Bounding and Approximating Intersectional Fairness through Marginal Fairness

Mathieu Molina
Inria
FairPlay team
91120 Palaiseau, France
mathieu.molina@inria.fr

Patrick Loiseau
Inria
FairPlay team
91120 Palaiseau, France
patrick.loiseau@inria.fr

Abstract

Discrimination in machine learning often arises along multiple dimensions (a.k.a. protected attributes); it is then desirable to ensure *intersectional fairness*—i.e., that no subgroup is discriminated against. It is known that ensuring *marginal fairness* for every dimension independently is not sufficient in general. Due to the exponential number of subgroups, however, directly measuring intersectional fairness from data is impossible. In this paper, our primary goal is to understand in detail the relationship between marginal and intersectional fairness through statistical analysis. We first identify a set of sufficient conditions under which an exact relationship can be obtained. Then, we prove bounds (easily computable through marginal fairness and other meaningful statistical quantities) in high-probability on intersectional fairness in the general case. Beyond their descriptive value, we show that these theoretical bounds can be leveraged to derive a heuristic improving the approximation and bounds of intersectional fairness by choosing, in a relevant manner, protected attributes for which we describe intersectional subgroups. Finally, we test the performance of our approximations and bounds on real and synthetic data-sets.

1 Introduction

Research on fairness in machine learning has been very active in recent years, in particular on fair classification under *group fairness* notions, see e.g., [16, 30, 34, 33, 7, 28]. Such notions define demographic groups based on so-called *protected attributes* (e.g., gender, race, religion), and impose that some statistical quantity be constant across the groups. For instance, demographic parity imposes that the class-1 classification rate is the same for all groups, but other notions were defined such as equal opportunity [16] or calibration by group [7]—see a survey in [4]. As exact fairness is too constraining, one often measures *unfairness*, which roughly quantifies the distance to the fairness constraint.

Most works on fair classification consider a single protected attribute and hence only two (or a small number of) groups. Then, they use measures of unfairness to evaluate and penalize classifiers in order make them more fair. This is making an implicit but very fundamental assumption that one can estimate the unfairness measure from the data at hand. With only a few groups, this assumption is indeed easily satisfied as there are sufficiently many data points for each group.

In many—if not most—real-world applications, there are multiple protected attributes (typically 10-20) along which discrimination is prohibited [1, 2]. It is then desirable to consider the strong notion of *intersectional fairness*, which roughly specifies that no subgroup (defined by an arbitrary combination of protected attributes) is unfavorably treated. In that case, however, estimating the unfairness measure becomes very challenging: as the number of groups is exponentially large (e.g., 2^{10} for 10 binary protected attributes), it is very likely that the dataset has at least one subgroup

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

for which there are very few (or zero) data point. A potential solution is to treat each protected attribute separately through its *marginal unfairness* (which is easy to estimate); but it was observed in several real-world and algorithmic examples that it is not sufficient to ensure intersectional fairness [8, 5, 21, 22]. This raises the question: *How to estimate intersectional fairness from data, and what is its precise relation to marginal fairness?* To date, only very few papers have tackled this issue. [21, 22] adopt a definition of intersectional fairness that weights the unfairness of each group by its size. This allows them to get large-samples generalization guarantees of empirical estimates (hence solving the estimation issue), but then it does not protect minorities since it allows a very high unfairness for tiny subgroups—which is contradictory to the intuitively desired behavior.

[17] makes a similar assumption by considering only subgroups above a minimum size, which eases estimate generalization. [14] on the other hand uses the more natural definition of intersectional fairness based on the worst treated group irrespective of its size; but they consider only a few protected attributes, precisely to have enough data points on each subgroup to estimate intersectional unfairness. [13] extends this work by proposing methods to interpolate for subgroups for which too few points are available, based on Bayesian machine learning models. However, this work is empirical and does not give any guarantee on the estimates obtained. In this paper, we also use the natural (strong) definition of intersectional fairness but we take instead a purely statistical approach. We view the protected attributes as random variables to understand intersectional fairness and how it related to marginal fairness more finely.

Contributions: We identify sufficient conditions under which intersectional fairness can be exactly derived from marginal densities, which clarifies when marginal unfairness is a good estimate of intersectional unfairness. We prove probabilistic bounds on intersectional unfairness based on marginal densities and independence measures of the protected attributes, that we show are easy to estimate. We propose a method to improve the approximation of intersectional unfairness and the theoretical bound based on grouping carefully some of the protected attributes together, which we do through a heuristic by leveraging the independence measures exhibited in our bounds. We perform experiments on real and synthetic datasets that illustrate the performance of our approach. In particular, we show that grouping with our heuristic does improve the approximation of intersectional unfairness. To the best of our knowledge, our work is the first work to exploit statistical information to better understand and estimate intersectional (un)fairness. Our work is fairly general and can be instantiated for a variety of standard fairness notions (demographic parity, equal opportunity, etc.). For simplicity, we focus on discrete protected attributes and on classification, but most of the core results can be extended to other cases.

Further Related Works: [31] proposes a unified framework to train a fair classifier under intersectional fairness metrics, but without taking into account regimes with sparse group membership data. Some works propose to audit the accuracy of fairness metrics in contexts other than intersectionality, when there are missing data [35] or when there are unlabeled examples [18]. Others tackle the problem of intersectionality beyond group fairness, e.g., [32] considers causal intersectional fairness. Finally there has been some interest [24, 10] in a different formulation of intersectional group fairness as a multi-objective optimization problem where each objective is the discrimination faced by a given protected group. Another interesting approach to fairness is individual fairness developed in [12], however this is quite different from group fairness metrics on which we focus on and our techniques do not apply.

2 Setting and Models

2.1 Basic Setting

Notational convention: Wherever useful, for any two random variables V and W, we will use the shorthand $p_V(v) = \Pr(V = v)$, $p_{V,W}(v, w) = \Pr(V = v, W = w)$ and $p_{V|W}(v \mid w) = \Pr(V = v \mid W = w)$.

Consider a multi-class classification task. A given individual is described by a tuple of random variables (X, A, Y) drawn according to a distribution $\mathcal{D}$ where X is the features vector, Y is the label with values in $\mathcal{Y}$, and $A = (A_1, ..., A_d)$ is a d-tuple of protected attributes. The only variable used to make a prediction is X and the only variable to measure unfairness is A, but otherwise there are no constraints and A can be a part of X. We denote the support of X, Y, A and A_k for $1 \leq k \leq d$, by $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{A}$ and $\mathcal{A}_k$ respectively. We assume that $\mathcal{A}$ is finite (hence discrete). For a deterministic classifier h, $\hat{Y} = h(X)$ is the predicted class for a random individual. The classifier h is fixed, as we are interested in measuring its fairness and not finding a fair classifier.

