measure of variability, unevenness or non-uniformity. In contrast to risk measures, which can work sensibly on continuous probability spaces, inequality measures are usually only defined for populations of individuals of some finite size $n \in \mathbb{N}$. Let $x \in \mathbb{R}^n$, and represent by x_i the income of the ith individual. Without loss of generality we assume incomes are nonnegative, and thus an *inequality measure* is a function $\mathcal{I} \colon \mathbb{R}^n_{\geq 0} \to \mathbb{R}_{\geq 0}$. It is often assumed that the ordering of the individuals is such that $i < j \Rightarrow x_i \leq x_j$. This assumption will be relaxed later by imposing a symmetry condition on $\mathcal{I}$.

Merely requiring $\mathcal{I}$ to be a function $\mathbb{R}^n_{\geq 0} \to \mathbb{R}_{\geq 0}$ is hardly illuminating or satisfactory. In order to capture what is intuitively understood by "inequality" a range of conditions are additionally imposed. Depending upon which conditions are used, this leads to a large variety of measures of inequality, some of which are surveyed by Cowell (2000) and Sen (1997).

In order to introduce these conditions on $\mathcal{I}$ and to relate inequality measures to fairness measures, we introduce some notation and terminology due to Marshall et al. (2011). Suppose $x \in \mathbb{R}^n$. We write $x_{[i]}$ for the ith component of the *decreasing rearrangment* of x which satisfies $i < j \Rightarrow x_{[i]} \geq x_{[j]}$. For $x, y \in \mathbb{R}^n$, we say x is *majorized* by y on A and write $x \succ_{\text{M}} y$ on A if $x, y \in A$ and

$$\sum_{i=1}^{k} x_{[i]} \leq \sum_{i=1}^{k} y_{[i]}, \forall k \in [n-1] \quad \text{and} \quad \sum_{i=1}^{n} x_{[i]} = \sum_{i=1}^{n} y_{[i]}.$$

Suppose $A \subset \mathbb{R}^n$. A function $\phi \colon A \to \mathbb{R}$ is *Schur-convex on* A if

$$x \succ_{\text{M}} y \text{ on } A \;\; \Rightarrow \;\; \phi(x) \leq \phi(y).$$

If, furthermore, $\phi(x) < \phi(y)$ whenever $x \succ_{\text{M}} y$ but y is not a permutation of x, then ϕ is *strictly Schur-convex on* A. If $A = \mathbb{R}^n$ we simply say *Schur-convex* or *strictly Schur-convex*. Let $\mathcal{C}(X)$ denote the set of continuous convex functions on

pertinent common message we distill via the following four brief quotations:

- "Adam Smith, who has strong claim to being both the Adam and the Smith of systematic economics, was a professor of moral philosophy and it was at that forge that economics was made. Even when I was a student, economics was still part of the moral sciences tripos at Cambridge University. It can claim to be a moral science, therefore, from its origin, if for no other reason. Nevertheless, for many economists the very term 'moral science' will seem like a contradiction" (Boulding, 1969).
- "Economics is, and always has been, essentially a moral science whatever the protestations to the contrary by some of its practitioners" (Cochran, 1974).
- "[E]conomic science must merge with moral science" (Hodgson, 2001, page 309).
- "Morality is not something we add to a model constructed without it; one built without morality as its integral component is an unsound structure. Restoring economics to its former habitat as a moral science, a science of practical knowledge, cannot be evaded" (Rona & Zsolnai, 2017, page 9).

Amartya Sen, in his *On Ethics and Economics*, argued that "economics has had two rather different origins, both related to politics, but related in rather different ways, concerned respectively with 'ethics', on the one hand, and with what may be called 'engineering', on the other" (Sen, 1987, pages 2–3). Sen uses "ethics" to mean choices of values and goals, and "engineering" as mere means to achieve them. Ironically there has been a steady growth in soul searching in the discipline of engineering itself asking these same questions. Most recently, the engineering field of machine learning is waking up to the same realisation; perhaps soon we shall see papers entitled *Machine Learning as Moral Science*!

X. We have (Marshall et al., 2011, page 14) that

$$x \succ_{\mathrm{M}} y \;\Leftrightarrow\; \sum_{i=1}^{n} \phi(x_i) \leq \sum_{i=1}^{n} \phi(y_i) \;\; \forall \phi \in \mathcal{C}(\mathbb{R}^n).$$

Every convex symmetric function is Schur convex. A special case of this are *convex seperable* functions of the form $\phi(x) = \sum_{i=1}^{n} g(x_i)$, where $g\colon \mathbb{R} \to \mathbb{R}$ is convex and continuous. But not all Schur-convex functions are convex seperable. The significance of Schur-convexity for our current purpose is that the Pigou-Dalton condition on an inequality measure $\mathcal{I}$ is equivalent to requiring that $\mathcal{I}$ is strictly Schur-convex (Marshall et al., 2011, page 560).

A widely used notion in the theory of economic inequality is the Lorenz curve. We associate with an income vector $x \in \mathbb{R}^n_{\geq 0}$ a probability distribution μ_x defined via

$$\mu_x(\{x_i\}) := \tfrac{1}{n}|\{j \in [n]\colon x_j = x_i\}|.$$

Let F_x denote the corresponding cumulative distribution function: $F_x(t) := \frac{1}{n}|i\colon x_i \leq t|$ and F_x^{-1} its quasi inverse $F_x^{-1}(p) := \inf\{x\colon F(x) \geq p\}$. The *Lorenz curve* for x is then defined as $\{(p, L_x(p)\colon p \in [0,1]\}$; i.e. the graph of the function

$$L_x(p) := \frac{1}{m_x} \int_0^p F^{-1}(t)\mathrm{d}t \quad p \in [0,1],$$

where $m_x = \int_{-\infty}^{\infty} t\mathrm{d}F_x(t)$ is the corresponding mean. Let $\mathbb{R}^{\uparrow n} := \{(x_1, \ldots, x_n) \in \mathbb{R}^n\colon x_1 \leq x_2 \leq \cdots \leq x_n\}$. The Lorenz curve induces a partial order on $\mathbb{R}^{\uparrow n}$ via pointwise domination of the corresponding functions: $x \preccurlyeq_{\mathrm{L}} y \Leftrightarrow L_x \geq L_y$. This is called the *Lorenz ordering*.

We can interpret $x \in \mathbb{R}^n$ as a map $x\colon [n] \to \mathbb{R}$ and thus for a permutation π, we write $x \circ \pi = (x_{\pi(1)}, \ldots, x_{\pi(n)})$. We denote the Kronecker product by $\otimes$ and $1_r \in \mathbb{R}^r$ the all ones vector of length r. Let $x \in \mathbb{R}^n$, $r \in \mathbb{N}$, and $x^r := 1_r \otimes x$. If S is a set and $J \in \mathbb{N}$, then $S_1, \ldots, S_J$ is called a *partition* of S if $S_i \cap S_j = \emptyset$ for $i, j \in [J]$, $i \neq j$, and $S = \bigcup_{j \in [J]} S_j$.

In order to make sense of the zoo of potential inequality measures, a range of conditions (often called "axioms") are imposed (Núñez-Velázquez, 2006). We first state the axioms formally, and then discuss them, providing the intuition and justification behind them[9]. In the inequality measure literature, often the domain is consistently taken to be $\mathbb{R}^{\uparrow n}$ but since we will only consider measures that satisfy symmetry, apart from axiom I1 we presume $\mathcal{I}$ is defined on $\mathbb{R}^n$.

I1 Symmetry For any permutation $\pi\colon [n] \to [n]$, $\mathcal{I}(x \circ \pi) = \mathcal{I}(x)$ for all $x \in \mathbb{R}^{\uparrow n}$.

I2 Scale Invariance For any $\lambda > 0$, $\mathcal{I}(\lambda x) = \mathcal{I}(x)$ for all $x \in \mathbb{R}^n$.

I3 Pigou-Dalton Principle $\mathcal{I}$ is strictly Schur-convex on $\mathbb{R}^n$.

I4 Dalton Population Principle $\mathcal{I}(x^{(r)}) = \mathcal{I}(x)$ for all $x \in \mathbb{R}^n$ and $r \in \mathbb{N}$.

I5 Normalization $\mathcal{I}(x) \geq 0$ for all $x \in \mathbb{R}^n$ and $\mathcal{I}(x) = 0 \Leftrightarrow x = c1_n$ for some $c \in \mathbb{R}$.

I6 Constant Addition $\mathcal{I}(x + c1_n) \leq \mathcal{I}(x)$ for all $x \in \mathbb{R}^n$ and $c > 0$.

I7 Lorenz Compatibility $\mathcal{I}(x) \leq \mathcal{I}(y) \Leftrightarrow x \preccurlyeq_{\mathrm{L}} y$.

I8 Additive Decomposition Let $J \in \mathbb{N} \setminus \{1\}$, $S_1, \ldots, S_J$ be a partition of $[n]$, and for $j \in [J]$, let $n_j := |S_j|$ and $x^{(j)} := (x_i)_{i \in S_j}$. An inequality measure $\mathcal{I}$ satisfies *additive decomposition* if

$$(\forall x \in \mathbb{R}^{\uparrow n}) \;\; \mathcal{I}(x) = \sum_{j \in [J]} w_j \mathcal{I}(x^{(j)}) + B,$$

where w_j and B depend only on the subgroup sizes n_j and means $m_j := \frac{1}{n_j} \sum_{i \in [n_j]} x_i^{(j)}$.

[9] We make no claim that the list of axioms presented here is complete. But they do appear to be the most significant and widely used ones

I9 Subgroup Consistency Let S_1, S_2 be a partition of $[n]$. An inequality measure $\mathcal{I}$ satisfies *subgroup consistency* if there exists an aggregation function Φ increasing in its first two arguments, such that

$$(\forall x \in \mathbb{R}^{\uparrow n}) \quad \mathcal{I}(x) = \Phi\left(\mathcal{I}(x^{(1)}), \mathcal{I}(x^{(2)}); m_1, m_2; n_1, n_2\right).$$

I10 Seperable We say $\mathcal{I}$ is *seperable* if it can be written as $\mathcal{I}(x) = \sum_{i=1}^{n} g(x_i)$, where $g\colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$.

I11 Constant Sum Convexity For $c > 0$ $T_c = \{x \in \mathbb{R}^n\colon \sum_{i\in[n]} x_i = c\}$. Then $\mathcal{I}$ is *constant sum convex* (resp. *constant sum strictly convex*) if for any $c > 0$ it is convex (resp. strictly convex) on T_c.

We make some comments regarding each of these axioms.

I1 Symmetry captures the notion that the identity of individuals should not change the perceived degree of inequality.

I2 Scale invariance is also known as 0-homogeneity or homotheticity. It captures the idea that if all incomes (say) are multiplied by $\lambda > 0$ then whilst the welfare and well-being of individuals would change, the inequality across the population would not.

I3 Pigou-Dalton Principle is usually described in terms of the "principle of transfers," whereby if income is transferred from a rich to a poor person in a "mean-preserving" manner, then inequality can noot go up; see (Marshall et al., 2011, page 6) or (Cowell, 2000, page 98) for an exposition. For our purposes, the statement in terms of strict Schur convexity is more convenient.

