The Intersectionality Problem for Algorithmic Fairness

Johannes Himmelreich
Syracuse University
jrhimmel@syr.edu

Arbie Hsu
University of San Francisco
whsu10@dons.usfca.edu

Kristian Lum
University of Chicago
kristianl@uchicago.edu

Ellen Veomett
University of San Francisco
eveomett@usfca.edu

Abstract

A yet unmet challenge in algorithmic fairness is the problem of intersectionality, that is, achieving fairness across the intersection of multiple groups—and *verifying* that such fairness has been attained. Because intersectional groups tend to be small, verifying whether a model is fair raises statistical as well as moral-methodological challenges. This paper (1) elucidates the problem of intersectionality in algorithmic fairness, (2) develops desiderata to clarify the challenges underlying the problem and guide the search for potential solutions, (3) illustrates the desiderata and potential solutions by sketching a proposal using simple hypothesis testing, and (4) evaluates, partly empirically, this proposal against the proposed desiderata.

1 Introduction

That intersectionality matters is a point of consensus in the algorithmic fairness literature. A model's performance might be much worse for women of color than for women and people of color considered separately [1]. In this paper, we elucidate a problem that intersectionality raises for algorithmic fairness in practice: Because data on intersectional groups is often severely limited, *verifying* that algorithmic fairness—under various definitions thereof—has been attained is difficult. Although this problem is recognized in the literature [2, 3, 4, 5], its challenges do not appear to be fully appreciated and many existing contributions violate minimal moral or methodological desiderata.

Our contribution is fourfold: We (1) elucidate the problem of intersectionality in algorithmic fairness, and (2) develop desiderata to clarify the challenges that underlie the problem of intersectionality and to guide the search for potential solutions. Moreover, we (3) illustrate the desiderata and potential solutions by presenting a statistical setup that uses simple hypothesis testing, and (4) evaluate this proposal, partly empirically, in light of the desiderata.

Our larger aim is to advance the literature on algorithmic fairness more broadly. The approach that we propose in response to the problem of intersectionality differs fundamentally from the typical way of "measuring" algorithmic fairness.[1] We hence advance the debate, by pointing out possibilities of approaching fairness differently: as accounting for uncertainty (instead of concentrating on point estimates) and as a matter of sufficiency (instead of equality).

---

[1]We use fairness "measure," "metric" and their cognates with two caveats. First, the problem is one of estimation not measurement. Second, fairness metrics are *meta-metrics* since they aggregate a higher-dimensional vector of model performance into a lower-dimensional summary [6].

2 Preliminaries

2.1 Algorithmic Fairness

In the literature on algorithmic fairness, "fairness" is typically defined as model performance (such as accuracy or false positive rate) that is roughly equal across all relevant groups. Many versions of algorithmic fairness consider fairness to have been achieved if

$$|m(G) - m(\cdot)| < \epsilon \quad \text{for some small } \epsilon, \; \forall\, G \tag{1}$$

Where G denotes a subgroup of the population, $m(G)$ a model's performance (however understood) on only the subset of the data that belongs to group G, and $m(\cdot)$ the model's performance calculated across the entire dataset, irrespective of group membership. Membership in G typically corresponds to a sensitive or protected attribute such as race, sex, age, disability or marital status but G may also be defined intersectionally as a *combination* of such attributes.

Equation (1) generalizes a large family of definitions or—when aggregating $|m(G) - m(\cdot)|$ for all groups—metrics of fairness. We thus take (1) to represent the *typical* way of understanding algorithmic fairness. This typical way of understanding fairness faces the problem of intersectionality.

2.2 The Problem of Intersectionality

As the number of attributes that define subgroups grows, the amount of data available for each subgroup shrinks rapidly. After all, the number of subgroups grows *exponentially* with the number of protected attributes: For n binary attributes, there are 2^n intersectional groups. This, in turn, entails a data problem: When social identities are constituted by intersections of increasingly many attributes, and when these constituting attributes are not just binary, the data within each of the intersections can become very small. In Europe, where discrimination is highly intersectional and fairness audits are encouraged by legislation,[2] fairness audits may need to account for several thousand subgroups.[3] And because gathering the data necessary for fairness audits is typically costly—e.g., the "ground truth" needs to be established to assess whether a prediction is correct—such data tend to be scarce to begin with.

In short, the intersectionality problem of algorithmic fairness is a problem of statistical uncertainty due to small data and, subsequently, raises problems for how "fairness" is typically defined.

Intersectionality renders fairness metrics, as they are typically defined, meaningless. These metrics, such as (1), rely on point estimates of model performance (e.g., whether this performance is roughly the same for all groups). But point estimates become nonsensical with small data [2].[4] The challenge posed by intersectionality for algorithmic fairness is to define a fairness metric that provides meaningful estimates of fairness even when groups are very small and audit data are sparse.

Our discussion hence adds to the existing technical and critical objections against (intersectional) algorithmic fairness [10, 11], acknowledging that a commitment to intersectionality and fairness likely requires a broader set of actions than estimating certain properties of models [12, 13, 14].

3 Existing Work

Various statistical methods have been proposed for intersectionality in algorithmic fairness.

---

[2]Recital 49 of the EU Artificial Intelligence Act [7] encourages "the development of benchmarks and measurement methodologies for AI systems" [8]. Yet the statistical problems of intersectional fairness are, in some way, greater in Europe. Since nationality groups are already comparatively small, intersectional groups are even smaller subgroups within already small nationality groups. For example, Hungarian Roma face discrimination in the housing market, Maghrebi French in the labor market, whereas people of African descent in England and Wales face discrimination in the criminal justice system [9].

[3]Assuming 3 binary attributes (e.g., non-white, cis-gender, same-sex orientation) 1 three-valued attribute (e.g., gender as 'male,' 'female,' and 'neither'), 9 different ethnic backgrounds (e.g., Roma, Chinese, Turkish), and 12 different nationalities or localities (e.g., Hungarian, German, French), yields 2,592 intersectional subgroups. And this number may be conservative since the number of discernible ethnic groups is larger than 9, of nationalities is larger than 12, and the legally protected attribute of age is not even included.

[4]For example, in binary classification, an individual prediction is either 1 or 0; and the model accuracy for each singleton group is thus either 1 or 0.

3.1 Kearns et al.

An early identification and statement of the problem of intersectional fairness arising from small groups is due to Kearns et al. [2]. The approach of Kearns et al. involves an audit algorithm that learns to classify models as fair or unfair instead of defining a fairness metric. The process of learning this audit algorithm is subject to a fairness constraint that is weighted depending on the proportion of the population belonging to a particular group G.

Kearns et al. define $\alpha(G) = Pr(G)$ and reformulate fairness in (1) as

$$\alpha(G)|m(G) - m(\cdot)| < \epsilon \quad \forall\, G \tag{2}$$

Essentially, the addition of $\alpha(G)$ relaxes the original fairness metric of (1) depending on the proportion of G as a share of the overall population. The smaller G is, the more the condition is relaxed. As Kearns et al. explain, this addition is necessary to enable statistical estimation, given the increasing statistical uncertainty with decreasing group size. We discuss the implications in Section 4, and give the results of an empirical study regarding this formulation in Appendix E.

3.2 Foulds et al. and Morina et al.

Foulds et al. [3] provide an alternative approach based on ratios of model performance metrics. An expanded version of which is, in turn, given by Morina et al. [4].[5]

These definitions require that the ratio of some metric value between two groups be within a fixed interval. For example, suppose $m(G)$ measures the true positive rate (TPR) for subgroup G. Then the ϵ-differential intersectional definition of TPR parity (equal opportunity) given in [4] is that

$$e^{-\epsilon} \leq \frac{m(G)}{m(G')} \leq e^{\epsilon} \quad \forall\, G, G' \tag{3}$$

Morina et al. [4] note that $\epsilon = 0$ corresponds to "perfect fairness" ($m(G) = m(G')$).

3.3 Molina and Loiseau

Molina and Loiseau use a statistical approach to addressing intersectionality and fairness [5]. They call a classifier (ϵ, δ)-probably intersectionally fair if "the expected number of people that faces a discrimination more than ϵ is less than $n\delta$" (n is the population size).[6]

3.4 Cherian and Candès

Cherian and Candès [16] address fairness auditing for many subpopulations within the framework of hypothesis testing, as we do here. They use a bootstrap process to provide statistical performance bounds for many subpopulations at once. Our addition to this study is the illumination and discussion of desiderata (in Section 4), a clear description of how one can design fairness metrics using hypothesis testing (Section 5), and an empirical study showing that these metrics encourage (rather than discourage) the gathering of additional data to improve model performance (Section 6).