To compare the discrimination between groups, we consider a second random variable A' such that $(A' \mid \hat{Y})$ is independent and identically distributed (i.i.d.) to $(A \mid \hat{Y})$. Some authors look at the difference in the treatment of protected groups as a ratio [14], and some others as a difference [21]. Here we choose to study discrimination in terms of ratio. We further apply a logarithm to symmetrize the discrimination measure between two protected groups and for ease of computation. We will consider Statistical Parity for simplicity of exposition, but other group fairness metrics can be either derived directly or adapted using the methods developed in this paper (see Appendix A.1). We define our measure of unfairness as follows:

Definition 2.1. For a distribution $\mathcal{D}$ and a classifier h, we define the *intersectional unfairness* and the kth *protected attribute marginal unfairness* as:

$$u^* = \sup_{(y,a,a') \in \mathcal{Y} \times \mathcal{A}^2} u(y,a,a'), \quad \text{and} \quad u_k^* = \sup_{(y,a_k,a'_k) \in \mathcal{Y} \times \mathcal{A}_k^2} u_k(y,a_k,a'_k) \tag{1}$$

$$\text{with } u(y,a,a') = \left| \log\left(\frac{\Pr(\hat{Y}=y \mid A=a)}{\Pr(\hat{Y}=y \mid A'=a')} \right) \right|, \; u_k(y,a_k,a'_k) = \left| \log\left(\frac{\Pr(\hat{Y}=y \mid A_k=a_k)}{\Pr(\hat{Y}=y \mid A'_k=a'_k)} \right) \right|. \tag{2}$$

One could think that if the marginal unfairness of each protected attribute is smaller than some $\epsilon > 0$, then the overall unfairness is smaller than ϵ; measuring $\mathrm{u}^{\mathrm{M}} = \sup_k u_k^*$ corresponds to this idea. As stated in the introduction this is not sufficient to describe unfairness and we can still have $u^* > \mathrm{u}^{\mathrm{M}}$. We can rewrite (1) as $u^* = \sup_{\mathcal{Y}} \log(\sup_{\mathcal{A}} p_{\hat{Y}|A} / \inf_{\mathcal{A}} p_{\hat{Y}|A})$, and similarly for u_k^*. This means that to measure unfairness we only need to analyze the function $p_{\hat{Y}|A}$.

2.2 Estimation of Unfairness

If we want to estimate unfairness, the most straightforward approach is to estimate the probability mass function $p_{A,\hat{Y}}$ and then to compute the unfairness over these estimated distributions. The main difficulty in estimating the unfairness is estimating $\inf p_{\hat{Y}|A}$, as we can upper bound the $\sup$ by 1, but we cannot easily lower bound the $\inf$. For a data-set of n samples and d protected attributes, we denote for $(a,y) \in \mathcal{A} \times \mathcal{Y}$ the counts by group and prediction as $N_{a,y} = \sum_{i=1}^{n} \mathbb{1}[(A^{(i)}, \hat{Y}^{(i)}) = (a,y)]$ where $(A^{(i)}, \hat{Y}^{(i)})$ is the i-th i.i.d. realization of $(A, \hat{Y})$. The empirical probability is then defined as $\hat{\Pr}(\hat{Y}=y, A=a) = N_{a,y}/n$. [14] shows in Theorem VIII.3 that the error made by using empirical estimates is decreasing in N_a, which means that there needs to be sufficient data for each protected group to estimate u^*. When there are many protected groups the probability that at least one group receives no sample is high, and in this case there is at least one a in $\mathcal{A}$ for which the empirical probability $\hat{\Pr}(\hat{Y}=y \mid A=a)$ is undefined, hence the $\inf$ and $\sup$ cannot be computed. [13] and [14] alleviate this issue of 0-counts by using a Dirichlet prior of uniform parameter $\alpha > 0$. This yield the Bayesian estimates $(N_{a,y} + \alpha)/(n + |\mathcal{A}||\mathcal{Y}|\alpha)$, that are then used to compute the estimator u_B. They also propose other methods to estimate $p_{\hat{Y}|A}$ which empirically performs better, but without guarantees; whereas u_B has the nice property that u_B is a consistent estimator of u^*. This is because of the consistency of the Bayesian probability estimates and by applying the Continuous Mapping Theorem for $\max$ and $\min$ which are continuous. Note that the empirical estimator (with $\alpha = 0$ is also consistent, but has infinite bias.

Nonetheless, u_B has the drawback that for a low amount of samples and when the number of protected groups is high, it is almost determined deterministically by the parameter α and cannot be trusted. If $N_a = 0$ for a protected group a, the estimated distribution is uniform on $\hat{Y} \mid A=a$ and this group does not affect the computation of the $\sup$ and $\inf$. Hence if the most discriminated group is among the undiscovered one, we risk making an important error on the estimation. When N_a increases, we gain more information on the distribution of $\hat{Y} \mid A$. However, when N_a is still low for all groups, the estimated distribution of the $\inf$ of $\hat{Y} \mid A=a$ is almost entirely determined by the prior parameter α.

2.3 Probabilistic Unfairness

When the number of protected subgroups grows arbitrarily large, it may be useless to try to guarantee fairness for every single one of them, regardless on how many people this truly affects. Should a decision maker sacrifice any potential predictive performance in order to guarantee fairness? It could be argued that an algorithm which discriminates 1 person among a 1000 can be described as fair to an extent. We may even be able to directly compensate the small amount of persons discriminated against if possible. Let us consider another example: if a company has clients on which it leverages machine learning predictions to make decisions, it would seem very limiting to guarantee fairness

for clients among specific protected groups for whom we will almost never deal with. Nevertheless, if the underlying clients distribution changes, our decision making process should also reflect this change in terms of fairness. This motivates looking at unfairness probabilistically in $(\hat{Y}, A, A')$. To do that we define $U = u(\hat{Y}, A, A')$ the random variable which corresponds to randomly choosing a prediction, and then independently selecting two protected groups according to $p_{A|\hat{Y}}$ to compare them. We now define our notion of probabilistic unfairness:

Definition 2.2. For $\epsilon \geq 0$ and $\delta \in [0, 1]$, we say that classifier h over distribution $\mathcal{D}$ is (ϵ, δ)-probably intersectionally fair if $\Pr(U > \epsilon) \leq \delta$.

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. Probable intersectional fairness corresponds to searching for quantiles of U. We define the δ-probabilistic unfairness as $\epsilon^*(\delta) = \min\{\epsilon \in \mathbb{R} \mid \Pr(U > \epsilon) \leq \delta\}$. It is the $(1 - \delta)$-quantile of U. We also know by definition that any classifiers over any distributions is $(u^*, 0)$-probably intersectionally fair, as we have $U \leq u^*$ with probability 1. This shows that probabilistic fairness is a relaxed version of the hard intersectional unfairness as $\lim_{\delta \to 0} \epsilon^*(\delta) = u^*$, and thus can be made arbitrarily close to intersectional fairness. In order to give more intuition on what this measure of fairness represents, we will briefly only for this paragraph consider discrimination of protected groups compared to the predictions distribution $p_{\hat{Y}}$ instead of between groups, meaning that we now measure $|\log(p_{\hat{Y}|A}/p_{\hat{Y}})|$. Suppose that a prediction model will be deployed over a population of n individuals. Then if the classifier is (ϵ, δ)-probably intersectionally fair, this means that $\mathbb{E}_{A,\hat{Y}}[\sum_{i=1}^{n} \mathbb{1}[u(\hat{Y}^{(i)}, A^{(i)}) > \epsilon]]$ the expected number of people that faces a discrimination more than ϵ is less than $n\delta$. This allows us to measure and control the size of the population that may face a difference in treatment that would be deemed too high. It corresponds to the notion of fairness we were searching for. For more comparisons between these different notions, see Appendix A.3.