I4 Dalton Population Principle is motivated by the idea that if a finite population is replicated r times (for each individual with given wealth, for example, r copies are created with the same wealth) then the inequality measure should not change. This principle, as stated, hides a significant subtelty. Implicit in its statement is that we actually have a *set* $\{\mathcal{I}^n\}_{n\in\mathbb{N}}$ of inequality measures, where $\mathcal{I}^n\colon \mathbb{R}^n_{\geq 0} \to \mathbb{R}_{\geq 0}$. The relationship between $\mathcal{I}^n$ and $\mathcal{I}^m$ for $n \neq m$ obviously needs to be specified. Usually this is done by writing a functional form (i.e. a symbolic formula) that holds for all $n \in \mathbb{N}$. Of course if, for a fixed $n^* \in \mathbb{N}$, $\mathcal{I}^{n^*}$ is specified, then there is no unique way to induce the full set $\{\mathcal{I}^n\}_{n\in\mathbb{N}}$; confer (Chakravarty, 1999, page 165). In some of the literature, this point is glossed over (e.g. (Núñez-Velázquez, 2006, page 205)).

I5 Normalization guarantees the inequality measure is non-negative, and only equals zero (no inequality) when all x_i ($i \in [n]$) are identical.

I6 Constant Addition captures the idea that a constant positive addition to everyone's income does not increase inequality.

I7 Lorenz Compatibility is of interest because of the historically wide use of the Lorenz curve and the induced Lorenz ordering; the axiom requires that the partial ordering induced by an inequality measure is the same as that induced by pointwise domination of the corresponding Lorenz curves.

I8 Additive Decomposition is motivated by the idea that the inequality of an entire population must be related to the inequalties of its constituent sub-populations. The condition is very strong and, in contrast say to I6, is stated as an *equality* rather than as an *inequality*. Shorrocks (1980) has shown that when combined with I2 and I4, the only admissible inequality measures are those in a single parameter generalised entropy family.

I9 Subgroup consistency introduced in (Shorrocks, 1984), is a significant weakening of I8 (I8 implies I9). Rather than insisting on an additive decomposition across subgroups, it is merely required that the inequalities of subgroups be aggregatable to recover the inequality of the whole population. Surprisingly, there is a sort of converse whereby an inequality measure satisfying I9 can be transformed via a monotonic transformation to one satisfying I8. Axiom I9 can be applied recursively to hold for partitions of arbitrary size. The inequality in the definition of I9 excludes inequality measures with the property that inequality in every subgroup rises, but overall inequality falls. Properties of inequality measures from the perspective of subgroup decomposition are explored in depth by Deutsch & Silber (1999).

I10 Seperability Is a special case of additive decomposition, and is an oft imposed condition on Social Welfare functions from which inequality measures can be derived (Cowell, 2000).

I11 Constant Sum Convexity Does not appear to be widely assumed in the inequality measure literature, with the exception of (Kolm, 1976b, Section IX) who provides an extensive justification of it and related notions of convexity. Constant sum convexity is arguably a better assumption than I8, I9, and I10 for three reasons. First, it allows a neater connection with risk measures. Second, it is mathematically more elegant. The third reason is perhaps more interesting. Axioms I8 and I9 are stated as *equalities* and this allows the application of the theory of functional equations in analysing the implications. Capturing the intent of I9 (one can not make inequality look better by dividing into subgroups) via an *inequality* precludes analysis via the theory of functional equations, but actually makes the arguments easier via the theory of convexity. Kolm (1976a, page 423) also proposed subadditivity as an axiom. If $\mathcal{I}$ is assumed 1-homogeneous, this is equivalent to convexity (recall [subadditivity and positive homogeneity] implies [convexity and positive homogeneity]).

The motivation of scholars of inequality in adopting the axiomatic approach varies. Many use it as a device to whittle down the large number of possibilites to a few or even one single inequality measure. But the arguments regarding the social desirability of the various axioms are not so clear cut. Kolm (1976a;b) in particular has argued at length against a simple-minded acceptance of any of these axioms; confer (Cowell, 2011). Kampelmann (2009) has observed that the appeal of such analyses is sometimes that of a pursuit of objectivity in making a choice that is seemingly "opinion-free". More fundamentally, any notion of *inequality* is predicated on some notion of *equality*, which turns out to be more subtle than one might think. Even in the realm of pure mathematics it is a subtle concept, which Mazur (2008) has argued is better construed as *equivalence up to a canonical isomorphism*. One could even argue that the proposals of Sen (1992) to take *capabilities* into account is partially grappling with this purely logical issue: taking income equality as "equality" implies quotienting out everything else! (Confer (Loewenstein, 2009) for a complementary argument against the notion that simple scalar measures can capture what is important.).

Fortunately, we can sidestep these weighty questions because our goal is not to prescribe which are suitable for the social and economics purposes to which they are typically applied, but rather to simply explore the relationship with risk measures in order that we can sensibly compare fair machine learning approaches built upon inequality measures with those built upon risk measures.

B.3. Relating properties of risk measures and properties of inequality measures

The theory of risk measures and the theory of inequality measures have been developed independently. We will now show that there is a straight-forward connection between them, which, up to some minor qualifications, provides a 1:1 correspondence between a fairness risk measure $\mathcal{R}$ (or equivalently a symmetric deviation measure $\mathcal{D}$) and an inequality measure $\mathcal{I}$.

In order to achieve this goal we first we need to deal with the different type signatures: $\mathcal{R}$ and $\mathcal{D}$ take a random variable $\mathsf{X}\colon S \to \mathbb{R}_{\geq 0}$ as an argument; inequality measures $\mathcal{I}$ take a vector $x \in \mathbb{R}^{\uparrow n}_{\geq 0}$ or $\mathbb{R}^n_{\geq 0}$ as an argument. We demand that $\mathcal{I}$, $\mathcal{R}$ and $\mathcal{D}$ all be symmetric which means henceforth we can take the domain of $\mathcal{I}$ as $\overline{\mathbb{R}}^n_{\geq 0}$. Risk measures can be defined on an arbitrary sample space S, but to construct the connection with inequality measures we set $S = [n]$.

Risk and deviation measures are defined on a space of random variables, and there needs to be some base measure on the sample space S for this to make sense. While as mentioned in the main body of the paper, the freedom to choose different ν is a feature of the framework we propose, we will restrict ourselves to $\nu = \nu^n_{\mathrm{u}}$, the uniform probability measure on $[n]$: for $B \subseteq [n]$, $\nu^n_{\mathrm{u}}(B) = |B|/n$. There are two reasons for this. First, it immediately guarantees symmetry of the induced inequality measures. Second, one can extend the argument due to Harsanyi (1955; 1958; 1975; 1978) to justify the choice of a uniform distribution as being morally warranted (when S corresponds to sensitive features) a uniform distribution corresponds to treating each value of the sensitive feature equally. In the limit where each person who is represented by data corresponds to an element of S, we precisely recover Harsanyi's setting and a uniform distribution corresponds to the moral principle of treating people without regard for their identity.

A vector $x \in \mathbb{R}^n_{\geq 0}$ can be identified with a a function $[n] \to \mathbb{R}_{\geq 0}$. We henceforth identify random variables $\mathsf{X}\colon [n] \to \mathbb{R}_{\geq 0}$ with vectors $x \in \overline{\mathbb{R}}^n_{\geq 0}$ via the identity

$$x_i = \mathsf{X}(i), \quad i \in [n].$$

Since our sample space S is now assumed finite, there are no measurability issues to concern us. We thus write $\mathcal{I}(x) = \mathcal{I}(\mathsf{X})$ as synonyms; likewise we write $\mathcal{R}(\mathsf{X}) = \mathcal{R}(x)$ and $\mathcal{D}(\mathsf{X}) = \mathcal{D}(x)$ as synonyms.

We know from the quadrangle theorem (Theorem 14) that regular risk measures $\mathcal{R}$ and regular deviation measures $\mathcal{D}$ are in

1:1 correspondence via the relationship

$$\mathcal{R}(\mathsf{X}) = \mathbb{E}(\mathsf{X}) + \mathcal{D}(\mathsf{X}). \tag{32}$$

Motivated by (32), we define an inequality measure $\mathcal{I}$ in terms of a given $\mathcal{R}$ or $\mathcal{D}$, and conversely define a risk measure $\mathcal{R}$ or deviation measure $\mathcal{D}$ in terms of a given inequality measure $\mathcal{I}$. Given a deviation measure $\mathcal{D}$, and a random variable $\mathsf{X}\colon S \to \mathbb{R}_{\geq 0}$, let

$$\mathcal{I}_{\mathcal{D}}(\mathsf{X}) := \begin{cases} \mathcal{D}(\mathsf{X})/\mathbb{E}(\mathsf{X}), & \text{if } \mathbb{E}(\mathsf{X}) \neq 0 \\ 0 & \text{if } \mathbb{E}(\mathsf{X}) = 0 \end{cases} \tag{33}$$

(Observe that since $\mathsf{X} \geq 0$, we have that $\mathbb{E}(\mathsf{X}) = 0 \Rightarrow \mathcal{D}(\mathsf{X}) = 0$ for any regular deviation measure, and thus the second case above actually follows from the first by considering $\lim_{c\downarrow 0} \mathcal{D}(\mathsf{X}_c)/\mathbb{E}(\mathsf{X}_c)$, where $\mathsf{X}_c := c\mathsf{X}$, and X is an arbitrary random variable such that $\mathsf{X} \geq 0$ and $\mathcal{D}(\mathsf{X}) > 0$.) Using (32), we can equivalently define

$$\mathcal{I}_{\mathcal{R}}(\mathsf{X}) := \frac{\mathcal{R}(\mathsf{X})}{\mathbb{E}(\mathsf{X})} - 1, \tag{34}$$

which is guaranteed to be well defined for any $\mathcal{R}$ satisfying F6. Conversely, given an inequality measure $\mathcal{I}$, we define

$$\mathcal{D}_{\mathcal{I}}(\mathsf{X}) := \mathbb{E}(\mathsf{X})\mathcal{I}(\mathsf{X}) \tag{35}$$

$$\mathcal{R}_{\mathcal{I}}(\mathsf{X}) := \mathbb{E}(\mathsf{X}) + \mathbb{E}(\mathsf{X})\mathcal{I}(\mathsf{X}) = \mathbb{E}(\mathsf{X})[1 + \mathcal{I}(\mathsf{X})]. \tag{36}$$

As an example, consider $\mathcal{D}(\mathsf{X}) = \sigma(\mathsf{X})$ (the standard deviation), in which case we get $\mathcal{I}_\sigma(X) = \sigma(\mathsf{X})/\mathbb{E}(\mathsf{X})$, which is known as the *coefficient of variation*.

We will now justify the above definitions by showing, under suitable assumptions on $\mathcal{I}$, that $\mathcal{D}_{\mathcal{I}}$ (resp. $\mathcal{R}_{\mathcal{I}}$) is a deviation (resp. risk) measure satisfying a subset of axioms F1—F8; and conversely, given suitable assumptions on $\mathcal{D}$ (resp. $\mathcal{R}$), that $\mathcal{I}_{\mathcal{D}}$ (resp. $\mathcal{I}_{\mathcal{R}}$) is an inequality measure satisfying a subset of axioms I1–I11.