3.5 Khan et al. and Agrawal et al.

Khan et al. [17] consider metrics of fairness, accuracy, and variance for model estimators. They empirically show that there tends to be a tradeoff between these three values. In a similar vein, Agrawal et al. study debiasing methods, and in doing so show both theoretically and empirically that estimation variance tends to be higher in small subgroups [18]. Additionally, they prove results suggesting that partial debiasing results in both less variance and better fairness properties.

---

[5]We note that Foulds et al. (the same group as in [3]) also study the usage of Bayesian modeling to more accurately measure fairness metrics than point estimates [15]. While these Bayesian models for measuring fairness metrics may give more accurate estimates than point estimates, they do not allow for for the same kind of statistical analysis and ethical evaluation as a confidence interval (that we propose in Sections 4 and 5).

[6]Molina and Loiseau moreover highlight the issue of estimating fairness of a model on subgroups for whom the set of predicted values on that subgroup is a proper subset of the set of all predicted values. This becomes an issue because they use a ratio similar to that in Equation (3), which is undefined if $m(g') = 0$ for some group g'. Our models do not suffer from this issue; however, extremely tiny subgroups do come with their own statistical uncertainty issues, as we highlight in Section 5.

4 Desiderata

Although the problem of intersectionality is recognized in the literature, how difficult this problem is may not have been fully appreciated. At least some of the existing contributions violate minimal moral or methodological desiderata, as we shall see in Sections 4.1, 4.2, and 4.3 (and Appendix E).

A core tenet of building ethical algorithms is that machine-learned models need to be consistent with "human values," which can be formulated as desiderata. We see the following desiderata for intersectional fairness metrics.

4.1 Minimal Justice

A first desideratum we call "minimal justice." The idea is, roughly, that a standard of fairness should not be lower for certain groups, such as those historically targeted for discrimination or facing structural injustice. Intuitively, minimal justice is a form of minority protection that says "don't disadvantage the disadvantaged."

This desideratum is a weak form of prioritarianism. Recent work in algorithmic fairness has identified a similar prioritarian idea in "predictive justice" [19]. Whereas prioritarianism, a theory of distributive justice for well-being, demands that "benefitting people matters more the worse off these people are" [20], minimal justice requires only that those "worse off" should be given *at least the same* weight in aggregating a fairness metric. The desideratum does *not* require that greater weight be given to any group, and is hence met when a standard of fairness is identical for all groups.

To illustrate the desideratum, consider an example. Notwithstanding its merits, the proposal of Kearns et al. [2] may violate minimal justice. As noted above, the addition of $\alpha(G)$ in (2) relaxes the fairness constraint proportional to the size of a group. The smaller a group is (as a share of the data), the worse a model performance can be and still certify the model as fair. The fairness standard is hence lowered for small groups. On the assumption that these small groups include historically disadvantaged or oppressed groups, (2) violates minimal justice.

And drastically so: For a group G' that is c times smaller than group G (i.e. $\frac{\alpha(G)}{\alpha(G')} = c$), a model can be certified as "fair" if the disparity between the average performance and the performance for group G' is as much as c-times worse than it is for group G. Furthermore, for some value of ϵ there are groups that are proportionally so small that there is no model performance poor enough to certify the model as unfair. For example, if $\epsilon = .01$, for a binary classifier, any group whose proportion of the total population is less than ϵ is protected by essentially no fairness constraint at all.[7]

The ethical impact can be immense. A group might *look* relatively small in the data but be, in fact, large in absolute numbers in the population. Indeed, disadvantaged groups tend to be under-represented in data [21, 22]. Thus, the approach of Kearns et al. may lower the standard of fairness for precisely those groups that fairness is meant to protect.

4.2 Consistent Conceptualization

Any fairness metric operationalizes a certain idea, or concept, of fairness. A second desideratum is that fairness metrics should operationalize a concept of fairness consistently.

This desideratum may resemble that of Minimal Justice. But whereas Minimal Justice is a moral desideratum, Consistent Conceptualization is a methodological one. Minimal Justice *assumes* a standard of fairness as given (and requires that it not be lower for certain groups). Consistent Conceptualization ensures that this standard has minimal construct validity, i.e., that the formal operationalization of fairness represents the informal, intended conception of fairness (whichever that may be). The importance of construct validity for fairness is already established in the literature [23].

Typically, fairness metrics in algorithmic fairness operationalize the idea of *equality*. This is particularly evident in (1) which, for each group G, restricts the absolute disparity of $m(G)$ from overall

---

[7]The average model performance $m(G)$ and $m(\cdot)$ is constrained to be less or equal to 1. But the maximum deviation of accuracy in the binary setting is 1. Thus, even if the model is entirely inaccurate for this population and perfectly accurate for the rest of the population, the constraint is still satisfied.

mean performance $m(\cdot)$. This is one—albeit a very simple—way of operationalizing inequality (for alternatives see [24]). Likewise, (3) operationalizes fairness as equality [3, 4].[8]

Moreover, (1) and the definition by Foulds et al. and Morina et al. operationalize equality *consistently*. The fairness metrics apply an equality condition without bounds or exceptions.

Not so the proposal by Molina and Loiseau [5], which explicitly *bounds* equality. Effectively, the fairness measure permits that some small number of people faces severe discrimination, as long as the likelihood of discrimination or their relative size as a share of the overall population is small.[9] This fairness metric thus fails the desideratum of operationalizing the concept of equality consistently.

Typically fairness metrics, and all instances of (1), operationalize fairness as equality. Alternatives, well-known from distributive justice, include *prioritarianism*, stating that more of some good, such as model performance, should be given to those in greater need [20], and *sufficientarianism*, requiring that everyone has *enough* of some good (instead of the same) [25, 26].

4.3 Incentive Compatibility

The final desideratum starts with the recognition that metrics specify incentives. Anyone who wants to increase their models' fairness may want to maximize a fairness metric. The final desideratum thus requires that a fairness metric not have "perverse" incentives of two kinds: discouraging data collection and allowing "gaming."

First, a fairness metric should not discourage data collection. Any fairness metric that indicates greater *un*fairness only because further data are sampled from some group would fail to be incentive compatible. Likewise, inversely, any fairness metric would fail the desideratum that indicates greater *fair*ness only because data based on group identity are dropped.

The fairness metric (2), of Kearns et al., likely violates this desideratum of incentive compatibility. This is because collecting more data on a minority population G tightens the constraint by increasing $\alpha(G)$, thus making a certification of "fairness" at a given level of ϵ more difficult. Specifically, suppose that $m(G) = .15$ and $m(\cdot) = .85$. If $\alpha(G) = .01$, then the performance would be deemed "fair" for all $\epsilon > 0.7 \times 0.01 = .007$. However, if we collect more data for group G such that $\alpha(G) = .2$, then the model would be "fair" only for $\epsilon > 0.7 \times 0.2 = .014$. Unless the additional data results in material improvements to $m(G)$, for any ϵ such that $.007 < \epsilon < 0.014$, the fairness metric (2) would certify a given model as fair prior to further data collection, but as unfair afterwards. In short, under (2), fairness for hard-to-predict groups could be attained simply by under-representing them in the training data. We see this effect in our empirical study, described in Appendix E. We leave the details of this study to Appendix E, but the results show that the fairness metric suggested by Kearns et al. appears to indeed disincentivize additional data collection, violating *Incentive Compatibility*.

This is a "perverse" effect because, in practice, additional data collection about a minority group will help improve the model performance for that group. In other words, the metric gives an incentive to do the opposite of what it is meant to achieve.[10]

Whether other metrics [4, 3, 5] violate this desideratum depends on whether the estimated performance disparity is greater than the true disparity (which further data would likely help approximate). Fairness metrics that operationalize fairness as *equality* (e.g., as model performance disparity across groups), incentivize $m(G)$ to be nearly the same for all subgroups G. If the true model performance is nearly equal among groups, then these metrics incentive further data collection in order to have more accurate estimates of $m(G)$.

Second, a fairness metric should not encourage knowingly erroneous predictions. But some metrics (e.g., statistical or demographic parity) have exactly this property: Even if the label that we want to predict is known (which it generally, of course, isn't), "fairness" as these metrics define it can be improved by erroneous predictions. This is an undesirable property of fairness metrics [27].

---

[8]Compared to (1), (3) aims for equality between groups (as opposed to minimizing disparity with $m(\cdot)$), and measures *relative* disparity (a performance ratio instead of performance difference).

[9]Molina and Loiseau [5] write: "It can be seen for some given ϵ as a statement on the expected size of the population that is not being discriminated too much against."