As a remark, looking at $\mathbb{E}[U]$, it can serve as a lower bound of u^* because $\mathbb{E}[U] = \sum_{y,a,a'} p_{\hat{Y},A,A'}(y, a, a')u(y, a, a') \leq u^*$. This represents the average discrimination in a population between two protected groups. This is weaker than the notion presented above and is only mentioned in passing.

Probabilistic fairness can be especially relevant in the context where A are continuous sensitive attributes. Indeed, even for a very basic multivariate normal distribution on A, we will end up with $u^* = \infty$ which is unhelpful. Yet by considering this notion of probabilistic fairness we end up with finite (hence comparable) measures of unfairness where the discriminated population size can be explicitly controlled; see Appendix A.4 for some examples. All in all, this notion of probabilistic unfairness, beyond its main interest of being a relaxed version of intersectional unfairness, could be in itself helpful for decision makers.

3 Measures of Independence and Theoretical Bounds

We now focus on providing valid (ϵ, δ) couples for probable intersectional fairness. First note that while the intersectional unfairness u^* is hard to estimate, it is much easier to estimate the marginal unfairness u^{M}. The work done by [22] in the different setting of weighted unfairness, however, shows through experiments that across multiple classifiers and data-sets, u^* and u^{M} can be uncorrelated, correlated, or even equal. Building on this observation, we would like to approach u^* using marginal quantities estimable for reasonably-sized data-sets.

3.1 Intersectional Unfairness with Independence

Since the intersectional unfairness takes into account the interactions between all the protected attributes A_k, one could guess that if the A_k are mutually independent, this implies that u^* is close to u^{M}. Our first result is not far from this intuition, but we also need to take into account the influence from the classifier h. Indeed, even if the protected attributes are independent, since the classifier makes predictions based on X which may encode redundant information from some A_k, there can be interaction between those protected attributes through the classifier. See Appendix B.1 for a counter example with the independence of the A_k only but no clear relationship between marginal and intersectional fairness.

Proposition 3.1. *If the protected attributes A_k are mutually independent and mutually independent conditionally on $\hat{Y}$, then*

$$u^* = \sup_{y \in \mathcal{Y}} \sum_{k=1}^{d} \sup_{(a_k, a'_k) \in \mathcal{A}_k^2} u_k(y, a_k, a'_k) \leq \sum_{k=1}^{d} u_k^*. \tag{3}$$

Sketch of proof. The main idea is to decompose p_A and $p_{A|\hat{Y}}$ as their products of marginals using the independence assumptions, and using the fact that the $\sup$ taken over a product of functions with independent variables is distributed over the product. The inequality is obtained because the $\sup$ of a sum is smaller than the sum of the $\sup$. See proof in Appendix B.2. □

This theorem gives us a first sense on how intersectional unfairness relates with marginal unfairness in some contexts. This shows us that if the independence conditions are fulfilled, then u^* becomes easy to estimate. What we provide here are conditions and a equation to derive a direct relationship between the intersectional unfairness and the marginal unfairness of each A_k. These are unfortunately too strong conditions to actually expect and are almost never randomly satisfied, but they help us give insight into the relationship between marginal and intersectional fairness. It also drives the analysis conducted in the next sub-section. We would like to relax the independence criteria while still using marginal information from the problem.

3.2 Bounds on Probable Intersectional Fairness

In order to bound the probable intersectional unfairness and relate it with the strictly independent case, we want to use some measure of independence. We want to bound in probability the joint probability density $\Pr(A=a)$ with the product of its marginals $\prod_{k=1}^{d}\Pr(A_k=a_k)$. We will use one of the possible multivariable generalization of Mutual Information known as Total Correlation [29]:

$$C(A)=\mathbb{E}_A\Big[\log\Big(\frac{p_A(A)}{\prod_{k=1}^{d}p_{A_k}(A_k)}\Big)\Big]=\sum_{a\in\mathcal{A}}p_A(a)\log\Big(\frac{p_A(a)}{\prod_{k=1}^{d}p_{A_k}(a_k)}\Big)=\Big(\sum_{k=1}^{d}H(A_k)\Big)-H(A), \quad (4)$$

where $H(A)$ is the Shannon Entropy of A. Similarly we define the conditional total correlation as $C(A\,|\,\hat{Y})=\mathbb{E}_{A,\hat{Y}}[\log(p_{A|\hat{Y}}(A\,|\,\hat{Y})/\prod_k p_{A_k|\hat{Y}}(A_k\,|\,\hat{Y}))]=(\sum_{k=1}^{d}H(A_k\,|\,\hat{Y}))-H(A\,|\,\hat{Y})$ where $H(A\,|\,\hat{Y})$ is the conditional entropy of A given $\hat{Y}$. Note that both can also be written in terms of a KL or expectation in $\hat{Y}$ over conditional KL divergence, which means that $C(A)\geq 0$ and $C(A\,|\,\hat{Y})\geq 0$. From these measures of independence, we intuitively define the following two random variables, $L=\log(p_A(A)/\prod_k p_{A_k}(A_k))$ and $L_y=\log(p_{A|\hat{Y}}(A\,|\,\hat{Y})/\prod_k p_{A_k|\hat{Y}}(A_k\,|\,\hat{Y}))$. By definition we have that $\mathbb{E}[L]=C(A)$ and $\mathbb{E}[L_y]=C(A\,|\,\hat{Y})$. We denote σ and σ_y the standard deviation of these two variables. We have the following property:

$$\perp\!\!\!\perp_{k=1}^{d} A_k \Leftrightarrow C(A)=0\Leftrightarrow\sigma=0 \quad \text{and} \quad \perp\!\!\!\perp_{k=1}^{d} A_k|\hat{Y}\Leftrightarrow C(A\,|\,\hat{Y})=0\Leftrightarrow\sigma_y=0. \quad (5)$$

The equivalence between independence and $C(A)=0$ comes from rewriting $C(A)$ as a KL and the fact that $\mathrm{KL}(P\|Q)=0$ if and only if $P=Q$ almost everywhere. For $C(A\,|\,\hat{Y})=\mathbb{E}_y[\mathrm{KL}(p_{A|\hat{Y}=y}\|\otimes p_{A_i|\hat{Y}=y})]$ we also use that the expectation of a positive random variable is 0 if and only the variable is 0 almost everywhere. When $\sigma=0$ then $L=c$ is a constant which means that $p_A=\prod_k p_{A_k}e^c$, and using that the probabilities must sum to 1 we have $e^c=1$ hence $L=c=0$. The same arguments apply for σ_y. We denote $I(V,W)=H(V)-H(V\mid W)$ the mutual information between a variable V and W. With these definitions, we can now derive the following theorem which bounds the probable intersectional fairness with independence measures and functions of marginal densities:

Theorem 3.2. *For $\delta\in(0,1]$, any classifier h over a distribution $\mathcal{D}$ is (ϵ_1,δ) and (ϵ_2,δ)-probably intersectionally fair with*

$$\epsilon_1=2\sqrt{2}\frac{s^*}{\sqrt{\delta}}+\sup_{y\in\mathcal{Y}}\Big\{\sum_{k=1}^{d}\sup_{(a_k,a_k')\in\mathcal{A}_k^2}u_k(y,a_k,a_k')\Big\} \quad (6)$$

$$\epsilon_2=\sqrt{2}\frac{s^*}{\sqrt{\delta}}+\gamma+\sup_{y\in\mathcal{Y}}\Big\{\sum_{k=1}^{d}\log\Big(\frac{p_{\hat{Y}}^{1-1/d}(y)}{\inf_{a_k\in\mathcal{A}_k}p_{\hat{Y}|A_k}(y\,|\,a_k)}\Big)\Big\} \quad (7)$$

$$\textit{where}\quad s^*=(\sigma^{2/3}+\sigma_y^{2/3})^{3/2}\quad \textit{and}\quad \gamma=C(A)-C(A\,|\,\hat{Y})=\Big(\sum_{k=1}^{d}I(A_k,\hat{Y})\Big)-I(A,\hat{Y}). \quad (8)$$

Sketch of proof. We apply Chebyshev's inequality to L and L_y for some introduced parameters $\boldsymbol{\alpha}$ to bound the tails of these random variables, while making sure that overall the probability bounds stay larger than $1-\delta$. We can then compute inequalities on p_A and $p_{A|\hat{Y}}$, and take the $\inf$ for a and $\sup$ for $\boldsymbol{\alpha}$. This leads to a constrained minimization problem that can be solved, which yields s^*. The full proof is in Appendix B.3. For ϵ_2 we additionally use that $p_{\hat{Y}|A}\leq 1$ as $\mathcal{Y}$ is discrete. □

We observe that both ϵ_1 and ϵ_2 are composed of one term in s^* related with the δ-confidence, and a quantity with marginal information. Aditionnaly ϵ_2 also includes a term in γ that corresponds to some form of mutual information correction. We can control the confidence in this bound with the parameter δ. Because $s^* = 0$ if and only if $\sigma = \sigma_y = 0$ and combined with (5) we can see that s^* somewhat measures how far we are from the conditions of Proposition 3.1. With ϵ_1 we see that when s^* goes to zero, we recover exactly the conditions of Proposition 3.1.

In order to prove Theorem 3.2, we used Chebyshev's inequality. We can derive a similar proof for other concentration inequalities, specifically with Chernoff bounds through the estimation of the moment generating function, which often leads to tighter bounds. However this leads to harder quantities to estimate in addition to having to solve a non-convex optimization problem, see Appendix B.4.

To conclude this section we provide additional intuition on the relationship between marginal and intersectional fairness. For this we temporarily change our definition of unfairness for this paragraph only: suppose we are now only interested in the unfairness regarding one outcome $y \in \mathcal{Y}$, say $y = 1$, and let us redefine our unfairness, probabilistic unfairness, γ and s^* accordingly (see in Appendix (15)). We can then derive the following corollary from the proof of the above Theorem:

Corollary 3.3. *Denoting $(\Omega, \mathcal{T}, \Pr)$ the probability space on which (A, A') is defined, there exists an event F so that for $f(a) = \prod_{k=1}^{d} p_{\hat{Y}=y|A_k}(a)/p^d_{\hat{Y}=y}$ we have*

$$\sup_{\omega \in F} p_{\hat{Y}=y|A}(A(\omega)) \in [p_{\hat{Y}}(y) e^{\frac{-2\sqrt{2}s^*}{\sqrt{\delta}} - \gamma} \sup_{\omega \in F} f(A)(\omega), p_{\hat{Y}}(y) e^{\frac{2\sqrt{2}s^*}{\sqrt{\delta}} - \gamma} \sup_{\omega \in F} f(A(\omega))], \tag{9}$$

$$\inf_{\omega \in F} p_{\hat{Y}=y|A}(A(\omega)) \in [p_{\hat{Y}}(y) e^{\frac{-2\sqrt{2}s^*}{\sqrt{\delta}} - \gamma} \inf_{\omega \in F} f(A)(\omega), p_{\hat{Y}}(y) e^{\frac{2\sqrt{2}s^*}{\sqrt{\delta}} - \gamma} \inf_{\omega \in F} f(A(\omega))], \tag{10}$$

$$\textit{and} \quad \Pr(F \mid \hat{Y} = y) \geq 1 - \delta, \tag{11}$$

and the same inequalities hold for $\sup_{\omega \in F} p_{\hat{Y}=y|A}(A')(\omega)$ for the same event F.

The proof can be found in Appendix B.3. This means that there is a fraction of the relevant pairs population of size bigger than $1 - \delta$, for which we can give intervals for the extreme values of $p_{\hat{Y}=y|A}A$ and $p_{\hat{Y}=y|A}A'$ over this fraction F. These intervals are centered and reduce around a unique quantity as s^* goes to 0. We also have that $|\log(\sup_{\omega \in F} p_{\hat{Y}=y|A}(A)(\omega)/\inf_{\omega \in F} p_{\hat{Y}=y|A}(A)(\omega))|$ goes to u^* and that $|\log(\sup_{\omega \in F} f(A)(\omega)/\inf_{\omega \in F} f(A)(\omega))|$ goes toward the quantity derived in the corresponding Proposition 3.1 as δ goes to 0. This tells us that in a way we approach the quantity derived in Proposition 3.1 when s^* and δ go to 0.

3.3 Estimation of the measures of independence

Theorem 3.2 trades the precise estimation of u^* with an upper bound, but with much easier quantities to estimate. More specifically, as they are information measures, we can leverage the extensive literature on statistical estimators and entropy estimation. We can intuitively see that the estimation of s^* and γ will be easier to handle because even the estimation with the empirical distribution $\hat{p}_{A,\hat{Y}}$ is always well defined, and is a Maximum Likelihood Estimator (MLE) as continuous functions of MLE. They are well defined because s^* and γ are functions of entropies and of the quantities $Q(P) = \sum_i p_i \log(p_i)^2$ for a probability distribution P, which is finite event for $p_i = 0$ because $x \mapsto x\log(x)$ and $x \mapsto x\log(x)^2$ are continuous at 0. Contrarily to u_B we do not have to use any prior to obtain a well defined estimator. In addition, using the delta method on the sum of entropies, for which the MLE is asymptotically normal (See [27] 3.1), shows that $\hat{\gamma}$ is asymptotically normal. For more information on the estimation of entropy, mutual information or total correlation we defer to [27, 26, 3, 6, 15] to name but a few. Moreover even with the very simple MLE, we can obtain L_2 error upper-bounds for $H(P)$ in $\mathcal{O}(\log(|P|)^2/n)$ where $|P|$ is the number of outcomes for a discrete distribution P [20]. This bound depends only on the number of outcomes (supposed known), and not the actual distribution. Using the same tools, we derive a rough error bound for $Q(P)$:

Proposition 3.4.

$$\mathbb{E}[(Q(P) - Q(\hat{P}))^2] = \mathcal{O}\Big(\frac{\log^4(n)}{n}\Big). \tag{12}$$

Sketch of proof. We apply the methods described in [20] that bounds the bias using approximation theory for Bernstein polynomials and bounds the variance using the Efron-Stein inequality. See proof in Appendix B.6. □

More efficient estimators can be created using methods of [19], nevertheless the main interest of this proposition is to show that these quantities have an error rate depending on the number of samples n, and not the number of samples per group N_a, which is much better.