B.3.1. I1 SYMMETRY

The symmetry constraints on fairness measures and inequality measures are easily related:

Lemma 17. *Suppose $S = [n]$ and $\nu = \nu_{\mathrm{u}}^n$.*

1. *If $\mathcal{I}$ is symmetric (I1) then $\mathcal{R}_{\mathcal{I}}$ is law invariant (F7).*
2. *If $\mathcal{R}$ is law-invariant (F7) then $\mathcal{I}_{\mathcal{R}}$ is symmetric (I1).*

For regular risk measures, law invariance of $\mathcal{R}$ implies law-invariance of $\mathcal{D}$ so an analogous results holds with $\mathcal{R}$ replaced by $\mathcal{D}$ throughout.

Proof. Let Π_n be the set of all permutations $\pi\colon [n] \to [n]$. Observe that for any $\pi \in \Pi_n$, $\nu_{\mathrm{u}}^n \circ \pi^{-1} = \nu_{\mathrm{u}}^n$. Given that $\mathcal{R}_{\mathcal{I}}(\mathsf{X}) = \mathbb{E}(\mathsf{X})[1 + \mathcal{I}(\mathsf{X})]$, we have $\mathcal{R}_{\mathcal{I}}$ is law-invariant (resp. symmetric) if and only if $\mathcal{I}$ is law-invariant (resp. symmetric) since $\mathbb{E}$ is law-invariant (resp. symmetric).

It convenient to rewrite the definitions of symmetry and law-invariance as follows.

$$\mathcal{I} \text{ is symmetric if } \mathcal{I}(\mathsf{L}) = \mathcal{I}(\tilde{\mathsf{L}}) \ \ \forall \mathsf{L},\ \forall \tilde{\mathsf{L}} \in \mathcal{L}_{\mathrm{sym}}(\mathsf{L}) \tag{37}$$

$$\mathcal{I} \text{ is law invariant if } \mathcal{I}(\mathsf{L}) = \mathcal{I}(\tilde{\mathsf{L}}) \ \ \forall \mathsf{L},\ \forall \tilde{\mathsf{L}} \in \mathcal{L}_{\mathrm{li}}(\mathsf{L}), \tag{38}$$

where $\mathcal{L}_{\mathrm{sym}}(\mathsf{L}) := \{\mathsf{L} \circ \pi\colon \pi \in \Pi_n\}$ and $\mathcal{L}_{\mathrm{li}}(\mathsf{L}) := \{\mathsf{L}\colon \mu_{\mathsf{L}} = \mu_{\tilde{\mathsf{L}}}\}$. Thus in order to prove the lemma it suffices to show that for any L, $\mathcal{L}_{\mathrm{sym}}(\mathsf{L}) = \mathcal{L}_{\mathrm{li}}(\mathsf{L})$ since then the conditions (37) and (38) are identical.

Suppose $\tilde{\mathsf{L}} = \mathsf{L} \circ \pi$ for some $\pi \in \Pi_n$. Then for any $A \in \mathcal{B}(\mathbb{R}_{\geq 0})$ (the Borel ssubsets of $\mathbb{R}_{\geq 0}$),

$$\begin{aligned} \mu_{\tilde{\mathsf{L}}}(A) &= (\nu_{\mathrm{u}}^n \circ \tilde{\mathsf{L}}^{-1})(A) \\ &= (\nu_{\mathrm{u}}^n \circ \pi^{-1} \circ \mathsf{L}^{-1})(A) \\ &= (\nu_{\mathrm{u}}^n \circ \mathsf{L}^{-1})(A) \\ &= \mu_{\mathsf{L}}(A). \end{aligned}$$

Thus $\mathcal{L}_{\mathrm{sym}}(\mathsf{L}) \subseteq \mathcal{L}_{\mathrm{li}}(\mathsf{L})$.

Suppose instead that $\mathsf{L}, \tilde{\mathsf{L}} \colon [n] \to \mathbb{R}_{\geq 0}$ are such that $\mu_{\mathsf{L}} = \mu_{\tilde{\mathsf{L}}}$. Then the image of S under L and $\tilde{\mathsf{L}}$ is identical: $\mathsf{L}(S) = \tilde{\mathsf{L}}(S)$. Let $\{t_1, \dots, t_k\} = \mathsf{L}(S)$ where $k \leq n$. For $i \in [k]$, $\mu_{\mathsf{L}}(\{t_i\}) = j_i/n$, where j_i is positive integer and $\sum_{i \in [k]} j_i = n$. Each elementary probability mass $1/n$ corresponds under L (resp. $\tilde{\mathsf{L}}$) to a particular $s_r \in S$ (resp. $s_{\tilde{r}} \in S$), $r, \tilde{r} \in [n]$. The only freedom in choosing $\tilde{\mathsf{L}}$ and L is in the indexing of elements of S. Thus one can write $\tilde{r} = \pi^{-1}(r)$ for some permutation π and thus $\tilde{\mathsf{L}} = \mathsf{L} \circ \pi$. Consequently $\mathcal{L}_{\mathrm{sym}}(\mathsf{L}) \supseteq \mathcal{L}_{\mathrm{li}}(\mathsf{L})$ which completes the proof. □

B.3.2. I2 Scale Invariance

The form of scaling behaviour of inequality measures and risk measures is also easily related:

Lemma 18. *If $\mathcal{D}$ (resp. $\mathcal{R}$) is positively homogeneous (i.e. satisfies F2) then $\mathcal{I}_{\mathcal{D}}$ (resp. $\mathcal{I}_{\mathcal{R}}$) is 0-homogeneous (I2); conversely, if $\mathcal{I}$ satisfies I2 then $\mathcal{D}_{\mathcal{I}}$ (resp. $\mathcal{R}_{\mathcal{I}}$) satisfies F2.*

Proof. Since $\mathbb{E}$ is positively homogeneous and the ratio of two positively homogeneous functions is 0-homogeneous, we obtain the first part. The second part follows since the product of a 1-homogeneous function with a 0-homogeneous function is 1-homogeneous. □

The fact that if $\mathcal{I}$ is 0-homogeneous then $\mathsf{X} \mapsto \mathbb{E}(\mathsf{X})\mathcal{I}(\mathsf{X})$ is 1-homogeneous was observed by Kolm (1976a, page 423).

B.3.3. I3 Pigou-Dalton (Schur-Convexity)

We need the following straightforward lemma.

Lemma 19. *Suppose $\phi \colon \mathbb{R}^n \to \mathbb{R}$. Let $\bar{x} := \frac{1}{n}\sum_{i=1}^n x_i$ and define $\tilde{\phi} \colon x \mapsto \phi(x)/\bar{x}$. Then $\tilde{\phi}$ is Schur-convex (resp. strictly Schur-convex) if and only if ϕ is Schur-convex (resp. strictly Schur-convex).*

Proof. Suppose $x, y \in \mathbb{R}^n$ and $x \succ_{\mathrm{M}} y$ which implies $\bar{x} = \bar{y} =: m$. Suppose $m > 0$ (we justify that this does not lose generality below). Let $\tilde{x} := x/m$ and $\tilde{y} = y/m$. Observe that $x \succ_{\mathrm{M}} y \Leftrightarrow \tilde{x} \succ_{\mathrm{M}} \tilde{y}$. We thus have

$$\begin{aligned}
\phi \text{ is Schur-convex} &\Leftrightarrow [x \succ_{\mathrm{M}} y \Leftrightarrow \phi(x) \leq \phi(y)] \\
&\Leftrightarrow [\tilde{x} \succ_{\mathrm{M}} \tilde{y} \Leftrightarrow \phi(\tilde{x}) \leq \phi(\tilde{y})] \\
&\Leftrightarrow [x \succ_{\mathrm{M}} y \Leftrightarrow \tilde{\phi}(x) \leq \tilde{\phi}(y)] \\
&\Leftrightarrow \tilde{\phi} \text{ is Schur-convex.}
\end{aligned}$$

Since [x is not a permutation of y] $\Leftrightarrow$ [$\tilde{x}$ is not a permutation of $\tilde{y}$] we similarly can conclude that strict Schur-convexity of ϕ is equivalent to strict Schur-convexity of $\tilde{\phi}$. □

Lemma 20. *Suppose $\mathcal{D}$ is convex (F1), then $\mathcal{I}_{\mathcal{D}}$ is strictly Schur-convex (I3).*

Proof. Combine Lemma 19 with the fact that every convex symmetric function is Schur-convex (Marshall et al., 2011). □

A similar result has been observed in a particular case that complements the situation of interest to us: a probability space S is called *atomless* is there exists a random variable on S with a continuous cumulative distribution function. Dana (2005) showed (confer (Grechuk & Zabarankin, 2012)) that on an atomless probability space, every law invariant risk functional $\mathcal{L}^p(S) \to \bar{\mathbb{R}}$ is Schur-convex. A converse to the above lemma is impossible because there are Schur-convex functions that are not convex (Marshall et al., 2011).

B.3.4. I4 Dalton Population Principle

Lemma 21. *If $\mathcal{R}$ (resp. $\mathcal{D}$) is law invariant (F7), then $\mathcal{I}_{\mathcal{R}}$ satisfies I4.*

Proof. It follows from the definition of $\mathcal{I}_{\mathcal{R}}$ that $\mathcal{I}_{\mathcal{R}}$ satisfies I4 if and only if $\mathcal{R}$ does (since expectation trivially does not change under replication). For $r \in \mathbb{N}$, given a random variable $\mathsf{L} \colon [n] \to \mathbb{R}_{\geq 0}$ define the r-replicated random variable $\mathsf{L}_{(r)} \colon [rn] \to \mathbb{R}_{\geq 0}$ via $\mathsf{L}_{(r)}(i) := \mathsf{L}(\psi_r(i))$, where $\psi_r \colon \mathbb{N} \to \mathbb{N}$ is given by $\psi_r(i) := \lceil i/r \rceil$. The inverse $\psi_r^{-1} \colon \mathbb{N} \rightrightarrows \mathbb{N}$ is

given by $\psi_r^{-1}(j) = \{s \in \mathbb{N}\colon \lceil s/r \rceil = j\} = \{rj, rj+1, \ldots, rj+(r-1)\}$. Thus the probability law of $\mathsf{L}_{(r)}$ is given by (for $A \in \mathcal{B}(\mathbb{R}_{\geq 0})$)

$$\begin{aligned}\mu_{\mathsf{L}_{(r)}}(A) &= (\nu_{\mathrm{u}}^{nr} \circ \mathsf{L}_{(r)}^{-1})(A) \\ &= (\nu_{\mathrm{u}}^{nr} \circ \psi_r^{-1} \circ \mathsf{L}^{-1})(A) \\ &= (\nu_{\mathrm{u}}^{n} \circ \mathsf{L}^{-1})(A) \\ &= \mu_{\mathsf{L}}(A),\end{aligned}$$

because $\nu_{\mathrm{u}}^{nr} \circ \psi_r^{-1} = \nu_{\mathrm{u}}^n$. Since $\mathcal{I}$ is law invariant, $\mathcal{I}_{\mathcal{R}}(\mathsf{L}_{(r)}) = \mathcal{I}_{\mathcal{R}}(\mathsf{L})$ and consequently $\mathcal{I}_{\mathcal{R}}$ satisfies the Dalton Population Principle I4. □

The Dalton Population Principle does not, on its own, imply law invariance. It can't, because if we start with a population of size n, it only makes assertions about populations of size nr for $r \in \mathbb{N}$. But given a distribution μ_{L}, there are many more sample spaces (not necessarily of cardinality nr) on which one can define random variables $\tilde{\mathsf{L}}$ that can give rise to the same distribution. In particular, if one does not impose symmetry, one can construct examples of $\mathcal{I}$ that are non-symmetric which satisfy the Dalton population principle but are not law-invariant.