[10]We do *not* contend that more data should be collected. Privacy considerations are important. Our point is instead that maintaining the appearance of a good fairness metric is a bad reason to not collect more data.

5 Two Alternative Metrics

We now illustrate how these desiderata can be met. We propose two alternative models, which we call the "optimist's" and "pessimist's model" respectively. Both define the problem using hypothesis testing. The optimist has the null hypothesis that the model is fair, and we have to prove it is not (similar to "innocent until proven guilty"); the pessimist inverts the "burden of proof" and has the null hypothesis that the model is unfair.[11]

5.1 Optimist's Model

We could formulate the problem of fairness for small groups as testing the joint hypothesis that

$$H_0 : m(G) > c \quad \forall\, G$$
$$H_1 : m(G) \le c \quad \exists\, G$$

Consider a group G of size n_G. Suppose $m(G)$ is accuracy. As a sample proportion, the standard error for our estimate of $m(G)$ is $\sqrt{\frac{m(G)(1-m(G))}{n_G}}$. Then, we would reject the null if the upper end of its confidence interval is less than c, i.e., if $m(G) + 1.64\sqrt{\frac{m(G)(1-m(G))}{n_G}} < c$ (ignoring multiple testing).[12] Under this formulation, we reject H_0 if $m(G)$ is sufficiently less than c, where "sufficiently less" has to do with our statistical power to detect that it is less. We would declare the model fair, if at given level c we cannot statistically reject that the model performs at least c well for all groups.

A minority population which is sufficient in number would easily reject the null if $m(G)$ is truly below c. Indeed, even with a population size of $n_G = 1000$, if $c = 0.7$, then a value of $m(G) < 0.67$ would reject the hypothesis that the model is fair.

5.2 Pessimist's Model

Depending on a model's deployment context, the optimistic approach might be problematic.[13] Consider instead the following pessimistic hypothesis test.

$$H_0 : m(G) < c \quad \exists\, G$$
$$H_1 : m(G) \ge c \quad \forall\, G$$

We would declare the model fair, if at a given level c we know with statistical certainty that the model performs at least c-well for all groups. In this case, (ignoring multiple testing again) we would require that $m(G) - 1.64\sqrt{\frac{m(G)(1-m(G))}{n_G}} > c$ for all G.

5.3 Fairness Metrics

The formulations can be extended from a hypothesis test to a fairness metric by finding the maximal c for which the respective null hypothesis cannot be rejected (for the optimist) or can be rejected (for the pessimist). In the optimist's model, choose the maximal c such that

$$c \le m(G) + 1.64\sqrt{\frac{m(G)(1 - m(G))}{n_G}} \tag{4}$$

for all relevant groups G. The fairness metric is the maximal c such that we cannot reject the hypothesis that the model performs at least c-well for all groups.

---

[11]Throughout, we assume that for the metric $m(\cdot)$ larger values are better (think accuracy, not error rates). Specifically, and without loss of generality, we use accuracy as our sample metric. This choice is for simplicity only; one could replace accuracy with any other metric for which higher values are preferred. For the hypothesis tests we describe, we use a z-score of 1.64, which corresponds to a 95% confidence interval for a one-sided hypothesis test. This is a conventional parameter choice and nothing in our argument depends on it.

[12]This ignores multiple hypothesis testing, which we address in Appendix B.

[13]Depending on the ethical risks involved in how a model is used, the more precautionary assumptions behind the pessimist's model might be more appropriate.

This metric can be read as saying that a model is "fair up to c." Intuitively, this means that, for all we know, the model performance $m(G)$ (say, accuracy) is likely as high as c for each group.

On the pessimist's model, we instead choose the maximal c such that

$$c \leq m(G) - 1.64\sqrt{\frac{m(G)(1 - m(G))}{n_G}} \tag{5}$$

for all relevant groups G. This fairness metric is the maximal c such that we reject the hypothesis that the model is *un*fair, that is, we reject that it does not perform at least c-well for each group.

This metric can be read as saying that a model is "unfair above c." The model likely performs at least c-well for each group; but for values above c, there likely is at least one group for which the model does not perform at least c-well—and we hence can't rule out that the model is unfair.

In summary, the fairness metrics are defined as bounds of the interval

$$\left(m(G) - 1.64\sqrt{\frac{m(G)(1 - m(G))}{n_G}}, m(G) + 1.64\sqrt{\frac{m(G)(1 - m(G))}{n_G}} \right)$$

This interval, of course, now has two interpretations. For one, it is the 90% confidence interval for the value of $m(G)$ for each G. Moreover, across all groups, it is also the interval in which we cannot reject the hypothesis that the model is unfair, nor can we reject the hypothesis that the model is fair.[14]

5.4 Discussion: Desiderata

Both metrics satisfy *Minimal Justice*. The bound c encodes a standard of fairness that is identical for all groups. Moreover, the relative size of groups doesn't matter. Whether a null hypothesis can be rejected changes with the absolute size of the group n_G (rather than the proportion $\frac{n_G}{n}$).

On the optimist's metric, for a small group, the difference between the actual (lower) model performance and the level up to which a model can be certified as fair might be large. But both of our metrics base their certification of "fairness up to c" on an aggregation that gives all groups the same weight. In fact, the pessimist's metric can be called "epistemically risk averse" insofar as it picks the *highest lower* bound out of all groups' confidence intervals (and hence is similar to the maximin decision rule).

On *Consistent Conceptualiztion*, both of our metrics conceptualize fairness as sufficiency. They understand fairness not as a matter of whether everyone has the same (as equality does), but whether everyone has *enough* [25, 26]. This idea is operationalized in (4) and (5) in a transparent and natural way: with an inequality. Moreover, the threshold c, what counts as "enough," is determined absolutely in the terms of model performance measure, and not depending on, e.g., how well the model performs on other groups. Thus, both of our metrics operationalize sufficiency consistently across all groups.

For *Incentive Compatibility* the picture is mixed: Both of our metrics discourage gaming (and thus satisfy Incentive Compatibility in this respect). This is because both fairness metrics determine (un)fairness as the highest (or lowest) expectable model performance across all groups. As such, improving model performance will never increase unfairness; and decreasing model performance will never increase fairness. In fact, decreasing model performance may lead to a decrease in fairness. It appears that operationalizing the idea of fairness as sufficiency is what makes our fairness metrics less susceptible to gaming—in particular, that the minimum level of model performance is defined in absolute terms and equally enforced for all groups.

But one of our metrics, namely the optimist's, may discourage further data collection (and thus violate Incentive Compatibility in its first respect). Because the optimist's model starts with the null hypothesis that a model is fair at a given c, gathering more data can make things "worse"; that is, with more data, we might come to reject the optimistic null hypothesis of fairness at a given c. A model might perform very poorly for certain groups, but we cannot reject the null hypothesis that the model is fair up to c, thanks to sparse data—and the metric thus results in an incentive to not sample more data but to instead "look the other way."

---

[14]The reader may go to Appendix A for an exploration on the impacts of changing m and n on these models.

6 Fairness Datasets Analysis

We evaluate empirically whether our metrics meet the desideratum of incentive compatibility. The question is: Do our metrics incentivize or disincentivize additional data collection?

To answer this question, we "simulate" additional data collection by experiment. We train models on increasingly larger subsamples of benchmark datasets and observe how metrics behave as the size of the training data increases. The behavior that we want to see is that the fairness metrics increase with the size of the training data sampled from the dataset. If, instead, a fairness metric *decreased* as greater shares of the dataset are sampled, the metric would *disincentivize* further data collection.

We seek to observe our metrics' behavior across the largest feasible range of benchmarks. To achieve this, we use lale, a Python library created by IBM [28]. Lale allows for the creation of consistent automated machine learning models across 20 well-known "fairness datasets" that can easily be fetched, modeled, and evaluated [29]. These datasets are all tabular with a categorical target variable. They each come with "fairness metadata," which includes protected attributes, along with ranges/values of those attributes that correspond to the privileged group.[15] Details on the methods of our analysis are in Appendix C. Here we only discuss the main result on testing whether our metrics incentivize against data collection.

For each of the datasets, we observe model performance $m(G)$, as well the optimist's c_1^g and the pessimist's fairness metric c_2^g respectively. For ease of interpretation we use accuracy as model performance; neither our results nor their interpretation depend on this.