Beyond the practical use of these inequalities and approximations, these theorems also show one crucial idea: we can relate intersectional and marginal unfairness with the help of information on the independence of the protected attributes.

4 Refined approximations and inequalities

In the previous section, we have derived conditions for marginal unfairness to directly relate to u^*, and bounds on probable intersectional fairness. We now would like to propose an approximation of u^* using similar ideas. Looking at (6), (3), and indirectly through Corollary 3.3 it seems natural to propose as one possible approximation of u^* the following quantity:

$$u_I = \sup_{y \in \mathcal{Y}} \sum_{k=1}^{d} \sup_{(a_k, a'_k) \in \mathcal{A}_k^2} u_k(y, a_k, a'_k). \tag{13}$$

For the rest of the article, we thus now only focus on s^* and u_I. Compared with u_B the estimator with Bayesian prior, it does not depend on a prior parameter, and is usually well defined as we only need $N_{a_i,y}$ the number of samples per a_i and y to be strictly positive instead of all $N_{a,y}$. However this estimator of u^* is not consistent. We will show that the previous bounds can be improved and that we can make our estimator consistent by gradually grouping together the protected attributes as the number of samples increases.

4.1 Grouping protected Attributes together

Until now, we have always decomposed the protected attributes A on their marginals A_i. However it may be that we have more than just marginal information available. Take the example of 4 protected attributes $A = (A_1, A_2, A_3, A_4)$. For a set $t \subseteq \{1, 2, 3, 4\}$, we define $A_t = (A_k)_{k \in t}$. We may not have enough data to compute the full intersectional unfairness, but it may be possible to compute it for the grouped protected attributes $A_{\{1,2\}} = (A_1, A_2)$ and $A_{\{3,4\}}$. We can use the same decomposition as we did before on the new marginals attributes (which corresponds to flattening A_1 and A_2 together) with support $\mathcal{A}_{\{1,2\}} = \mathcal{A}_1 \times \mathcal{A}_2$ and $\mathcal{A}_{\{3,4\}} = \mathcal{A}_3 \times \mathcal{A}_4$.

More generally, let q be a partition of $\{1, ..., d\} = [d]$. For a partition q, we denote $A^{(q)} = (A_t)_{t \in q}$. This is only a different way to group together the marginal attribute, and is the same as A. Whenever quantities are changed according to some partition q, it will be indicated with (q). For each of the new marginal attributes defined by a set t of q, the new marginal unfairness u_t^* corresponds to the intersectional unfairness of the $(A_k)_{k \in t}$. If the A_t are independent, and independent conditionally on $\hat{Y}$, we can apply Theorem 3.1 and obtain directly u^* through the newly defined marginals. If we relax the independence conditions, the same arguments of the previous section still apply, and we can look at the bounds and approximations defined by these new marginal densities. We denote the new approximation with partition q by $u_I^{(q)} = \sup_{y \in \mathcal{Y}} \sum_{t \in q} \sup_{(a_t, a'_t) \in \mathcal{A}_t^2} u_t(y, a_t, a'_t)$ where we are using the new marginals defined by q. If we use the partition q of singletons then $u_I^{(q)} = u_I$, and if we use the trivial partition then $u_I^{(q)} = u^*$. The constraints of independence for these new marginals should be more feasible than the original marginals, hence it is possible that the A_t fulfill the independence conditions, even if the A_k do not (the trivial partition is such an example). If we have enough data to compute the marginal densities derived from q and the A_t fulfill the independence conditions, we can then compute u^* through the partition q. Of course most of the time the independence conditions are not satisfied satisfied for a partition q. Nonetheless because s^* measures how far we are from the independence conditions, we can more carefully select a partition among those for which we can compute the new marginal densities.

4.2 Efficient Partition Selection

Let $\mathcal{Q}$ be the set of all *feasible* partitions q, that is $\mathcal{Q} = \{q \in \mathcal{P}([d]) \mid \forall t \in q, \forall (a_t, y) \in \mathcal{A}_t \times \mathcal{Y}, N_{a_t,y} > 0\}$ with $\mathcal{P}([d])$ the set of all partitions of $[d]$. This set represents the set of partitions for which we can compute the newly defined marginals without having to use a prior parameter. Note that $\mathcal{Q}$ is a random set that converges to $\mathcal{P}([d])$ almost surely as the number of samples n increases. If $q \in \mathcal{Q}$, then any partitions q' finer (meaning that any element of q' is a subset of an element of

q) than q is in $\mathcal{Q}$ as well. We will say that q can be merged further if there exists a partition $q' \in \mathcal{Q}$ so that q is finer than q'. Note that the choice of a partition q does not change the value of u^* but only that of $u_I^{(q)}$. We therefore want to find a good feasible partition q in $\mathcal{Q}$ so that we can expect heuristically $|u_I^{(q)} - u^*|$ to be the lowest among the partitions. There are two criteria that should help us decide which partition $q \in \mathcal{Q}$ to choose from.

Algorithm 1 Greedy Partition Finder

input: Protected attributes data and occurrences of $\hat{Y}$
require: The partition of singletons is feasible
$q^* \leftarrow$ the partition of singletons
repeat
  $\mathcal{M} = \{\{t_1 \cup t_2\} \cup q^* \setminus (\{t_1\} \cup \{t_2\}), (t_1, t_2) \in q^{*2}, t_1 \neq t_2\}$
  $s^*_{\text{min}} \leftarrow +\infty$
  for q in $\mathcal{M}$ **do**
    if q is feasible and $s^*(q) < s^*_{\text{min}}$ **then**
      $(s^*_{\text{min}}, q^*) \leftarrow (s^*(q), q)$
    end if
  end for
until $\mathcal{M} = \emptyset$ or $s^*_{\text{min}} = \infty$ (Nothing possible to merge)
return: q^*

If a partition q' is coarser than q (which means that q is finer than q'), then reasonably the approximation is better with q' than q. The reasoning is that by taking coarser partitions, we are taking more interactions between the protected attributes into account. For example the coarsest partition which is the whole set gives us the intersectional unfairness as mentioned earlier. However because the 'finer-than' relationship is only a partial order, we are not able to choose between any two sets. Because Theorem 3.2 seems to hint that there is a relationship between the error $|u_I^{(q)} - u^*|$, and the distance to the independence conditions s^*, the second criterion will be to select the partitions q with the smallest $s^*(q)$ defined as s^* but taking the marginals in q. These two criteria are closely linked. Selecting coarser partitions does tend to yield partitions with smaller s^*, but not always. We give some details on relationship between $s^*(q)$ of a partition q compared to a coarser one in Appendix C.2. More crucially, finding a good partition with a small $s^*(q)$ will also improve our inequalities as they are a function of $s^*(q)$ which decreases on average as shown in Figure 5 as the number of sample grows (and as the partitions get coarser).