The condition of $\mathcal{I}$ being law-invariant also has the subtlety of domain of definition that the Dalton population principle has, but this can be cleanly swept under the carpet as follows. Let $(S, \mathcal{S}, \nu)$ be a probability space and let $\mathcal{M}_S := \{\mathsf{L}\colon (S, \mathcal{S}) \to (\mathbb{R}_{\geq 0}, \mathcal{B}(\mathbb{R}_{\geq 0}))\}$ denote the set of random variables defined on $(S, \mathcal{S}, \nu)$, measureable with respect to the Borel sigma-algebra on $\mathbb{R}_{\geq 0}$. For a *given* S (with associated $\mathcal{S}$ and ν), one can talk precisely of $\mathcal{I}\colon \mathcal{M}_S \to \mathbb{R}_{\geq 0}$. However, one can *not* talk sensibly of "$\mathcal{I}\colon \bigcup_S \mathcal{M}_S \to \mathbb{R}_{\geq 0}$" where the putative union is over "all sample spaces" (i.e. over all possible sets, with all the difficulties such a notion implies). However, there is a simple fix that avoids such a definitional conundrum: the whole point of the notion of law invariance is that $\mathcal{I}$ depends upon $\mathsf{L}\colon S \to \mathbb{R}_{\geq 0}$ *only* in terms of the probability distribution $\mu_{\mathsf{L}}(A) = (\nu \circ \mathsf{L}^{-1})(A)$. That is, one can write $\mathcal{I}(\mathsf{L}) = \bar{\mathcal{I}}(\mu_{\mathsf{L}})$ for some function $\bar{\mathcal{I}}$. Whilst to *compute* μ_{L} one needs to specify a sample space S (and base measure ν) the distribution itself is simply a function $\mu_{\mathsf{L}}\colon \mathcal{B}(\mathbb{R}_{\geq 0}) \to [0,1]$ — and the sample space becomes invisible. We could thus construe $\mathcal{I}, \mathcal{R}$ and $\mathcal{D}$ as having the type signature $\mathcal{I}, \mathcal{R}, \mathcal{D}\colon \Delta(\mathbb{R}_{\geq 0}) \to \mathbb{R}_{\geq 0}$, where $\Delta(\mathbb{R}_{\geq 0})$ denotes the set of probability distributions on $\mathbb{R}_{\geq 0}$. (Confer the arguments of Le Cam (1986, Chapter 1) regarding avoiding explicit use of the sample space.)

B.3.5. I5 Normalization

We restate I5 in the language of random variables as: $\mathcal{I}(\mathsf{X}) \geq 0$ for all $\mathsf{X}\colon [n] \to \mathbb{R}_{\geq 0}$ and $\mathcal{I}(\mathsf{X}) = 0$ if and only if $\mathsf{X} = \mathsf{C}$ where C is a constant random variable, $\mathsf{C}(i) = C$ for all $i \in [n]$ for some $C \in \mathbb{R}_{\geq 0}$.

Lemma 22. *Suppose $\mathcal{I}$ satisfies I5, then $\mathcal{R}_{\mathcal{I}}$ satisfies F6 and F8. Conversely, Suppose $\mathcal{R}$ satisfies F6 and F8. Then $\mathcal{I}_{\mathcal{R}}$ satisfies I5.*

Proof. $\mathcal{I}$ satisfying I5 implies $[\mathcal{R}_{\mathcal{I}}(\mathsf{X}) = \mathbb{E}(\mathsf{X})(1 + \mathcal{I}(\mathsf{X})) \geq 0 \ \ \forall \mathsf{X}\colon [n] \to \mathbb{R}_{\geq 0}]$ and $[\mathcal{R}_{\mathcal{I}}(\mathsf{X}) = \mathbb{E}(\mathsf{X}) \text{ if and only if } \mathsf{X} = \mathsf{C}]$. Thus $\mathcal{R}_{\mathcal{I}}$ satisfies F6. Furthermore we have $\mathcal{R}(\mathsf{C}) = \mathbb{E}(\mathsf{C})[1 + \mathcal{I}(\mathsf{C})] = C$ and so $\mathcal{R}_{\mathcal{I}}$ satisfies F8.

Conversely suppose $\mathcal{R}$ satisfies F6 and F8 and $\mathcal{I}_{\mathcal{R}}(\mathsf{X}) = \frac{\mathcal{R}(\mathsf{X})}{\mathbb{E}(\mathsf{X})} - 1$. Suppose X is constant ($\mathsf{X} = \mathsf{C}$), then F8 implies $\mathcal{R}(\mathsf{C}) = C$. But $\mathbb{E}(\mathsf{C}) = C$ also and thus $\mathcal{I}_{\mathcal{R}}(\mathsf{X}) = 0$. Alternatively, if X is non-constant, then by F6, we have $\mathcal{I}_{\mathcal{R}}(\mathsf{X}) > 0$. Thus $\mathcal{I}_{\mathcal{R}}$ satisfies I5. □

Since we know (main body of paper) that F5 and F6 imply F8, we also have the second part of above lemma holding where the assumption is simply F5.

B.3.6. I6 Constant Addition

Lemma 23. *Suppose $\mathcal{R}$ satisfies F5 and [F6 or (F1 and F2) or (Subadditivity and F2)], then $\mathcal{I}_{\mathcal{R}}$ satisfies I6. Conversely, suppose $\mathcal{I}$ satisfies I5, then $\mathcal{R}_{\mathcal{I}}$ satisfies F6.*

Proof. Translating I6 to the language of random variables we require $\mathcal{I}_{\mathcal{R}}(\mathsf{X} + \mathsf{C}) \leq \mathcal{I}_{\mathcal{R}}(\mathsf{X})$ for all $\mathsf{X}\colon [n] \to \mathbb{R}_{\geq 0}$ and all constant random variables $\mathsf{C}(i) = C$ for all $i \in [n]$ for some constant $C > 0$. If $\mathsf{X}(i) = 0$ for $i \in [n]$, then F6 implies

$\mathcal{R}(\mathsf{X}) = \mathbb{E}(\mathsf{X})$ and thus $\mathcal{I}_\mathcal{R}(\mathsf{X}+\mathsf{C}) = \mathcal{I}_\mathcal{R}(\mathsf{C}) = 0 = \mathcal{I}_\mathcal{R}(\mathsf{X})$ and I6 is satisfied. Alternatively, if $\mathsf{X} > 0$ then $\mathbb{E}(\mathsf{X}) > 0$ and we have

$$\begin{aligned}
\text{F6} &\Rightarrow \mathbb{E}(\mathsf{X}) \le \mathcal{R}(\mathsf{X}) && \forall \mathsf{X} > 0 \\
&\Rightarrow C(\mathbb{E}(\mathsf{X}) - \mathcal{R}(\mathsf{X})) \ge 0 && \forall \mathsf{X} > 0,\ \forall C \in (0,\infty) \\
&\Rightarrow \frac{C(\mathbb{E}(\mathsf{X}) - \mathcal{R}(\mathsf{X}))}{\mathbb{E}(\mathsf{X})(\mathbb{E}(\mathsf{X}) + C)} \le 0 && \forall \mathsf{X} > 0,\ \forall C \in (0,\infty) \\
&\Rightarrow \frac{\mathcal{R}(\mathsf{X})\mathbb{E}(\mathsf{X}) + C\mathbb{E}(\mathsf{X}) - \mathcal{R}(\mathsf{X})\mathbb{E}(\mathsf{X}) - \mathcal{R}(\mathsf{X})C}{\mathbb{E}(\mathsf{X})(\mathbb{E}(\mathsf{X}) + C)} \le 0 && \forall \mathsf{X} > 0,\ \forall C \in (0,\infty) \\
&\Rightarrow \frac{\mathcal{R}(\mathsf{X}) + C}{\mathbb{E}(\mathsf{X}) + C} - \frac{\mathcal{R}(\mathsf{X})}{\mathbb{E}(\mathsf{X})} \le 0 && \forall \mathsf{X} > 0,\ \forall C \in (0,\infty) \\
&\Rightarrow \frac{\mathcal{R}(\mathsf{X}+\mathsf{C})}{\mathbb{E}(\mathsf{X}+\mathsf{C})} - 1 \le \frac{\mathcal{R}(\mathsf{X})}{\mathbb{E}(\mathsf{X})} - 1 && \forall \mathsf{X} > 0,\ \forall C \in (0,\infty),
\end{aligned}$$

where pulling the C into the argument of $\mathcal{R}$ is justified by F6 or (F1 & F2) or (Subadditivity & F2)

$$\begin{aligned}
&\Rightarrow \mathcal{I}_\mathcal{R}(\mathsf{X}+\mathsf{C}) \le \mathcal{I}_\mathcal{R}(\mathsf{X}) && \forall \mathsf{X} > 0,\ \forall C \in (0,\infty) \\
&\Rightarrow \text{I6}.
\end{aligned}$$

Conversely, given $\mathcal{I}$, consider $\mathcal{R}_\mathcal{I}(\mathsf{X}) = \mathbb{E}(\mathsf{X})(1+\mathcal{I}(\mathsf{X}))$. If $\mathsf{X} = \mathsf{C}$ is constant then $\mathcal{R}_\mathcal{I}(\mathsf{C}) = \mathbb{E}(\mathsf{C})(1+\mathcal{I}(\mathsf{C})) = \mathbb{E}(\mathsf{C}) = C$, which proves one part of F8. If X is not constant, then by I5, $\mathcal{I}(\mathsf{X}) > 0$ and thus $\mathcal{R}_\mathcal{I}(\mathsf{X}) > \mathbb{E}(\mathsf{X})$ which proves the other part of F8. □

B.3.7. I7 Lorenz Compatibility

Foster (1985) characterised Lorenz compatibility of inequality measures via the following:

Lemma 24. *An inequality measure is Lorenz compatible (I7) if and only if it satisfies I1, I2, I3 & I4.*

Relations between I7 and F1,...,F9 thus follow from Lemma (24) and the earlier lemmas.

B.3.8. I8, I9, I10 and I11 (Decomposability)

We consider I8–I11 together. It is now convenient to write inequality measures (and deviation and risk measures) in terms of vectors $x \in \mathbb{R}^n_{\ge 0}$. We write $\mathbb{E}(x) = \|x\|_1/n$, where $\|x\|_1 = \sum_{i=1}^n x_i$ for $x \in \mathbb{R}^n_{\ge 0}$.

We need the following lemma (Marshall et al., 2011, page 92, C.1.a) applied to the interval $(0,\infty)$:

Lemma 25. *Suppose $\mathcal{I}$ is seperable (I10). Then $\mathcal{I}$ is strictly Schur-convex if and only if g is strictly convex.*

A consequence of this lemma is that if $\mathcal{I}$ is seperable and satisfies I3, then $\mathcal{I}$ is strictly convex. We consider deviation measures induced from an inequality measure: $\mathcal{D}_\mathcal{I}(x) := \mathbb{E}(x)\mathcal{I}(x)$.