We ran two versions of this experiment. In one version, we subsample the entire dataset of each benchmark; whereas in another, we subsample only on the *critical subgroup*, which is the group that is right on the c threshold. The first version simulates additional data collection for *all* groups, whereas the latter for those groups that "drag down" the fairness metric. Here we concentrate on results from subsampling on the critical subgroup only, shown in Figure 1.[16] Full results for both versions are in Appendix C.4.
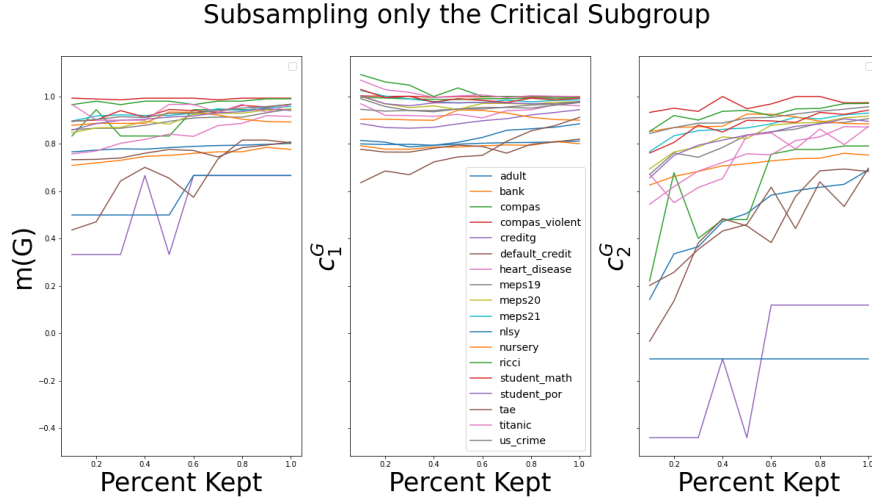

Subsampling only the Critical Subgroup

Figure 1: Plots of accuracy $m(G)$, optimist's metric c_1^g, and pessimist's metric c_2^g of critical subgroups G for each dataset. The x-axis corresponds to the percentage of the critical subgroup that is kept. Legend lists the dataset name.

---

[15]While there have been critiques of the usage of some of these datasets [30] [31], they are still appropriate for the purpose of testing whether our proposals incentivize or disincentivize the collection of additional data.

[16]Some of the plots are disconnected. This is because sometimes the subsampling of the dataset did not include any members of the critical subgroup; in those cases, the model could not predict for that subgroup, so no accuracy measurement could be taken. The most erratic curves (curves of $m(G)$ and c_2 for the `creditg` and `nlsy` datasets) correspond to either a subgroup of size 1 or 2.

The thing to note here is that there is a trend upwards in each of these plots. Most notably, the middle plot, on the optimist's metric c_1^g shows this upward trend.[17] This suggests that our optimist's metric—at least for the datasets tested—does *not* pose perverse incentives. The further we go on the x-axis (representing more data being "collected"), the model performance as well as the fairness metrics tend to improve.

Consider for example the behavior of the optimist's metric for the model trained on increasing amounts of data from the `tae` dataset (brown line that "starts" lowest in middle figure). Although the metric does not strictly increase as the training is based on greater data (the metric decreases slightly from 20% to 30% of data used), it shows a very strong upward trend.

7 Conclusion

Although the general idea of intersectionality seems easy to state, putting intersectionality to work in quantitative social science is, generally, far from straight-forward [32]. Likewise, intersectionality presents a problem for algorithmic fairness: Intersectionality requires estimating statistical properties across subgroups that are increasingly small, which gives rise to statistical as well as moral-methodological challenges.

Statistically, small groups are a challenge for estimation. As statistical uncertainty increases (due to more and smaller groups), the point estimates of model performance for these groups become meaningless. Any approach of intersectional fairness needs to account for statistical uncertainty. But some existing metrics do not seem to fully appreciate the moral-methodological challenges that underlie this problem and "lower the fairness bar" for smaller groups, i.e., the metrics violate desiderata such as Minimal Justice or Consistent Conceptualization.

With this paper, we elucidate this intersectionality problem for algorithmic fairness: We develop minimal desiderata to clarify the moral-methodological challenges underlying this problem; we argue that some existing fairness metrics fail these desiderata, but illustrate that the desiderata can be met. We propose fairness metrics that rely on hypothesis testing (instead of performance point estimates) and that understand fairness as sufficiency (instead of equality). On these proposed metrics, fairness is understood as a certain minimum level of expected model performance that is, for all we know, likely enjoyed by all groups. We empirically evaluate the metrics against the proposed desiderata, including on 18 datasets that are widely used for fairness benchmarks.

In light of their technical and normative-theoretical limitations, the metrics we propose should be seen as illustrations. Technically, the simple hypothesis testing needs to be extended to multiple hypothesis testing to allow for interdependent subgroup memberships (see Appendix B). Normative-theoretically, the desiderata that we develop are not exhaustive and they do not uniquely characterize the metrics we propose.

Nevertheless, overall, our findings extend the list of problems that statistical uncertainty raises for algorithmic fairness. Previous work observed that fairness metrics are biased: They "fail to account for statistical uncertainty . . . exaggerating the extent of performance disparities" between groups where such disparities exist and indicating disparities "in cases where model performance is . . . identical across groups" [6]. Our present findings add that with increasing statistical uncertainty fairness metrics risk becoming either nonsensical (if they aggregate point estimates) or morally inadequate (if they "lower the fairness bar" to enable statistical estimation).

However, we also offer ways of advancing the literature on algorithmic fairness: with desiderata that clarify the challenges at hand and guide the search for solutions, and with fairness metrics that suggest novel avenues for defining such metrics based on hypothesis testing and fairness as sufficiency.

Acknowledgements

---

[17]Some datapoints "overshoot" on the y-axis with values > 1, suggesting a negative trend, e.g., for the `compas` dataset (green line). But this behavior is an artifact of the standard way of calculating the confidence interval.

Veomett was in residence at the Simons Laufer Mathematical Sciences Institute (formerly MSRI) in Berkeley, California, during the Fall 2023 semester.

References

[1]   Joy Buolamwini and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification". In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Ed. by Sorelle A. Friedler and Christo Wilson. Vol. 81. Proceedings of Machine Learning Research. PMLR, 2018, pp. 77–91. URL: https://proceedings.mlr.press/v81/buolamwini18a.html.

[2]   M. Kearns et al. "Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness". In: *The 35 th International Conference on Machine Learning*. 2018.

[3]   James R. Foulds et al. "An Intersectional Definition of Fairness". In: *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. 2020, pp. 1918–1921. DOI: 10.1109/ICDE48307.2020.00203.

[4]   Giulio Morina et al. "Auditing and Achieving Intersectional Fairness in Classification Problems". In: *CoRR* abs/1911.01468 (2019). arXiv: 1911.01468. URL: http://arxiv.org/abs/1911.01468.

[5]   Mathieu Molina and Patrick Loiseau. "Bounding and Approximating Intersectional Fairness through Marginal Fairness". In: *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*. Ed. by Sanmi Koyejo et al. 2022. URL: http://papers.nips.cc/paper%5C_files/paper/2022/hash/6ae7df1f40f5faeda474b36b61197822-Abstract-Conference.html.

[6]   Kristian Lum, Yunfeng Zhang, and Amanda Bower. "De-biasing "bias" measurement". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. New York, NY, USA: Association for Computing Machinery, June 2022, pp. 379–389. ISBN: 978-1-4503-9352-2. DOI: 10.1145/3531146.3533105. URL: https://dl.acm.org/doi/10.1145/3531146.3533105 (visited on 10/16/2023).

[7]   European Union. "Regulation (EU) 2021/0106 of the European Parliament and of the Council: Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Ammending Certain Union Legislative Acts". In: *Official Journal of the European Union* (2021).

[8]   European Union. *Recital 49 of the EU AI Act*. https://artificialintelligenceact.eu/recital/49/. 2021.

[9]   Center for Intersectional Justice. *Intersectionality at a Glance in Europe*. https://www.intersectionaljustice.org/img/2020.4.14_cij-factsheet-intersectionality-at-a-glance-in-europe_du2r4w.pdf. 2020.

[10] Sam Corbett-Davies et al. "The measure and mismeasure of fairness". In: *J. Mach. Learn. Res.* 24.1 (Mar. 6, 2024). ISSN: 1532-4435.

[11] Youjin Kong. "Are "Intersectionally Fair" AI Algorithms Really Fair to Women of Color? A Philosophical Analysis". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. New York, NY, USA: Association for Computing Machinery, June 20, 2022, pp. 485–494. ISBN: 978-1-4503-9352-2. DOI: 10.1145/3531146.3533114. URL: https://dl.acm.org/doi/10.1145/3531146.3533114 (visited on 11/05/2024).