In principle, finding the best partition according to our criteria requires enumerating all feasible partitions which is computationally intractable. Instead we propose a greedy heuristic that we describe in Algorithm 1. We start from the finest partition (the partition of singletons), look at all the feasible partitions (with enough data) that can be obtained from merging two elements of the current partition, select the one with the smallest s^*, and repeat until there are no coarser partitions with enough data. Note that when we want to verify that there is enough data available, we may need to do it multiple times for the same subset of protected attributes. This is an expensive call so it is more efficient to do memoization and remember if there is enough data available for a given subset once encountered, which we do using a hash table to reduce the lookup time. We denote this partition q^*. We have the following property with the proof in Appendix C.1:

Proposition 4.1. *The estimator $u_I^{(q^*)}$ is a consistent estimator of u^*.*

This proposition shows that $u_I^{(q^*)}$ is relevant in estimating u^*, while not needing to use a Bayesian prior with parameters that may overwhelmingly affect the estimation. Note that instead of using $\mathcal{Q}$ which ensures that $N_{a_t,y} > 0$, we can instead use $\mathcal{Q}_\tau$ for $\tau \in \mathbb{N}$ which ensures that for any $q \in \mathcal{Q}_\tau$, $N_{a_t,y} > \tau$ for $t \in q$.

5 Experiments

In this section, we present experimental results that show how our inequalities and approximations perform on real and synthetic data-sets, and compare their estimation error rates as the number of samples grow. All the code used in our experiments can be found in the supplementary material or here.

5.1 Data-sets and processing

In order to compare how well u_B and u_I perform as estimators on data on datasets with a high number of protected attributes, we need to compute u^* which is as discussed above inherently difficult. We will always measure the unfairness with respects to the empirical distribution of the dataset. For this

empirical distribution to yield a well defined fairness measure, we need that $N_{a,y} > 0$ for all a and y. This means that if we want to take into account a high number of sensitive attributes, we have to pick a very large dataset.

We used US Census data from 1990 [11] which contains $n = 2,458,285$ samples, and for which we identified many potential protected attributes. We then train a Random Forest binary classifier on a poverty binary label, where we weight the labels differently so as to obtain about the same number of predictions for each outcome. However we still do not have $N_{a,y} > 0$ on the whole dataset. To alleviate this issue, we will consider subsets of the protected attributes for which this is true, and we will measure fairness with respect to these subsets. We obtain about 100 different subsets with $d = 8$ protected attributes, that we denote as D_i which is the original dataset where we only kept the i-th subset of protected attributes and the predictions $\hat{Y}$. Each of these subsets yield different values of u^* and s^*. We pick 12 (for computational reasons) different D_i with various values of u^* and s^*. Some examples of the final protected attributes include sex, not speaking English at home, being overweight, being Hispanic, and others. We will always take $\delta = 0.1$ when relevant.

We also conduct experiments on synthetic data. We generate $(A, \hat{Y})$ probability distributions from a Dirichlet distribution, thus we can directly compute u^* without dealing with a very large dataset. We take $d = 10$. This synthetic data is one of the worst case for the approximation of u^* with u_I, as the marginal distributions are a sum of 2^{d-1} i.i.d. random variables that all converges to $1/2$ as d grows. We therefore will not plot u_I for the synthetic data (it is close to 0). Nonetheless, this synthetic data remains useful in order to compare the error rates between u_B and $\hat{s}^*$. We denote by P_i a generated probability distribution. We generate 12 of them.

5.2 Experiments Results

We first want to compare the convergence rate of u_B, s^* and u_I to their asymptotic value. To do that, and because they can take different values, we compute for each estimator $\hat{T}_n$ that converges in probability to T the relative expected L_2 error rate $L_2^r(\hat{T}_n) = \mathbb{E}[(\hat{T}_n - T)^2/T^2]$. We fix a number of available samples n from 100 to $2,000$, and we sample without replacement from the datasets. From these available samples, we compute all our estimators. We denote by $\hat{u}_I$ the estimator of u_I computed with the empirical marginal densities for n samples. In order to compute L_2^r, for each subset and each sample size n, we sample from D_i and P_i 20 times for each fixed number of samples n.

We see in Figure 1 on the left-most plot, that $\hat{u}_I$ is reasonably close to u^* on average. Still the gap between ϵ^* and ϵ_1 is quite big. Other bounds such as ϵ_2 or with other concentration inequalities are generally a bit more efficient, but we focus here on comparing the error rates between s^* and u_B, and on how u_I performs. The other two plots look at the L_2^r for the various estimators, with the middle one being with the real datasets D_i, and the right on the synthetic datasets P_i. We can see that $\hat{s}^*$ and $\hat{u}_I$ converges much faster than u_B. Moreover, it seems that the difference in error rate will only grow bigger as d increases, as there is a bigger gap for P_i. We can also see that u_B is unreliable, because the error rate varies a lot depending on α, and can even increase. This is because the parameter α dominates the computation of u_B as discussed earlier.

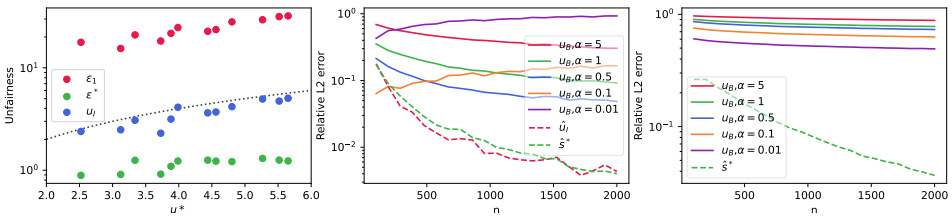

Figure 1: On the left-most plot, each point represents one real dataset D_i, and we compare ϵ_1, ϵ^*, and u_I with u^*. The dotted line corresponds to the equation $x = y$ for reference. The middle plot describes the average over the D_i of L_2^r as n increases for $\hat{u}_I$, $\hat{s}^*$, and $\tilde{u}^*$. u_B is computed for multiple values of α. The right-most plot is similar, but uses the synthetic datasets P_i.