Lemma 26. *Suppose $\mathcal{I}$ satisfies [I2 and I3 and I10] or [I2 and I11 (strict)]. Then $\mathcal{D}_\mathcal{I}$ (resp. $\mathcal{R}_\mathcal{I}$) is strictly convex and positively homogeneous (F1 and F2).*

Proof. For the first case, by Lemma 25, $\mathcal{I}$ is strictly convex. The second case simply assumes constant sum strict convexity of $\mathcal{I}$, so we can henceforth presume it. For $m > 0$ let $A_m := \{x \in \mathbb{R}^n_{\ge 0} \colon \mathbb{E}(x) = m\}$. We first show that $\mathcal{D}_\mathcal{I}$ is strictly convex on A_m for all $m > 0$. Fix $m > 0$ and pick $x, y \in A_m$. Then for any $\lambda \in (0,1)$,

$$\begin{aligned}
\mathcal{D}_\mathcal{I}(\lambda x + (1-\lambda)y) &= \mathbb{E}(\lambda x + (1-\lambda y))\mathcal{I}(\lambda x + (1-\lambda))y) \\
&= m\mathcal{I}(\lambda x + (1-\lambda)y) \\
&< m(\lambda\mathcal{I}(x) + (1-\lambda)\mathcal{I}(y)) \\
&= \lambda m\mathcal{I}(x) + (1-\lambda)m\mathcal{I}(y) \\
&= \lambda\mathcal{D}_\mathcal{I}(x) + (1-\lambda)\mathcal{D}_\mathcal{I}(y),
\end{aligned}$$

and thus $\mathcal{D}_\mathcal{I}$ is strictly convex on A_m for all $m > 0$. Lemma 18 implies that $\mathcal{D}_\mathcal{I}$ is 1-homogeneous on $\mathbb{R}^n_{\geq 0}$. We now show that these two facts imply $\mathcal{D}_\mathcal{I}$ is strictly convex on $\mathbb{R}^n_{\geq 0}$.

Pick $x, y \in \mathbb{R}^n_{\geq 0}$ and let $m_x := \mathbb{E}(x)$ and $m_y := \mathbb{E}(y)$. Observe that $\mathbb{E}(\lambda x + (1-\lambda)y) = \lambda m_x + (1-\lambda)m_y$. Furthermore, since $\mathcal{D}_\mathcal{I}$ is 1-homogeneous, for all $x \in \mathbb{R}^n_{\geq 0}$, $\mathcal{D}_\mathcal{I}(x) = \|x\|_1 \left(\frac{1}{\|x\|_1}\mathcal{D}_\mathcal{I}(x)\right) = \|x\|_1 \mathcal{D}_\mathcal{I}(x/\|x\|_1)$. Thus for all $\lambda \in (0,1)$

$$\begin{aligned}\mathcal{D}_\mathcal{I}(\lambda x + (1-\lambda)y) &= \|\lambda x + (1-\lambda)y\|_1 \mathcal{D}_\mathcal{I}\left(\frac{\lambda x + (1-\lambda)y}{\|\lambda x + (1-\lambda)y\|_1}\right) \\ &= (\lambda n m_x + (1-\lambda) n m_y)\mathcal{D}_\mathcal{I}(\lambda \bar{x} + (1-\lambda)\bar{y}),\end{aligned}$$

where $\bar{x} := x/(\lambda n m_x + (1-\lambda) n m_y)$ and $\bar{y} := y/(\lambda n m_x + (1-\lambda) n m_y)$ and $\bar{x}, \bar{y} \in A_n$. Since $\mathcal{D}_\mathcal{I}$ is strictly convex on A_n we have

$$\begin{aligned}\mathcal{D}_\mathcal{I}(\lambda x + (1-\lambda)y) &< (\lambda n m_x + (1-\lambda) n m_y)\mathcal{D}_\mathcal{I}(\bar{x}) + (1-\lambda)(\lambda n m_x + (1-\lambda) n m_y)\mathcal{D}_\mathcal{I}(\bar{y}) \\ &= \lambda \mathcal{D}_\mathcal{I}(x) + (1-\lambda)\mathcal{D}_\mathcal{I}(y),\end{aligned}$$

and thus $\mathcal{D}_\mathcal{I}$ is strictly convex. Since $\mathcal{R}_\mathcal{I}(x) = \mathbb{E}(x) + \mathcal{D}_\mathcal{I}(x)$, strict convexity and positive homogeneity of $\mathcal{R}_\mathcal{I}$ follows from the fact that these properties are preserved under summation with the convex and positively homogeneous function $x \mapsto \mathbb{E}(x)$. □

If $\mathcal{I}$ is not seperable, then I3 does not guarantee strict convexity (or even convexity) of $\mathcal{I}$ since not every strictly Schur-convex function is strictly convex in which case we can not guarantee the strict convexity (or even convexity) of $\mathcal{D}_\mathcal{I}$.

(Kolm, 1976b) has argued that one could equally normalise inequality measures via 1-homogeneity, in which case the bridge to risk and deviation measures is even simpler — such an 1-homogeneous inequality measure *is* a deviation measure.

B.4. Relating inequality measures and risk measures

Let FRM denote the set of fairness risk measures (i.e. $\mathcal{R}$ satisfying F1–F7) and let SIM denote the set of standardised inequality measures (i.e. satisfying Lorenz compatibility I7, normalisation I5 and constant addition I6). Let $\mathcal{R}_{\text{SIM}} := \{\mathcal{R}_\mathcal{I} \colon \mathcal{I} \in \text{SIM}\}$.

Theorem 27. *Suppose $S = [n]$ and $\nu = \nu^n_{\text{u}}$ and that $\mathcal{R}_\mathcal{I}$ and $\mathcal{I}_\mathcal{R}$ are defined by (33)–(36).*

1. *If $\mathcal{R} \in$ FRM then $\mathcal{I}_\mathcal{R} \in$ SIM.*
2. *$\mathcal{I} \in$ SIM then $\mathcal{R}_\mathcal{I}$ satisfies F2, F6, F7 and F8. If furthermore $\mathcal{I}$ satisfies I10, then $\mathcal{R}_\mathcal{I}$ also satisfies F1.*

Thus FRM $\subset \mathcal{R}_{\text{SIM}}$.

Proof. Part 1 follows from lemmas 17, 18, 20, 21, 22 and 23. Part 2 follows from lemmas 17, 18, 22, 23 and 26. □

The theorem suggests that fairness risk measures are a stronger notion than inequality measures. We now give some intuition as to why it is not plausible that the second part of the theorem can be strengthened, and that fairness risk measures are indeed a stronger (more restrictive) notion.

A crucial property of fairness risk measures is monotonicity (F3). Observe that if F3 is satisfied, not only do we have that (translating to vector notation for now) for $x, y \in \mathbb{R}^n_{\geq 0}$, $x \leq y \Rightarrow \mathcal{R}(x) \leq \mathcal{R}(y)$ but also $x \leq y \Rightarrow \mathbb{E}(x) \leq \mathbb{E}(y)$. Since $\mathcal{I}_\mathcal{R}(x) = \frac{\mathcal{R}(x)}{\mathbb{E}(x)} - 1$, we can conclude nothing about $\mathcal{I}(x+y)$ on the basis of these assumptions. Furthermore, to consider the converse direction, the crucial property of strict Schur-Convexity of $\mathcal{I}$ is equivalent to the requirement that $x \succ_{\text{M}} y \Leftrightarrow \mathcal{I}(x) \leq \mathcal{I}(y)$ for all such $\mathcal{I}$. But the relationship of majorisation conflicts with the pointwise domination required for monotonicity since if $x \leq y$ and $x \succ_{\text{M}} y$ then $x = y$ (Marshall et al., 2011, page 13). Thus the property of pointwise inequality $x < y$ is "invisible" to inequality measures. This failure matters for our motivating purpose — to find learned hypotheses that performs well in the traditional sense (average losses are small) *and* satisfies some notion of equity or fairness. While inequality measures can judge the unfairness, if one combines them with the average loss in a dimensionally sensible way (such as our 1-homogeneous proposal $\mathcal{R}_\mathcal{I}$), the resulting risk measure is not guaranteed to satisfy the highly desirable property of monotonicity.

Although most specific proposed inequality measures are continuous, continuity does not feature in the list of axioms typically applied to inequality measures. And thus we can not guarantee that $\mathcal{R}_{\mathcal{I}}$ satsifies F4. Of course one could trivially impose it upon $\mathcal{I}$ and the same continuity would immediately hold for $\mathcal{R}_{\mathcal{I}}$.

Finally there is an intrinsic difficulty in deriving an *equality* constraint (F5) on $\mathcal{R}_{\mathcal{I}}$ from an *inequality* constraint on $\mathcal{I}$ such as I6.

Thus in general we can not guarantee that $\mathcal{R}_{\mathcal{I}}$ satisfies F3, F4 and F5 and so $\mathcal{R}_{\mathcal{I}}$ is not guaranteed to be a fairness risk measure.

B.5. Consequences for Fair Machine Learning and Beyond

Our conclusion from the above analysis is that for the purposes of learning hypotheses from empirical data that perform well in terms of expected loss and allow the control of fairness, it is better to use the slightly more restricted class of fairness risk measures FRM rather than attempting to work with the larger class $\mathcal{R}_{\text{SIM}}$ which lacks many of the desirable properties of a fairness risk measure (convexity, monotonicity, and continuity). Since $\mathcal{R} \in \text{FRM}$ are also attractive from a computational standpoint, fairness risk measures seem to be preferrable to inequality measures as a basis for fair machine learning.

C. Additional experiments

We present some experimental results supplementing those in the body.

C.1. Results with categorical sensitive feature

We illustrate the viability of using the CVaR-fairness learner with a categorical sensitive feature S. We consider the `adult` dataset, but this time with `race` as the sensitive feature. This feature takes on 5 distinct values. We train the CVaR-fairness learner with fixed regularisation $C = 1$.

Figure 2 shows the average subgroup risk, and the difference between the maximal and minimal subgroup risks. We see that the average subgroup risk is maintained fairly constant. However, as α is increased, the gap between the maximal and minimal risks shrinks, i.e., the subgroup risks become more commensurate.

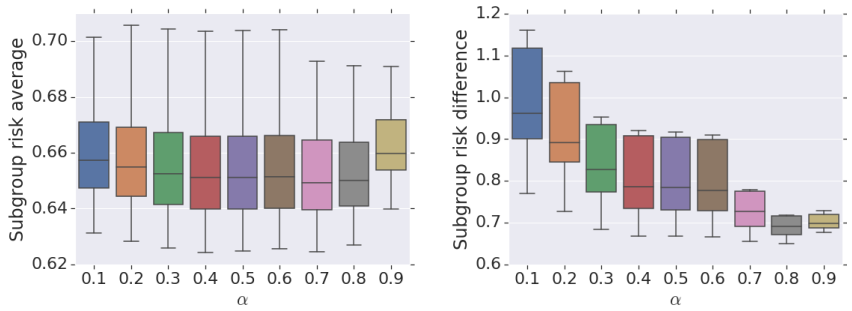

Figure 2. Results on `adult` dataset with `race` as the categorical sensitive feature. The panels show that as α is varied, CVaR-based optimisation results in the maximal and minimal subgroup risks being driven closer together.