[12] Angelina Wang, Vikram V Ramaswamy, and Olga Russakovsky. "Towards Intersectionality in Machine Learning: Including More Identities, Handling Underrepresentation, and Performing Evaluation". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. New York, NY, USA: Association for Computing Machinery, June 20, 2022, pp. 336–349. ISBN: 978-1-4503-9352-2. DOI: 10.1145/3531146.3533101. URL: https://dl.acm.org/doi/10.1145/3531146.3533101 (visited on 11/05/2024).

[13] Harini Suresh et al. "Towards Intersectional Feminist and Participatory ML: A Case Study in Supporting Feminicide Counterdata Collection". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. New York, NY, USA: Association for Computing Machinery, June 20, 2022, pp. 667–678. ISBN: 978-1-4503-9352-2. DOI:

10.1145/3531146.3533132. URL: https://dl.acm.org/doi/10.1145/3531146.3533132 (visited on 11/05/2024).

[14]   Goda Klumbytė, Claude Draude, and Alex S. Taylor. "Critical Tools for Machine Learning: Working with Intersectional Critical Concepts in Machine Learning Systems Design". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. New York, NY, USA: Association for Computing Machinery, June 20, 2022, pp. 1528–1541. ISBN: 978-1-4503-9352-2. DOI: 10.1145/3531146.3533207. URL: https://dl.acm.org/doi/10.1145/3531146.3533207 (visited on 11/05/2024).

[15]   James R. Foulds et al. "Bayesian Modeling of Intersectional Fairness: The Variance of Bias". In: *Proceedings of the 2020 SIAM International Conference on Data Mining (SDM)*. Society for Industrial and Applied Mathematics, 2020, pp. 424–432. DOI: 10.1137/1.9781611976236.48. eprint: https://epubs.siam.org/doi/pdf/10.1137/1.9781611976236.48. URL: https://epubs.siam.org/doi/abs/10.1137/1.9781611976236.48.

[16]   John Cherian and Emmanuel Candés. "Statistical Inference for Fairness Auditing". In: *arXiv* https://arxiv.org/pdf/2305.03712.pdf (2023).

[17]   Falaah Arif Khan, Denys Herasymuk, and Julia Stoyanovich. *On Fairness and Stability: Is Estimator Variance a Friend or a Foe?* 2023. arXiv: 2302.04525 [cs.LG]. URL: https://arxiv.org/abs/2302.04525.

[18]   Ashrya Agrawal et al. *Debiasing classifiers: is reality at variance with expectation?* 2021. arXiv: 2011.02407 [cs.LG]. URL: https://arxiv.org/abs/2011.02407.

[19]   Seth Lazar and Jake Stone. "On the Site of Predictive Justice". en. In: *Noûs* n/a.n/a (). ISSN: 1468-0068. DOI: 10.1111/nous.12477. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/nous.12477 (visited on 08/29/2023).

[20]   Derek Parfit. "Equality and Priority". In: *Ratio* 10.3 (Dec. 1997), pp. 202–221. ISSN: 0034-0006. DOI: 10.1111/1467-9329.00041. URL: http://www.blackwell-synergy.com/links/doi/10.1111%2F1467-9329.00041.

[21]   Jonas Lerman. "Big Data and Its Exclusions". en. In: *Stanford Law Review* (Sept. 2013). URL: https://www.stanfordlawreview.org/online/privacy-and-big-data-big-data-and-its-exclusions/ (visited on 08/17/2019).

[22]   Sarah Giest and Annemarie Samuels. "'For good measure': data gaps in a big data world". en. In: *Policy Sciences* (Apr. 2020). ISSN: 1573-0891. DOI: 10.1007/s11077-020-09384-1. URL: https://doi.org/10.1007/s11077-020-09384-1 (visited on 05/15/2020).

[23]   Abigail Z. Jacobs and Hanna Wallach. "Measurement and Fairness". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Mar. 2021). arXiv: 1912.05511, pp. 375–385. DOI: 10.1145/3442188.3445901. URL: http://arxiv.org/abs/1912.05511 (visited on 05/08/2021).

[24]   Amartya Sen. *On economic inequality*. Enl. ed., Oxford: Clarendon Press, 1997. ISBN: 978-0-19-829297-5.

[25]   Harry Frankfurt. "Equality as a Moral Ideal". In: *Ethics* 98.1 (Oct. 1987), pp. 21–43. ISSN: 00141704. DOI: 10.1086/292913. URL: http://www.jstor.org/stable/2381290 (visited on 05/31/2010).

[26]   Michael A. Slote. *Beyond Optimizing: A Study of Rational Choice*. en. Cambridge Mass.: Harvard University Press, 1989. ISBN: 978-0-674-06918-3.

[27]   Cynthia Dwork et al. "Fairness through awareness". In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ITCS '12. New York, NY, USA: Association for Computing Machinery, Jan. 2012, pp. 214–226. ISBN: 978-1-4503-1115-1. DOI: 10.1145/2090236.2090255. URL: https://dl.acm.org/doi/10.1145/2090236.2090255 (visited on 03/28/2024).

[28]   G. Baudart et al. "Lale: Consistent Automated Machine Learning". In: *AutoML Workshop at KDD*. 2020.

[29]   M. Hirzel and M. Feffer. *A Suite of Fairness Datasets for Tabular Classification*. https://arxiv.org/pdf/2308.00133.pdf. 2023.

[30]   Frances Ding et al. "Retiring Adult: New Datasets for Fair Machine Learning". In: *Advances in Neural Information Processing Systems*. Ed. by A. Beygelzimer et al. 2021. URL: https://openreview.net/forum?id=bYi_2708mKK.

[31] Michell Bao et al. "It's COMPASlicated: The Messy Relationship between RAI Datasets and Algorithmic Fairness Benchmarks". In: *NeurIPS Datasets and Benchmarks*. 2021. URL: https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/92cc227532d17e56e07902b254dfad10-Paper-round1.pdf.

[32] Liam Kofi Bright, Daniel Malinsky, and Morgan Thompson. "Causally Interpreting Intersectionality Theory". en. In: *Philosophy of Science* 83.1 (Jan. 2016). Publisher: Cambridge University Press, pp. 60–81. ISSN: 0031-8248, 1539-767X. DOI: 10 . 1086 / 684173. URL: https : / / www . cambridge . org / core / journals / philosophy - of - science / article / causally - interpreting - intersectionality - theory / E78BB6C33D0D7DF4316FCD3687912258 (visited on 02/12/2024).

[33] IBM/lale. *Lale Fairness Dataset Sample Notebook*. https://github.com/IBM/lale/blob/master/examples/demo_fairness_datasets.ipynb. 2023.

A Discussion: Impact of n and m on Each Model

To give the reader a feel for the mathematical impact of the choice between these two models, we share some hopefully informative plots in Figure 2.
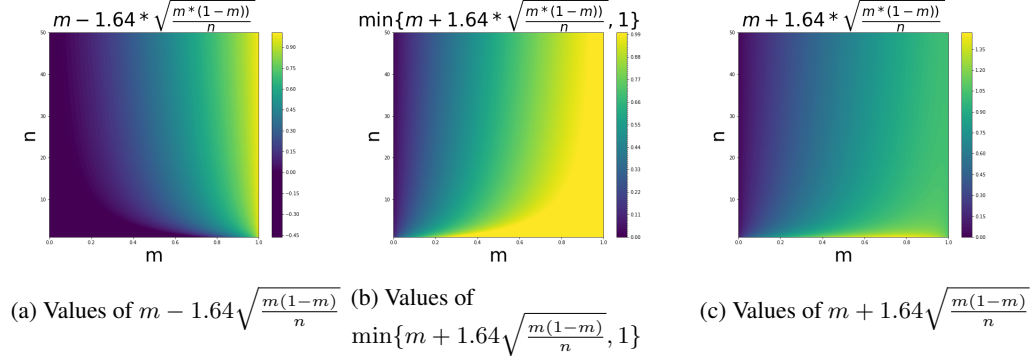

(a) Values of $m - 1.64\sqrt{\frac{m(1-m)}{n}}$

(b) Values of $\min\{m + 1.64\sqrt{\frac{m(1-m)}{n}}, 1\}$

(c) Values of $m + 1.64\sqrt{\frac{m(1-m)}{n}}$

Figure 2: Showing the relationship between m (metric), n (number in subgroup), and c (edge of confidence interval). Hues in 2a shows the values of c in the pessimist's model. Hues in 2b shows the values of c in the optimist's model, but with an upper limit of 1 (since no proportion can be larger than 1). Hues in 2c shows the values of c in the optimist's model, without limiting the value at 1 (so that we can see more easily where it is very difficult to reject the optimist's hypothesis that the model is fair).