We now conduct similar experiments, but this time using partitions. We can see in the middle plot of Figure 2 that $\hat{u}_I^{(q^*)}$ performs better. The choice of τ the count threshold for grouping always gives

reasonable approximations, with $\tau = 1$ being close to u_B, and τ big makes it close to u_I. Most importantly, the apparent good error rate of u_B is merely an artifact of the current range of u^* being above the starting values of u_B for these α. It is clear that u_B is unreliable by looking at the left plot in Figure 2: the estimation with u_B at $n = 2000$ for different values of α varies very little when u^* varies (it is almost not a function of u^*). This means that u_B depends very little on the data for low amount of samples. Even if it is not perfect, $u_I^{(q^*)}$ still has better performance and is more coherent. We note that the approximation performs well comparatively only when d is high, and considering more sensitive attributes should make an even bigger difference. These results combined with Proposition 4.1 show that $u_I^{(q^*)}$ is a relevant estimator of u^* with scarce data and high number of protected attributes. Concerning $s^*(q^*)$ the right-most plot shows that while it is not completely monotone, $s^*(q^*)$ does decreases on average when using partitions as the number of sample increases. The upper bound will become tighter as n grows, which will make bigger groupings of protected attributes possible.

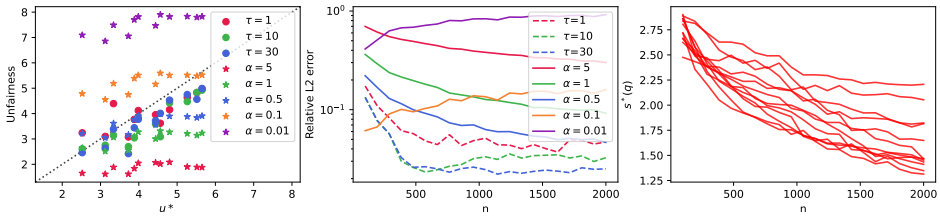

Figure 2: Each point of the same color and shape represent the estimation for one dataset D_i. The estimations are computed for $n = 2000$. The middle plot is the same as above, but with $\hat{u}_I^{(q^*)}$ this time. The rightmost plot is the average evolution of $s^*(q^*)$ for $\tau = 10$ as n increases.

6 Discussion

In this work, we presented new methods to approximate and to bound (in high probability) a strong intersectional unfairness measure, based on statistical information computable from a reasonable dataset. Our results highlight the key role of independence of the protected attributes conditionally to the classifier, and propose to approach it via a smart grouping of some attributes—which our theoretical bound allows us to compute via an efficient heuristic.

Our experiments show that the approximations proposed here perform reasonably well for data-sets with a high number of protected attributes, but that our bounds are not very effective. However their main interest is that it gives insight into the link between marginal and intersectional fairness, which was the main goal of this work. It also helps us derive the proposed approximation. We expect that more effective bounds could be derived for our notion of probabilistic fairness, for instance by making additional assumptions on the distribution, but presumably without an explicit dependence on independence measures and marginal densities, making the link between marginal and intersectional fairness harder to see.

In order to train fair models using the proposed approximations or bounds of this paper, we can use soft counts to compute the empirical densities (based on the classifier score for instance) as suggested in [14]. This makes the approximations and bounds differentiable, and ensure that we can apply gradient based methods so as to solve a constrained or penalized optimization problem using these quantities.

We hope that our approach will enable the development of improved bounds, raise interest in the proposed notion of probabilistic unfairness which we think is crucial to the development of fair algorithms, as well as the use of our approximations to penalize classifiers in order to train intersectionally fair classifiers.

Acknowledgments

This work has been partially supported by MIAI @ Grenoble Alpes (ANR-19- P3IA-0003), by the French National Research Agency (ANR) through grant ANR-20-CE23-0007, and by TAILOR (a project funded by EU Horizon 2020 research and innovation programme under GA No 952215).

References

[1] Equal Credit Opportunity Act, 1974.

[2] Code du Travail. Chapitre II : Principe de non-discrimination, 2020.

[3] Evan Archer, Il Memming Park, and Jonathan W. Pillow. Bayesian entropy estimation for countable discrete distributions. *Journal of Machine Learning Research*, 15(1):2833–2868, 2014.

[4] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. `http://www.fairmlbook.org`.

[5] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR, 2018.

[6] Pengyu Cheng, Weituo Hao, and Lawrence Carin. Estimating total correlation with mutual information bounds. *CoRR*, abs/2011.04794, 2020.

[7] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5 2:153–163, 2017.

[8] Kimberle Crenshaw. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist policies. *University of Chicago Legal Forum*, 1989(1):139–167, 1989.

[9] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Stochastic Modelling and Applied Probability. Springer Berlin Heidelberg, 2009.

[10] Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, and Aaron Roth. *Minimax Group Fairness: Algorithms and Experiments*, page 66–76. Association for Computing Machinery, 2021.

[11] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, page 214–226. Association for Computing Machinery, 2012.

[13] James R. Foulds, Rashidul Islam, Kamrun Keya, and Shimei Pan. Bayesian modeling of intersectional fairness: The variance of bias. In *SIAM International Conference on Data Mining (SDM)*, 2020.

[14] James R. Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. An intersectional definition of fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 1918–1921, 2020.

[15] Yanjun Han, Jiantao Jiao, and Tsachy Weissman. Minimax estimation of divergences between discrete distributions. *IEEE Journal on Selected Areas in Information Theory*, 1(3):814–823, 2020.

[16] Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[17] Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (Computationally-identifiable) masses. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1939–1948. PMLR, 2018.

[18] Disi Ji, Padhraic Smyth, and Mark Steyvers. Can i trust my fairness metric? assessing fairness with unlabeled data and bayesian inference. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18600–18612. Curran Associates, Inc., 2020.

[19] Jiantao Jiao, Kartik Venkat, Yanjun Han, and Tsachy Weissman. Minimax estimation of functionals of discrete distributions. *IEEE Trans. Inf. Theor.*, 61(5):2835–2885, 2015.

[20] Jiantao Jiao, Kartik Venkat, Yanjun Han, and Tsachy Weissman. Maximum likelihood estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory*, 63(10):6774–6798, 2017.

[21] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2564–2572. PMLR, 2018.

[22] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 100–109. Association for Computing Machinery, 2019.

[23] Elena Kosygina. Introductory examples and definitions. cramér's theorem, 2018. `https://sites.math.northwestern.edu/~auffing/SNAP/Notes1.pdf`.

[24] Natalia Martinez, Martin Bertran, and Guillermo Sapiro. Minimax pareto fairness: A multi objective perspective. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6755–6764. PMLR, 2020.

[25] Jeremie Mary, Clément Calauzènes, and Noureddine El Karoui. Fairness-aware learning for continuous attributes and treatments. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4382–4391. PMLR, 2019.

[26] Ilya Nemenman, F. Shafee, and William Bialek. Entropy and inference, revisited. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001.

[27] Liam Paninski. Estimation of entropy and mutual information. volume 15, pages 1191–1253. MIT Press, 2003.

[28] Yaniv Romano, Stephen Bates, and Emmanuel Candes. Achieving equalized odds by resampling sensitive attributes. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 361–371. Curran Associates, Inc., 2020.

[29] Satosi Watanabe. Information theoretical analysis of multivariate correlation. volume 4 of *IBM Journal of Research and Development*, pages 66–82, 1960.

[30] Blake Woodworth, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1920–1953. PMLR, 2017.

[31] Forest Yang, Mouhamadou Cisse, and Sanmi Koyejo. Fairness with overlapping groups; a probabilistic perspective. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4067–4078. Curran Associates, Inc., 2020.