C.2. Results with real-valued sensitive feature

We illustrate the viability of using the CVaR-fairness learner with a real-valued sensitive feature S. We consider the `adult` dataset, but this time with `fnlwgt` (an estimate of how representative an individual is) as the sensitive feature. Following 31, essentially all instances are placed into separate subgroups in forming the CVaR objective. We train the CVaR-fairness learner with fixed regularisation $C = 1$.

Figure 3 compares the histogram of margin scores $\{y_i \cdot f(x_i)\}_{i=1}^m$ for $\alpha = 0.1$ and $\alpha = 0.9$. We see that, as expected, setting $\alpha = 0.9$ encourages all scores to be roughly commensurate.

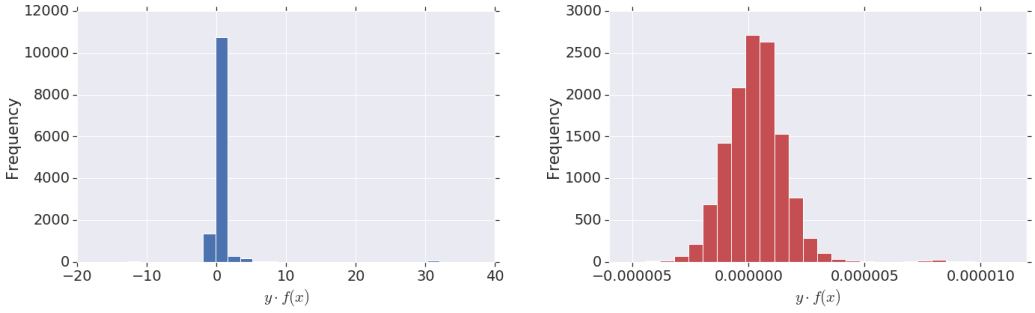

Figure 3. Results on `adult` dataset with `fnlwgt` as the continuous sensitive feature. The left and right panel are for $\alpha = 0.1$ and $\alpha = 0.9$ respectively. Since essentially each individual is a member of a singleton subgroup, the latter is seen to encourage commensurate model predictions, and thus margin scores across all instances.

C.3. Additional results on `synth` and `adult`

We present additional results for the experimental setting in the body. In Figures 4 — 9, we show the behaviour of the CVaR method as α is varied with respect to different metrics. In terms of fairness metrics, this includes the violation of the

balanced error, demographic parity and equality of opportunity, defined as

$$\begin{aligned}\Delta_{\mathrm{BER}}(f) &= |\mathrm{BER}(f; \mathbb{P}_{\mathsf{X},\mathsf{Y}|\mathsf{S}=1}) - \mathrm{BER}(f; \mathbb{P}_{\mathsf{X},\mathsf{Y}|\mathsf{S}=0})| \\ \Delta_{\mathrm{DP}}(f) &= |\mathbb{P}(f(\mathsf{X}) > 0 \mid \mathsf{S}=1) - \mathbb{P}(f(\mathsf{X}) > 0 \mid \mathsf{S}=0)| \\ \Delta_{\mathrm{EO}}(f) &= |\mathbb{P}(f(\mathsf{X}) > 0 \mid \mathsf{S}=1, \mathsf{Y}=1) - \mathbb{P}(f(\mathsf{X}) > 0 \mid \mathsf{S}=0, \mathsf{Y}=1)|.\end{aligned}$$

Here, $\mathrm{BER}(f; \mathbb{P})$ is the balanced error of a classifier on distribution $\mathbb{P}$. The difference in the balanced error across the subgroups is a particular instance of the lack of disparate mistreatment objective (5).

We present results using three different choices of subgroup risk $L_s(f)$ for the model:

$$\begin{aligned}L_s^{(1)}(f) &= \mathop{\mathbb{E}}_{\mathsf{X},\mathsf{Y}|\mathsf{S}=s}\left[\ell(\mathsf{Y}, f(\mathsf{X}))\right] \\ L_s^{(2)}(f) &= \mathop{\mathbb{E}}_{\mathsf{X},\mathsf{Y}|\mathsf{S}=s}\left[\mathbb{P}(\mathsf{Y})^{-1} \cdot \ell(\mathsf{Y}, f(\mathsf{X}))\right] \\ L_s^{(3)}(f) &= \mathop{\mathbb{E}}_{\mathsf{X},\mathsf{Y}|\mathsf{S}=s}\left[\mathbb{P}(\mathsf{Y} \mid \mathsf{S}=s)^{-1} \cdot \ell(\mathsf{Y}, f(\mathsf{X}))\right],\end{aligned}$$

where ℓ is the square-hinge loss. Here, $L^{(1)}$ is the standard subgroup risk, $L^{(2)}$ is the subgroup risk arising from balancing the class-labels *globally*, and $L^{(3)}$ is the subgroup risk arising from balancing the class-labels *within each subgroup*. Explicitly, using $L^{(2)}$ yields

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathsf{S}}\left[L_{\mathsf{S}}^{(2)}(f)\right] &= \mathop{\mathbb{E}}_{\mathsf{S}}\left[\mathop{\mathbb{E}}_{\mathsf{X},\mathsf{Y}|\mathsf{S}}\left[\mathbb{P}(\mathsf{Y})^{-1} \cdot \ell(\mathsf{Y}, f(\mathsf{X}))\right]\right] \\ &= \mathop{\mathbb{E}}_{\mathsf{X},\mathsf{Y}}\left[\mathbb{P}(\mathsf{Y})^{-1} \cdot \ell(\mathsf{Y}, f(\mathsf{X}))\right] \\ &= \sum_{y \in Y} \mathop{\mathbb{E}}_{\mathsf{X}|\mathsf{Y}=y}\left[\ell(y, f(\mathsf{X}))\right] \\ &\propto \mathop{\mathbb{E}}_{\mathsf{Y} \sim \mathrm{Unif}(Y)}\left[\mathop{\mathbb{E}}_{\mathsf{X}|\mathsf{Y}=y}\left[\ell(y, f(\mathsf{X}))\right]\right],\end{aligned}$$

so that the aggregate is a balanced version of the risk. On the other hand, using $L^{(3)}$,

$$\begin{aligned}L_s^{(3)}(f) &= \mathop{\mathbb{E}}_{\mathsf{X},\mathsf{Y}|\mathsf{S}=s}\left[\mathbb{P}(\mathsf{Y} \mid \mathsf{S}=s)^{-1} \cdot \ell(\mathsf{Y}, f(\mathsf{X}))\right] \\ &= \mathop{\mathbb{E}}_{\mathsf{Y}|\mathsf{S}=s}\left[\mathop{\mathbb{E}}_{\mathsf{X}|\mathsf{Y},\mathsf{S}=s}\left[\mathbb{P}(\mathsf{Y} \mid \mathsf{S}=s)^{-1} \cdot \ell(\mathsf{Y}, f(\mathsf{X}))\right]\right] \\ &= \sum_{y \in Y} \mathop{\mathbb{E}}_{\mathsf{X}|\mathsf{Y}=y,\mathsf{S}=s}\left[\ell(y, f(\mathsf{X}))\right] \\ &\propto \mathop{\mathbb{E}}_{\mathsf{Y} \sim \mathrm{Unif}(Y)}\left[\mathop{\mathbb{E}}_{\mathsf{X}|\mathsf{Y}=y,\mathsf{S}=s}\left[\ell(y, f(\mathsf{X}))\right]\right],\end{aligned}$$

so that the subgroup risk itself is balanced across the distribution of labels for the subgroup.

We present results for each of the three settings in turn.[10] As in the body, increasing α generally has the effect of reducing predictive accuracy of Y while also reducing the fairness violation. Further, using the subgroup-level label balancing per $L^{(3)}$ is seen to decrease the CVaR method's fairness violations. We also find that fair-ERM generally achieves low equality of opportunity violation, which is again unsurprising given the method is designed for this goal.

We reiterate here a comment in the body: by design, the CVaR objective seeks to ensure commensurate *surrogate* subgroup risk. This does not necessarily imply commensurate the subgroup risk with respect to 0-1 loss. Nonetheless, the above shows that such behaviour can be borne out empirically.

[10] For the `adult` dataset, in these plots we fix the regularisation parameter to $C = 10^{-5}$, since for larger C even the setting $\alpha = 0$ can achieve equitable subgroup risk; such results thus do not serve the purpose of illustrating the effect of increasing α when $\alpha = 0$ provides a suboptimal tradeoff.

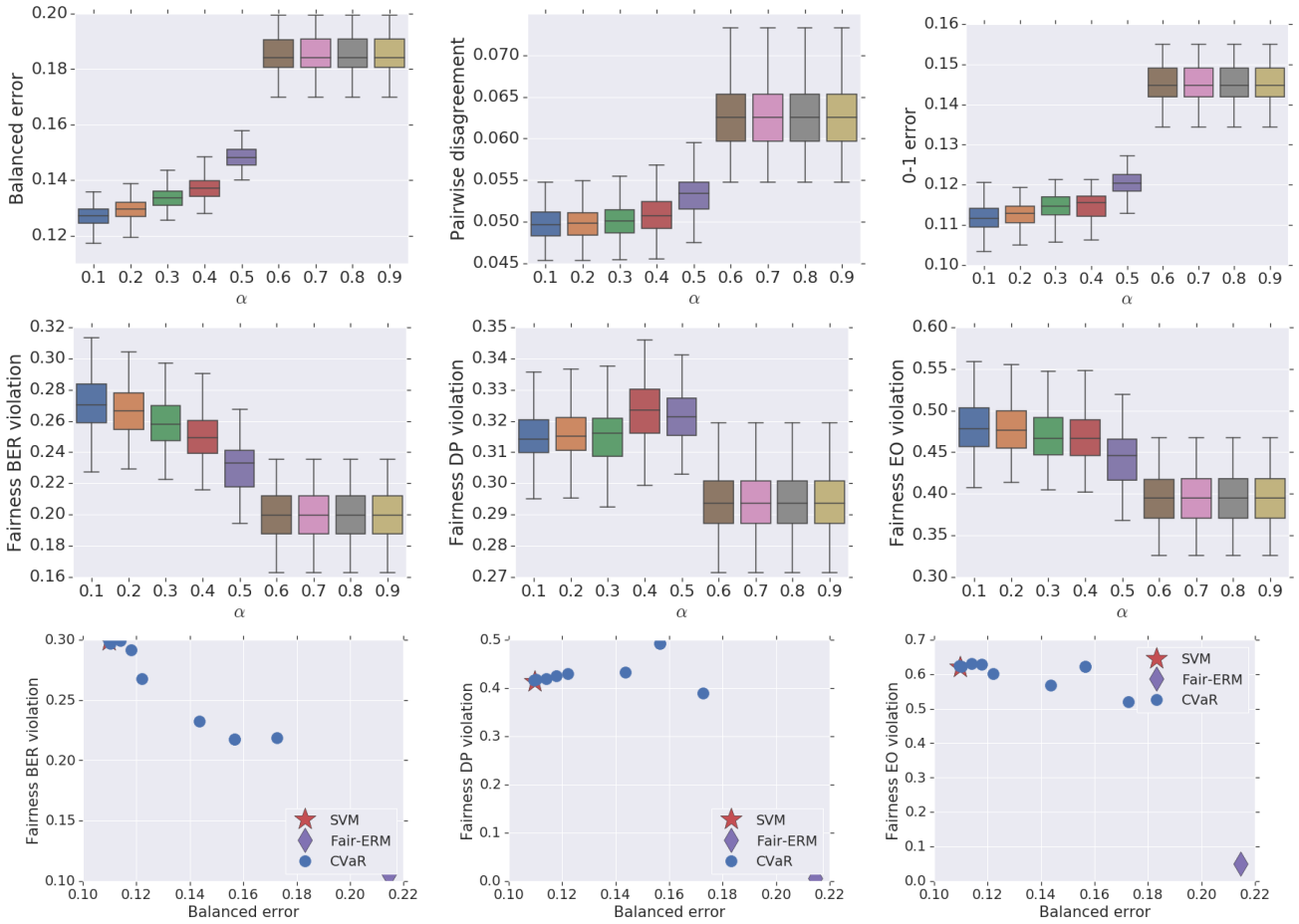

Figure 4. Results on `synth` dataset. The top panel shows three different measures of performance of CVaR in predicting the target Y: the balanced error, pairwise disagreement (one minus AUC-ROC), and 0-1 error. The middle panel show three different measures of fairness of CVaR in being agnostic towards the sensitive feature S: the difference in balanced error across subgroups, the violation of demographic parity, and violation of equality of opportunity. The bottom panel compares the CVaR model against baselines using different measures of fairness and accuracy.