The horizontal axis of each of these plots is m, the metric value, which is assumed to be a proportion for which higher values are preferred (such as accuracy). The vertical axis is n, the size of the subgroup. The hue at (m, n) in Figure 2a is the corresponding value of $m - 1.64\sqrt{\frac{m(1-m)}{n}}$. Here we can visually see that, for a fixed metric value m, the subgroup size must be reasonably large in order to reject the hypothesis that the model is "unfair above c" for c near m.

Similarly, the hue at (m, n) in Figure 2b is the corresponding value of $m + 1.64\sqrt{\frac{m(1-m)}{n}}$, capped at a value of 1 (since no proportion can be larger than 1). Here we can visually see that, for a fixed metric value m, the subgroup size must be reasonably large in order to reject the hypothesis that the model is "fair up to c" for c near m. To further understand the impact of small groups in this optimist's model, we include Figure 2c. In Figure 2c, the hue simply gives the value of $m + 1.64\sqrt{\frac{m(1-m)}{n}}$, even if it is larger than 1. This plot further highlights the fact that, in the optimist's model, it is very difficult to reject the hypothesis that the model is perfectly fair for very small subgroups.

B Limitations

We note that the issue of multiple hypothesis testing is one which we do not address in depth. If membership in the different groups in question is independent, one can use the Bonferroni correction to address the multiple hypothesis tests. Under this strict type of multiple hypothesis testing, the p-values that are calculated are using significance level $\frac{\alpha}{n}$, where n is the number of hypotheses that we are testing. This correction guarantees that the probability that we reject *one or more* null hypotheses is no more than α. Considering overlapping subgroups (such as considering fairness both for Black Women and for Latina Women) requires more care, and we do not delve into the issue of overlapping subgroups here. We thus, effectively, assume—counterfactually—that each person is a member of exactly one group. For the purposes of our empirical study (below), we fix the number of protected attributes to be as large as possible, as described in Section C.1.

C Methods

We here provide further details on our empirical methods.

For starters, we choose the lale library and its accompanying datasets for two reasons:

1. The number of "fairness datasets" in the lale library is larger than any other conglomeration of fairness datasets that we are aware of.

2. Because the lale library has built-in models, we can apply a consistent type of model to each dataset, so that our experiments are not muddied by differing model constructions.

C.1   Subgroup Identification

The models we created use a forest of boosted trees from the XGBoost library; the functions to easily create these models are also part of the lale library. We created three models using the lale pipeline, using 3-fold cross-validation. The three models can be accessed to evaluate their accuracy on various subgroups. However, since lale requires sklearn version 1.2, we do not have access to the train/test indices of each of the models. Thus, to evaluate the accuracy on group G, we do so on all of the members of G in the dataset.[18]

The set of subgroups G on which we calculated the model accuracy come in part from the fairness data that lale provides, and also from attributes that are well-understood to be sensitive. Specifically, all of the attributes that the lale library lists as "protected" are included in our master list of protected attributes. If the rows in the dataset correspond to individuals, and any of $\{\text{age}, \text{sex}, \text{race}\}$ were not in lale's list of protected attributes, we added them to the master list. From this master list, we created *all* subgroups using *all* categories in the master list. For example, if a dataset had race, sex, and age category, we included in G each triple (r, s, a), where r was a race in that dataset's race column, s was a sex in that dataset's sex column, and a was an age category for that dataset.

C.2   Data Pre-processing

For each of the 20 fairness datasets, we used the built-in lale data pre-processing with small adjustments.

We used the simple methods for imputing missing data which are provided with the sample notebook at [33].

In order to use XGBoost, we needed to change some of the predicted categories to integer type.

In order to make the results more understandable, we re-named some of the categories (for example, changing the 'sex' categories from 0/1 to male/female).

The "race" categories in the `nlsy` dataset were atypical, including both categories such as 'GERMAN' and 'BLACK.' We did not attempt to clean that data but left the categories as given.

We created groupings by age for those datasets that don't already come with age groupings (see Appendix C.3).

C.3   Age Grouping

For the age attribute, some of the datasets already come with age groupings. In those cases, we directly used those groupings as the age categories. For the datasets where age was a strictly numerical attribute, we used the following heuristic to create categories:

- If age was already listed by lale as a protected attribute, we used the ranges provided by lale (for priviledged/unpriviledged groups) to create the categories.

- If age was not already listed as a protected attribute:
    - We grouped by decade in all datasets where this produced at least 5 people of each decade.
    - The `law_school` dataset had fewer than 5 members of the [0,9] decade, and fewer than 5 members of the [10, 19] decade, so those were grouped into a 0-19 group

---

[18]Ideally, we would like to evaluate only on the members of G in the test set for that fold. However, our goal here is to assess our two proposed ideas to address small-sized subgroups, not to assess true model accuracy. Averaging the subgroup accuracy across the three folds provides appropriate information to do that. Thus, for this analysis, we set $m(G)$ to be the average accuracy of the three models for subgroup G.

After this initial analysis, we tossed out two of the datasets: `law_school` and `speeddating`. The standard lale models created by XGBoost were 100% accurate on those models, and thus did not provide interesting analysis for us.[19]

C.4 Analysis

For each such subgroup $G \in \mathcal{G}$, we calculated $m(G)$: the average accuracy of the three models on that subgroup. We then calculate the c values associated with each of those subgroups; indexed by c_1 for the optimist's and c_2 for the pessimist's metric. Specifically, for group G we calculate

$$c_1^G = m(G) + 1.64\sqrt{\frac{m(G)(1 - m(G))}{n_G}}$$

from the optimist's model and

$$c_2^G = m(G) - 1.64\sqrt{\frac{m(G)(1 - m(G))}{n_G}}$$

from the pessimist's model.

Once these are calculated for all subgroups, we calculate

$$acc_{min} = \min\{m(G) : G \in \mathcal{G}\}$$
$$c_1 = \min\{c_1^G : G \in \mathcal{G}\}$$
$$c_2 = \min\{c_2^G : G \in \mathcal{G}\}$$

We also find their corresponding subgroups:

$$G_{min\_acc} = \operatorname{argmin}\{m(G) : G \in \mathcal{G}\}$$
$$G_1 = \operatorname{argmin}\{c_1^G : G \in \mathcal{G}\}$$
$$G_2 = \operatorname{argmin}\{c_2^G : G \in \mathcal{G}\}$$

The group G_{min_acc} is the group with minimum estimated accuracy, while group G_1 (G_2) is on the cusp of rejecting the hypothesis that the model is fair (not being able to reject the hypothesis that the model is unfair) up to accuracy c_1 (c_2). Thus, we call groups G_{min_acc}, G_1, and G_2 the *critical subgroups* for a dataset. For some datasets, there are three distinct critical subgroups, while for other datasets, some of the critical subgroups are the same; see Tables 1, 2, 3, and 4 in Appendix D for details.

Once we had the (up to) three critical subgroups of each dataset, we did two additional analyses.

C.4.1 Subsample Just the Critical Group

Suppose G is a critical subgroup of a dataset. We then created 10 models (each a set of three 3-fold cross-validated models), where we include 10%, 20%, ..., 100% of the subgroup in the dataset used to create the model. We then evaluated that group's critical value (whether it be $m(G)$, c_1^G, or c_2^G) on each of those 10 models, to see how those values change. The intention here is to mimic increasing samples from just the critical group, and how that additional data collection impacts the fairness evaluation of the model. These results of this analysis were in Figure 1.

C.4.2 Subsample the Entire Dataset

Suppose G is a critical subgroup of a dataset. We also created 10 models (each a set of three 3-fold cross-validated models), where we included 10%, 20%, ..., 100% of the entire dataset to create the

---

[19]We suspect that these datasets might be included in lale's list because they have low scores on other fairness metrics, such as the "symmetric class imbalance" metric in the sample notebook at [33], or because the model must use protected attributes in order to be accurate.

model. We then evaluated that group's critical value (whether it be $m(G)$, c_1^G, or c_2^G) on each of those 10 models, to see how those values change. The intention here is to mimic increasing sampling overall, and how that additional data collection impacts the fairness evaluation of the model. We note that, for the `nursery` dataset, one of the predicted categories (recommend) had only two data points with that category. In order for XGBoost to successfully create a model, we needed to add back both of those two data points into each subsample (if they had been removed in that random subsample). The results of this analysis are in Figure 3.
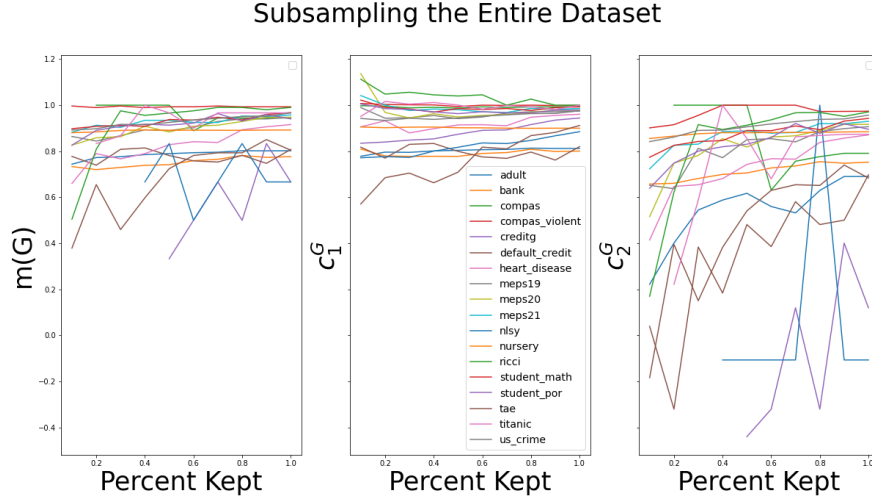

Figure 3: Plots of $m(G), c_1^G$, and c_2^G of critical subgroups G for each dataset. Here we subsampled the entire dataset, and the x-axis corresponds to the percentage of the entire dataset that is kept. Legend lists the dataset name.

D Tables

Dataset	Min $m(G)$ Subgroup	Min c_1 Subgroup	Min c_2 Subgroup
adult	(White, Male, 50's)	(White, Male, 50's)	(Other, Male, 60's)
bank	<=24	<=24	<=24
compas	(Male, Native American, 25 - 45)	(Male, African-American, 25 - 45)	(Male, Native American, 25 - 45)
compas_violent	(Female, African-American, <25)	(Male, African-American, 25-45)	(Male, Other, >45)
creditg	(male div/sep, male, <=25)	(male single, male, >25)	(male div/sep, male, <=25)
default_credit	(male, 40's)	(male, 40's)	(female, 70's)
heart_disease	(female, >54)	(female, >54)	(female, >54)
meps19	(White, 80's, female)	(White, 80's, female)	(Non-White, 80's, male)
meps20	(Non-White, 80's, female)	(Non-White, 80's, female)	(Non-White, 80's, female)
meps21	(Non-White, 80's, female)	(Non-White, 80's, female)	(Non-White, 80's, female)
nlsy	(Female, <18, GREEK)	(Female, >=18, GERMAN)	(Female, <18, HAWAIIAN)
nursery	great_pret	great_pret	great_pret
ricci	W	B	W
student_math	(M, <18)	(M, <18)	(M, <18)
student_por	(M, >=18)	(M, <18)	(M, >=18)
tae	1.0	2.0	1.0
titanic	(female, 60's)	(female, 30's)	(female, 60's)
us_crime	TRUE	TRUE	TRUE

Table 1: Subgroups with minimum $m(G)$, c_1, or c_2 for each dataset.

Dataset	Subgroup	Subgroup Category	n	$m(G)$
adult	(White, Male, 50's)	[race, sex, age_cat]	4256	0.8020050125313280
bank	<=24	age_cat	809	0.7766790276060980
compas	(Male, Native American, 25 - 45)	[sex, race, age_cat]	6	0.9444444444444450
compas_violent	(Female, African-American, <25)	[sex, race, age_cat]	95	0.9929824561403510
creditg	(male div/sep, male, <=25)	[personal_status, sex, age_cat]	2	0.6666666666666670
default_credit	(male, 40's)	[sex, age_cat]	2771	0.8078912546613740
heart_disease	(female, >54)	[sex, age_cat]	103	0.9158576051779940
meps19	(White, 80's, female)	[RACE, age_cat, SEX]	184	0.947463768115942
meps20	(Non-White, 80's, female)	[RACE, age_cat, SEX]	146	0.9474885844748860
meps21	(Non-White, 80's, female)	[RACE, age_cat, SEX]	142	0.9577464788732400
nlsy	(Female, <18, GREEK)	[gender, age_cat, race]	2	0.666666666666667
nursery	great_pret	parents	4320	0.8922839506172840
ricci	W	race	68	0.9901960784313730
student_math	(M, <18)	[sex, age_cat]	134	0.9676616915422890
student_por	(M, >=18)	[sex, age_cat]	73	0.9406392694063930
tae	1.0	whether_of_not _the_ta_is_a_native _english_speaker	29	0.8045977011494250
titanic	(female, 60's)	[sex, age_cat]	10	0.9666666666666670
us_crime	TRUE	blackgt6pct	970	0.9663230240549830

Table 2: Subgroups with minimum accuracy value $m(G)$

Dataset	Subgroup	Subgroup Category	n	c_1
adult	(White, Male, 50's)	[race, sex, age_cat]	4256	0.8120224969943970
bank	<=24	age_cat	809	0.8006925092362380
compas	(Male, African-American, 25 - 45)	[sex, race, age_cat]	1563	0.994278290695671
compas_violent	(Male, African-American, 25 - 45)	[sex, race, age_cat]	932	0.999219907516058
creditg	(male single, male, >25)	[personal_status, sex, age_cat]	492	0.94429531762294
default_credit	(male, 40's)	(sex, age_cat)	2771	0.820164957561691
heart_disease	(female, >54)	(sex, age_cat)	103	0.96071630150011
meps19	(White, 80's, female)	[RACE, age_cat, SEX]	184	0.974437789794083
meps20	(Non-White, 80's, female)	[RACE, age_cat, SEX]	146	0.977763384001414
meps21	(Non-White, 80's, female)	[RACE, age_cat, SEX]	142	0.985432236390149
nlsy	(Female, >=18, GERMAN)	[gender, age_cat, race]	179	0.88480628727837
nursery	great_pret	parents	4320	0.9000195456124770
ricci	B	race	27	1.0
student_math	(M, <18)	(sex, age_cat)	134	0.992723469193729
student_por	(M, <18)	(sex, age_cat)	193	0.978258629541195
tae	2.0	whether_of_not _the_ta_is_a_native _english_speaker	122	0.912074688695555
titanic	(female, 30's)	(sex, age_cat)	86	0.99964644295967
us_crime	TRUE	blackgt6pct	970	0.975822194666121

Table 3: Subgroups with minimum c_1 value

Dataset	Subgroup	Subgroup Category	n	c_2
adult	(Other, Male, 60's)	[race, sex, age_cat]	10	0.6903719639038720
bank	<=24	age_cat	809	0.7526655459759590
compas	(Male, Native American, 25 - 45)	[sex, race, age_cat]	6	0.7910815916767240
compas_violent	(Male, Other, Greater than 45)	[sex, race, age_cat]	49	0.9739395574376140
creditg	(male div/sep, <=25)	[personal_status, age_cat]	2	0.1200000000000000
default_credit	(female, 70's)	[sex, age_cat]	12	0.6973855176357850
heart_disease	(female, >54)	[sex, age_cat]	103	0.8709989088558780
meps19	(Non-White, 80's, male)	[RACE, age_cat, SEX]	67	0.9066848240754210
meps20	(Non-White, 80's, female)	[RACE, age_cat, SEX]	146	0.9172137849483570
meps21	(Non-White, 80's, female)	[RACE, age_cat, SEX]	142	0.9300607213563300
nlsy	(Female, <18, HAWAIIAN)	[sex, age_cat, race]	1	-0.10643674743062500
nursery	great_pret	parents	4320	0.8845483556220900
ricci	W	race	68	0.9706008691220500
student_math	(M, <18)	[sex, age_cat]	134	0.9425999138908480
student_por	(M, >=18)	[sex, age_cat]	73	0.8952823458833590
tae	1.0	whether_of_not _the_ta_is_a_native _english_speaker	29	0.6838443819651760
titanic	(female, 60's)	[sex, age_cat]	10	0.8735726878662690
us_crime	TRUE	blackgt6pct	970	0.9568238534438450

Table 4: Subgroups with minimum c_2 value

E   Analysis of Metric from Kearns et al.

As noted in Section 4.3, we hypothesize that the fairness metric outlined by Kearns et al. [2] violates *Incentive Compatibility*. The fairness metric likely "looks worse" as additional data is gathered about a small subgroup (i.e., a group whose size in proportion to the entire dataset is small). The fairness metric includes a factor which is the proportion of the subgroup within the dataset. Thus, as additional data is collected from that subgroup alone, this proportion increases, making the model more likely to violate the fairness criteria, hence potentially disincentivizing additional data collection on that subgroup. Here we empirically examine this hypothesis.