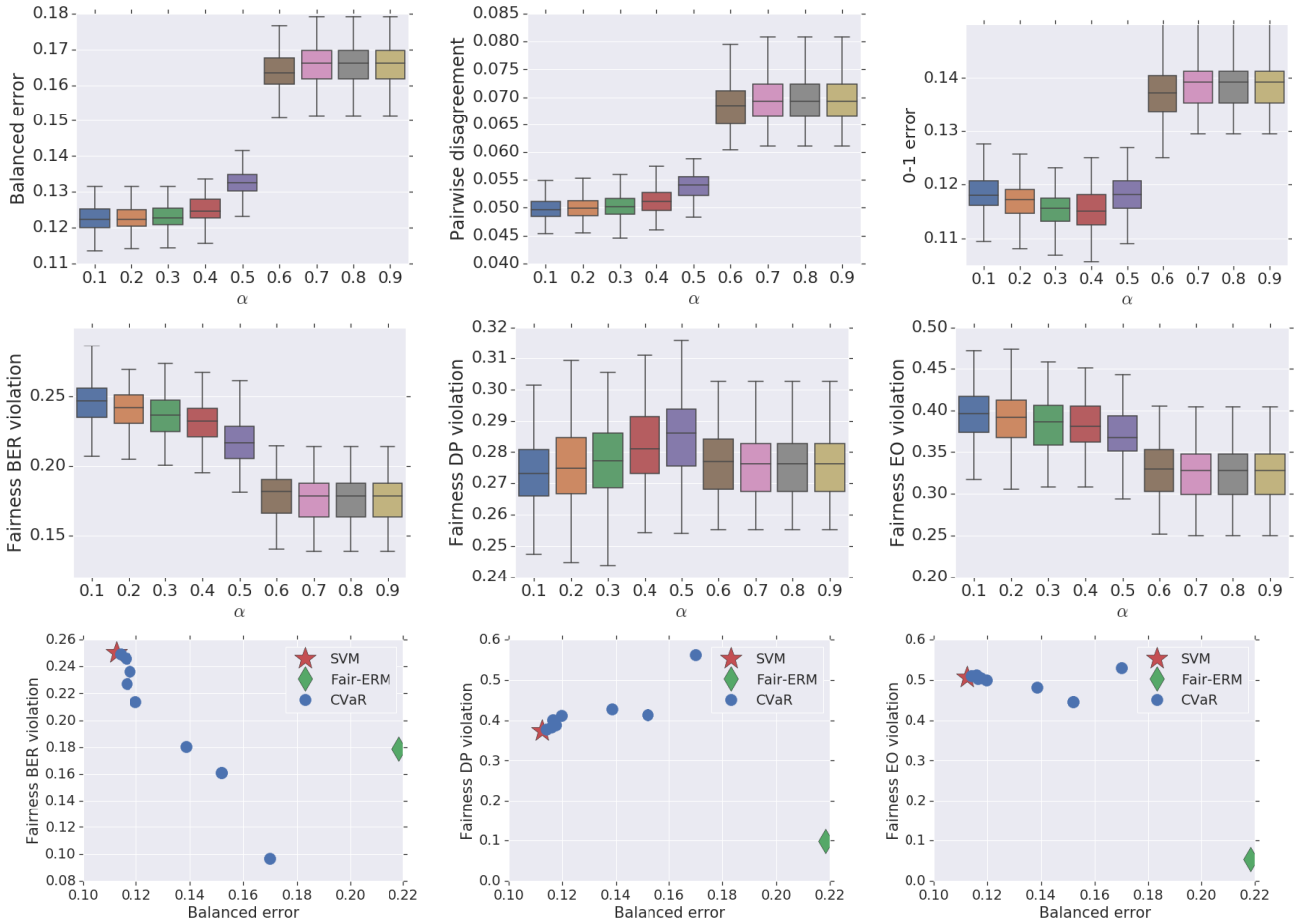

Figure 5. Results on `synth` dataset with global label balancing of the loss. The top panel shows three different measures of performance of CVaR in predicting the target Y: the balanced error, pairwise disagreement (one minus AUC-ROC), and 0-1 error. The middle panel show three different measures of fairness of CVaR in being agnostic towards the sensitive feature S: the difference in balanced error across subgroups, the violation of demographic parity, and violation of equality of opportunity. The bottom panel compares the CVaR model against baselines using different measures of fairness and accuracy.

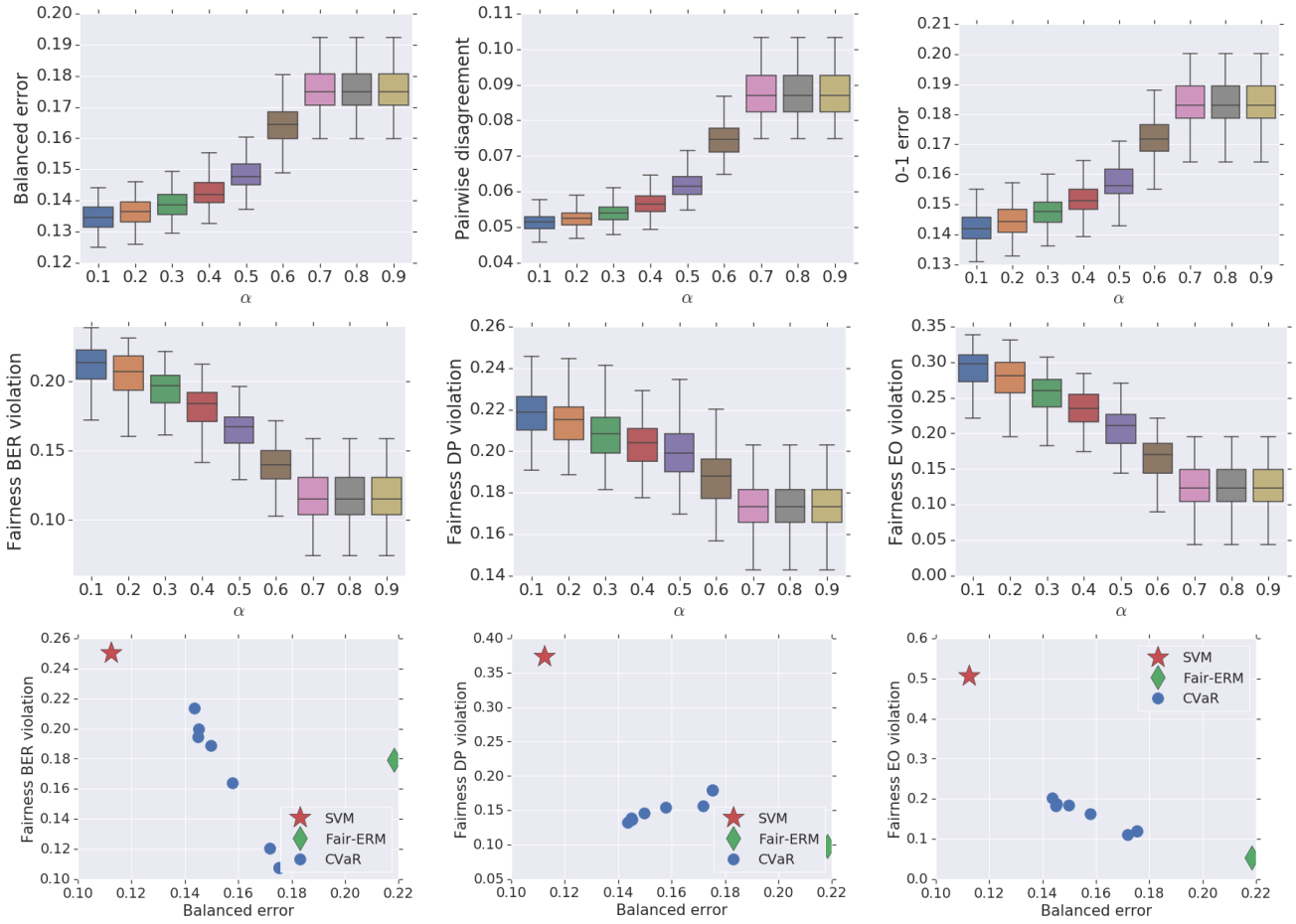

Figure 6. Results on `synth` dataset with subgroup-level label balancing of the loss. The top panel shows three different measures of performance of CVaR in predicting the target Y: the balanced error, pairwise disagreement (one minus AUC-ROC), and 0-1 error. The middle panel show three different measures of fairness of CVaR in being agnostic towards the sensitive feature S: the difference in balanced error across subgroups, the violation of demographic parity, and violation of equality of opportunity. The bottom panel compares the CVaR model against baselines using different measures of fairness and accuracy.

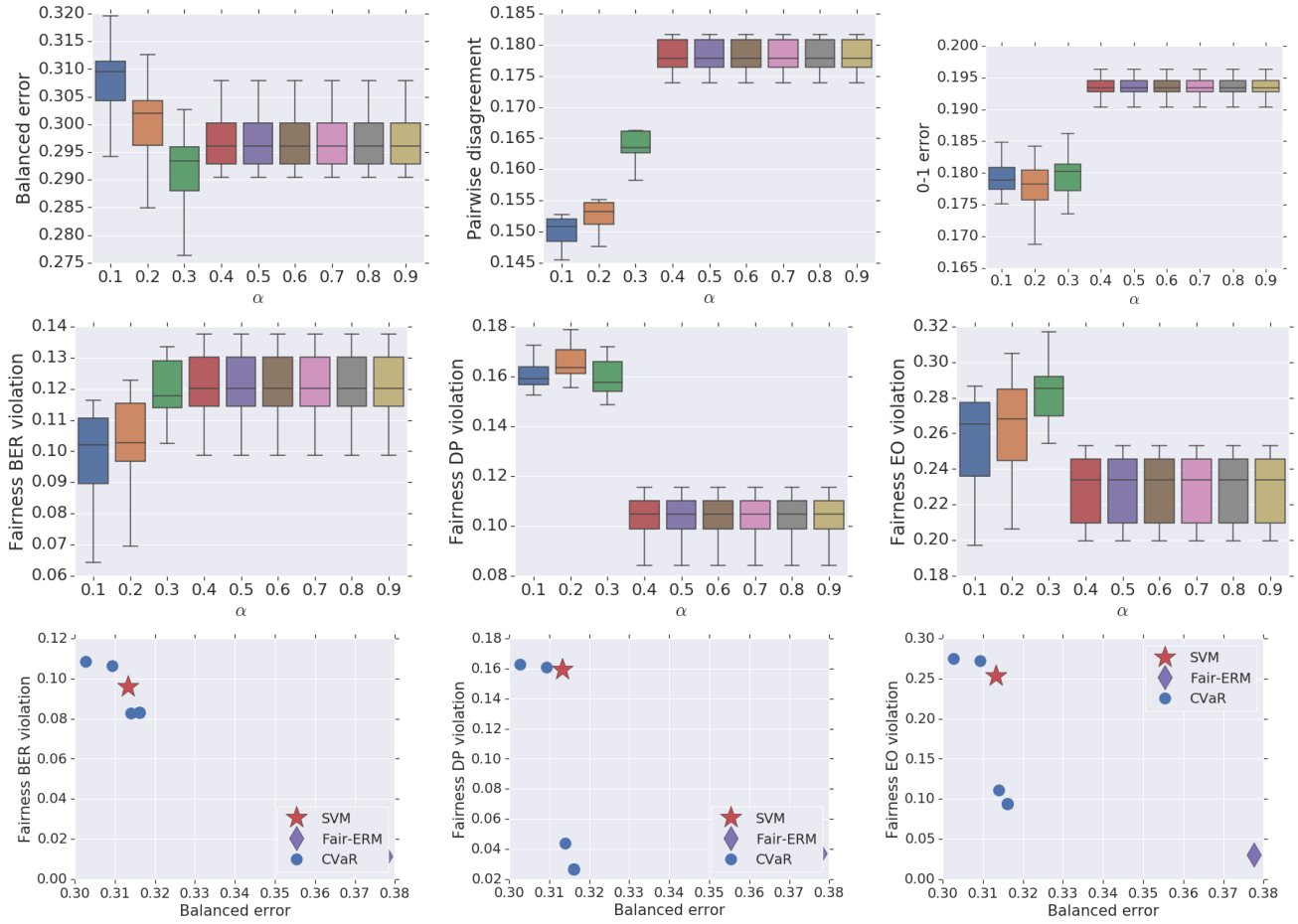

Figure 7. Results on `adult` dataset. The top panel shows three different measures of performance of CVaR in predicting the target Y: the balanced error, pairwise disagreement (one minus AUC-ROC), and 0-1 error. The middle panel show three different measures of fairness of CVaR in being agnostic towards the sensitive feature S: the difference in balanced error across subgroups, the violation of demographic parity, and violation of equality of opportunity. The bottom panel compares the CVaR model against baselines using different measures of fairness and accuracy.

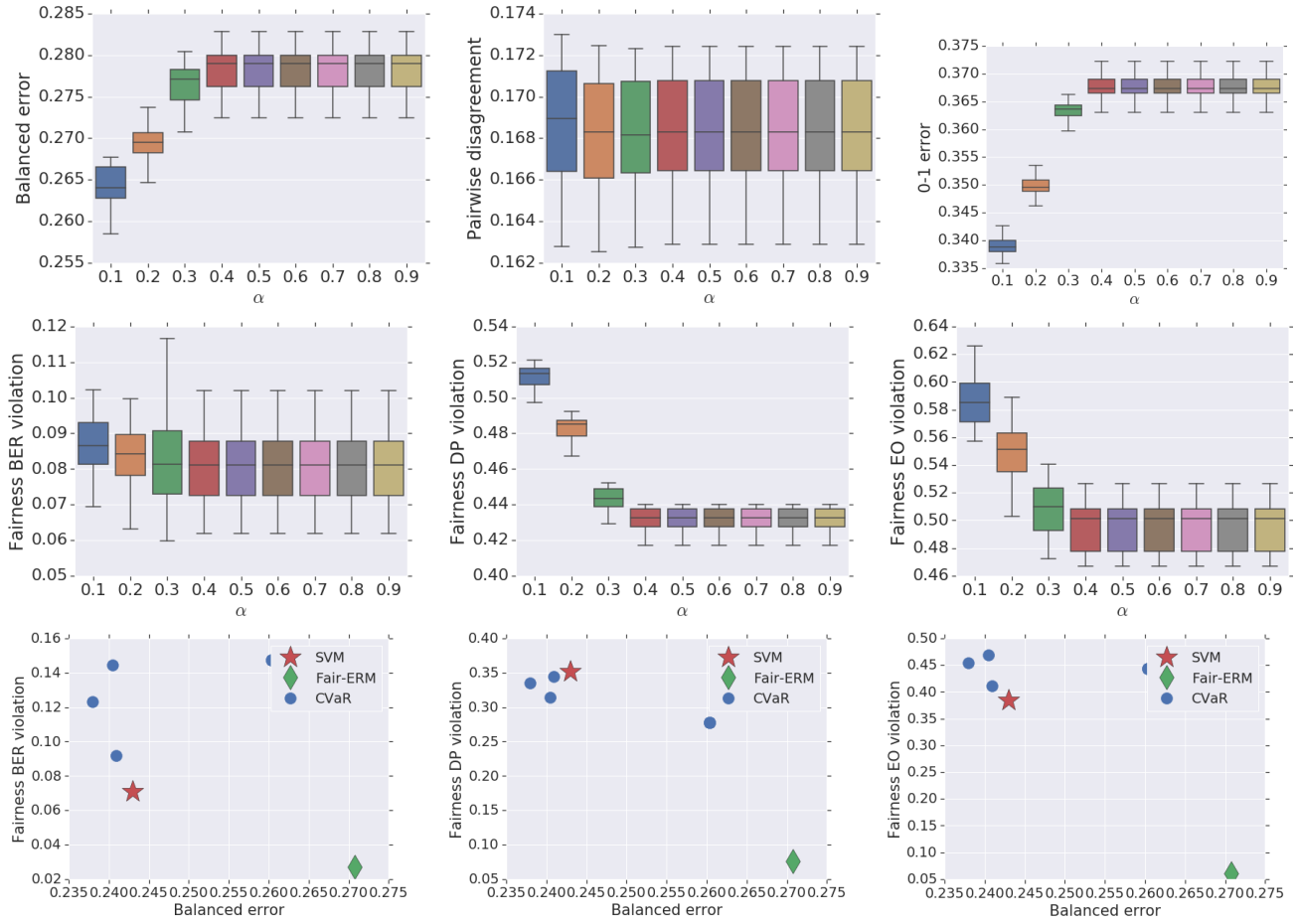

Figure 8. Results on `adult` dataset with global label balancing of the loss. The top panel shows three different measures of performance of CVaR in predicting the target Y: the balanced error, pairwise disagreement (one minus AUC-ROC), and 0-1 error. The middle panel show three different measures of fairness of CVaR in being agnostic towards the sensitive feature S: the difference in balanced error across subgroups, the violation of demographic parity, and violation of equality of opportunity. The bottom panel compares the CVaR model against baselines using different measures of fairness and accuracy.

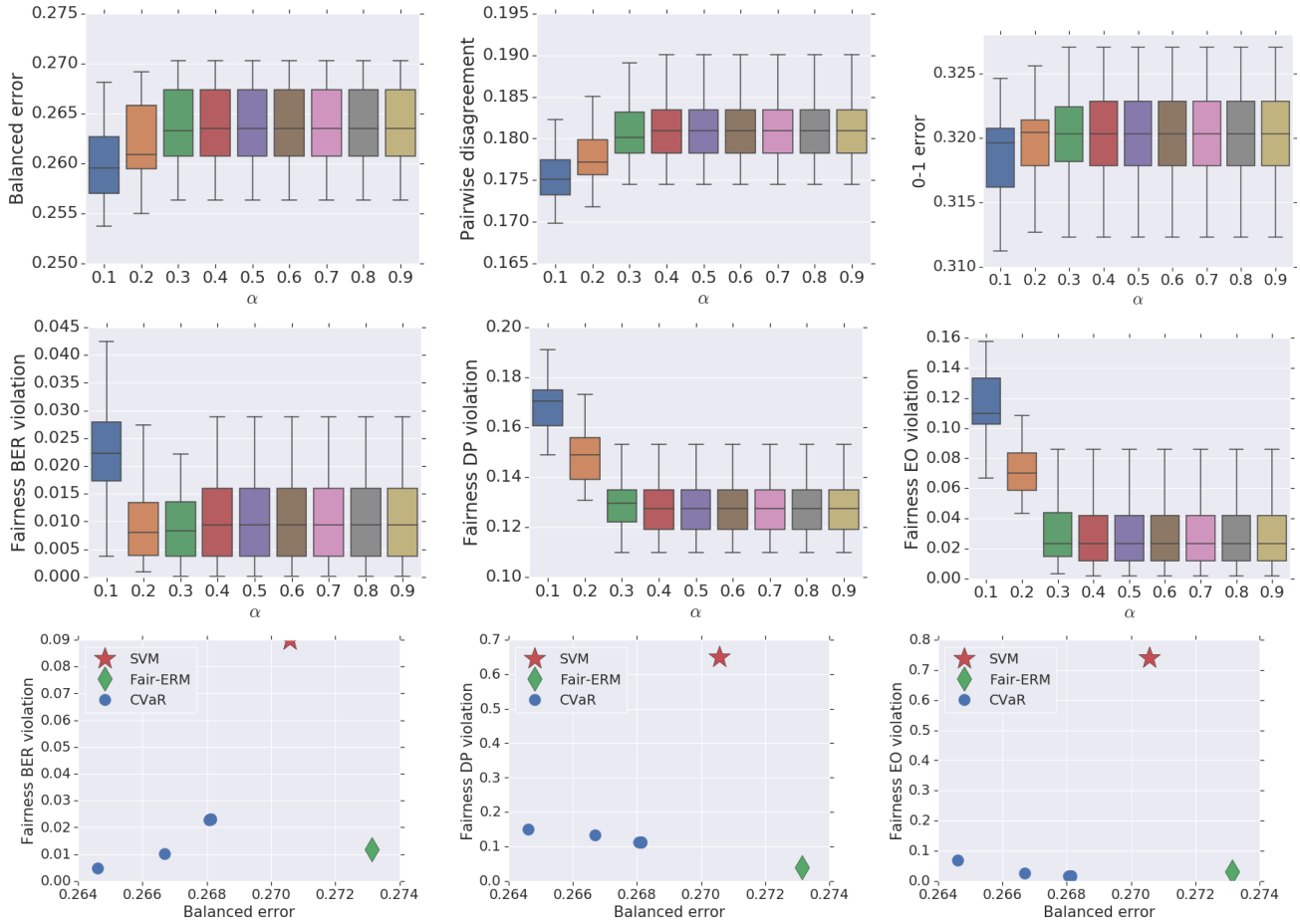

Figure 9. Results on `adult` dataset with subgroup-level label balancing of the loss. The top panel shows three different measures of performance of CVaR in predicting the target Y: the balanced error, pairwise disagreement (one minus AUC-ROC), and 0-1 error. The middle panel show three different measures of fairness of CVaR in being agnostic towards the sensitive feature S: the difference in balanced error across subgroups, the violation of demographic parity, and violation of equality of opportunity. The bottom panel compares the CVaR model against baselines using different measures of fairness and accuracy.