E.1   Study Description

Using the same datasets, subgroups, pre-processing, cleaning, and models outlined in Appendix C, we calculate the value of the following expression from Equation (2):

$$\alpha(G)|m(G) - m(\cdot)| \tag{6}$$

Recall that $\alpha(G)$ is the proportion of group G within the total population, and that m is some model performance metric (as in Appendix C, we use accuracy as our sample metric m for this study). The value $m(G)$ is the model performance metric evaluated only on subgroup G, while $m(\cdot)$ is the value of the model performance metric on the entire dataset.

Kearns et al. [2] use an auditing process wherein the value calculated from expression (6) must be below some threshold ϵ in order for a model to be considered fair. Thus, we can think of expression (6) as describing *unfairness* for group G.[20]

E.2   Methods

We calculate the value in expression (6) on increasing subsamples of each dataset. We concentrate on small sugbroups G that comprise no more than 10% of the total population. Just as in the experiments described in Appendix C, we examine two subsampling scenarios: We subsample the subgroup in question only (to simulate gathering more subgroup data), and we subsample the entire dataset (to simulate gathering more population data). Note that, of course, the values of m depend on the model created, which depends on the subsample of the data used to create the model.

Many of the datasets (`heart_disease`, `nursery`, `ricci`, `student_math`, `student_por`, `tae` and `us_crime`) don't have any subgroups comprising less than 10% of the total population, and thus we exclude those datasets from this analysis. From the other datasets, we concentrate on four: the `adult, bank, meps20`, and `titanic` datasets. The results for all other datasets are similar.

In Section 4.3, we hypothesize that the protocol of Kearns et al. violates *Incentive Compatibility*. Specifically, we hypothesize that when subsampling only small groups, the *unfairness* value of expression (6) would *increase*. Since the value $\alpha(G)$ does not change significantly when subsampling the entire population, we do not expect expression (6) to change much when subsampling the entire dataset, aside from the fact that potentially a better model might make expression (6) decrease.

E.3   Results and Discussion

The results of subsampling just the small subgroup can be found in Figure 4, and the results of subsampling the entire dataset can be found in Figure 5.

When only the small subgroup is subsampled (as in Figure 4), we see the value of (6) increasing for all subgroups in the `adult` and `bank` datasets. The picture is slightly more muddled in the `meps20` and `titanic` datasets, but these still show either a consistent increase or an initial increase for nearly all of the small subgroups. In other words, the value (6) of unfairness *increases*, indicating that the Kearns et al. auditing process discourages additional data collection of small subgroups, and thus violates *Incentive Compatibility*.

When the entire dataset is subsampled (as in Figure 5), values of (6) remain remarkably consistent in the `adult` dataset, and tend to decrease in the `bank, meps20`, and `titanic` datasets. We can thus

---

[20]All typical ways of defining "fairness" can be interpreted this way. A higher ϵ in (1) is interpreted as a *de*crease in fairness and thus an increase in unfairness.
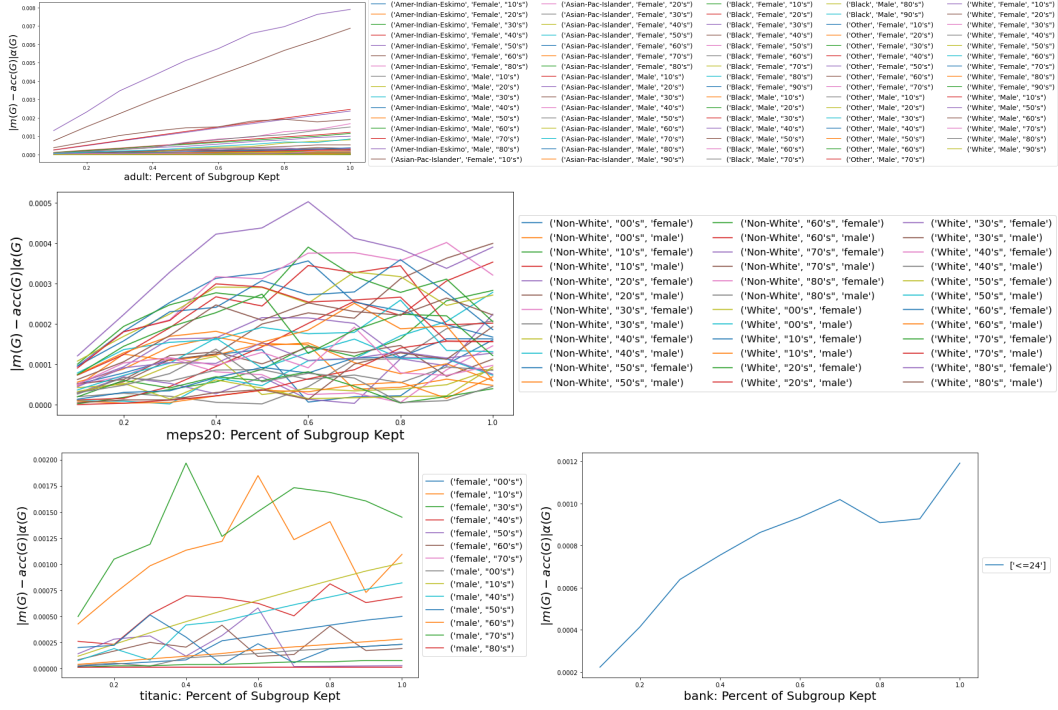
Figure 4: Values of expression (6) on the `adult`, `meps20`, `titanic`, and `bank` datasets. Horizontal axis is the percent of the subgroup, vertical axis is unfairness (i.e., the value of expression (6)).

conclude that the Kearns et al. approach, while it discourages collecting additional data from only the smallest subgroups in a dataset (thereby not satisfying *Incentive Compatibility*), does not appear to discourage additional data collection when each subgroup's proportion within the population stays consistent.
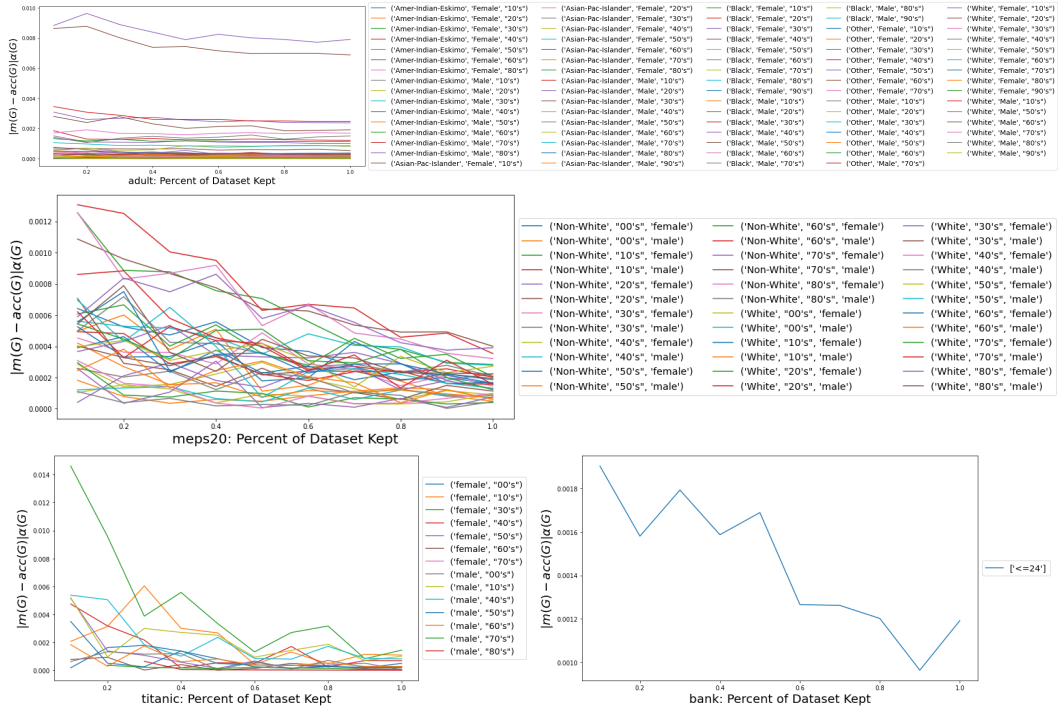
Figure 5: Values of expression (6) on the `adult`, `meps20`, `titanic`, and `bank` datasets. Horizontal axis is the percent of the entire dataset kept, vertical axis is unfairness (i.e., the value of expression (6)).