[32] Ke Yang, Joshua R. Loftus, and Julia Stoyanovich. Causal Intersectionality and Fair Ranking. In Katrina Ligett and Swati Gupta, editors, *2nd Symposium on Foundations of Responsible Computing (FORC 2021)*, volume 192 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 7:1–7:20. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2021.

[33] Muhammad B. Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P. Gummadi. Fairness Constraints: Mechanisms for Fair Classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 962–970, 2017.

[34] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1171–1180. International World Wide Web Conferences Steering Committee, 2017.

[35] Yiliang Zhang and Qi Long. Assessing fairness in the presence of missing data. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 16007–16019. Curran Associates, Inc., 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We discuss in Section 5 and in Section D limitations of our approach.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] This work is specifically related to Fairness, and as such we highlight in the Introduction existing works which already discuss some of the societal issues that intersectional fairness raises.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] Each of the statement made in the main text is referenced to the complete proof in the Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We included the code in the supplementary material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We specified relevant data processing details, the actual supervised learning model used not being the main interest. More details are given in the code.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] The plots with error bars are deferred to Appendix D to make the plots readable in the main text.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix D.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We used a dataset from the UCI dataset repository, and accordingly cited it.
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See Appendix D.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] See Appendix D.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Additional elements on Measures of Fairness

A.1 Generalization to other measures of fairness

Throughout the whole paper, we used a specific measure of fairness for simplicity. Nevertheless, the same arguments apply to a broader set of fairness measures, by modifying u^*.

To define u^*, we decided to take the $\log$ of a ratio. We note that when taking the $\sup$ over all possibles a and a' in $\mathcal{A}$, $\sup_{\mathcal{A}^2} u(1,\cdot,\cdot) \leq \epsilon$ is equivalent to the definition in [14], that is to say:

$$\forall (a,a') \in \mathcal{A}^2, \quad e^{-\epsilon} \leq \frac{\Pr(\hat{Y}=1 \mid A=a)}{\Pr(\hat{Y}=1 \mid A=a')} \leq e^{\epsilon}. \tag{14}$$

However if we want to define a measure of unfairness between two protected groups, it is reasonable for it to be symmetric in the groups considered. We make it so by applying $\log$ and an absolute value function to the above middle quantity. Other distances and pseudo distances can also be chosen, such as $|\Pr(\hat{Y}=1 \mid A=a) - \Pr(\hat{Y}=1 \mid A=a')|$. It is also symmetric, but may be less useful in comparing models with many protected attributes that are not designed to be fair, as this measure will be close to 1 most of the time, making the comparison less precise between two different models.

We can also modify U and the definition of probabilistic fairness accordingly to obtain other desirable measures of unfairness. Say that we are only interest in the unfairness related to the outcome $\hat{Y}=1$. Taking $U' = u(1, A, A')$ and $\Pr(U' > \epsilon \mid \hat{Y}=1) \leq \delta$ the new definition of probabilistic fairness in this case, we can derive similar propositions and theorems as done in this paper. We only need to take the expectation and variance with respect to $\Pr(\cdot \mid \hat{Y}=1)$ for L and L_y. This yields statistical quantities which are harder to interpret ($\sum_{a\in\mathcal{A}} \Pr(A=a \mid \hat{Y}=1) \log(\Pr(A=a)/\prod_{k=1}^{d} \Pr(A_k = a_k))$ but that should remain easy to estimate as they are always well defined because, considering the empirical distribution $\hat{p}$ we have $\hat{p}_A(A=a)=0 \implies \hat{p}_A(A=a \mid \hat{Y}=1)=0$. The changed definitions would be the following if we are only interested in the outcome $y \in \mathcal{Y}$:

$$u^* = \sup_{(a,a')\in\mathcal{A}^2} u(a,a'), \quad \text{and} \quad u_k^* = \sup_{(a_k,a_k')\in\mathcal{A}_k^2} u_k(a_k,a_k') \tag{15}$$

$$\text{with } u(a,a') = \Big|\log\Big(\frac{\Pr(\hat{Y}=y \mid A=a)}{\Pr(\hat{Y}=y \mid A'=a')}\Big)\Big|, \; u_k(a_k,a_k') = \Big|\log\Big(\frac{\Pr(\hat{Y}=y \mid A_k=a_k)}{\Pr(\hat{Y}=y \mid A_k'=a_k')}\Big)\Big|, \tag{16}$$

$$\gamma = \mathbb{E}[L - L_y \mid \hat{Y}=y], \quad \sigma = \sqrt{\mathrm{Var}(L \mid \hat{Y}=y)}, \quad \sigma_y = \sqrt{\mathrm{Var}(L_y \mid \hat{Y}=1)}. \tag{17}$$

Notably, we obtain a much nicer variant of Proposition 3.1, with $u^* = \sum_{k=1}^{d} u_k^*$.

Similarly we can also change only the underlying probability distribution. We can replace the underlying probability $\Pr$ by $\Pr_{Y=1}$. Using this new probability measure we see that u^* is a relaxed version of Equality of Opportunity for a binary predictor defined in [16] by

$$\Pr(\hat{Y}=1 \mid A=a, Y=1) = \Pr(\hat{Y}=1 \mid Y=1). \tag{18}$$

Indeed, $u^*=0$ is now equivalent with Equality of Opportunity. Practically, changing the underlying probability does not make much difference as showed in [21] because this amounts to measuring unfairness on the part of the dataset for which $Y=1$.

We compared the treatment faced by groups between them, such as looking at the discrimination between men and women. Another possibility is to measure the difference in treatment faced by a group compared to a reference value. This reference value is most of the time taken to be the population average of the decision criterion $\mathbb{E}_A[p_{\hat{Y}|A}(y \mid A)] = p_{\hat{Y}}(y)$. Therefore instead of evaluating $p_{\hat{Y}|A}(y \mid a)/p_{\hat{Y}|A}(y \mid a')$ we evaluate $p_{\hat{Y}|A}(y \mid a)/p_{\hat{Y}}(y)$.

Finally we can change over which treatment criterion we want to evaluate differences. In this paper we decided to look at the variable $\hat{Y}$. We can similarly define our fairness measure with Y. We can

actually use any (X, A, Y)-measurable random variable Z instead of $\hat{Y}$. For instance $|\hat{Y} - Y|$, which tells us whether or not the prediction is correct for binary classification, can be a good candidate.

Combining all of the above comments, we can consider a wider array of fairness metrics for which variations of the techniques and theorems described in this paper apply.

A.2 Some variants of Theorem 3.2 for modified fairness measures

Intersectional Fairness in terms of absolute difference: We consider the following definition of unfairness:

$$u^* = \sup_{y \in \mathcal{Y}} \sup_{(a,a') \in \mathcal{A}^2} u(y, a, a') \tag{19}$$

$$\text{with} \quad u(y, a, a') = \Big| \log \Pr(\hat{Y} = y \mid A = a) - \Pr(\hat{Y} = y \mid A' = a') \Big|. \tag{20}$$

The new version of Theorem 3.2 is

Theorem. For $\delta \in (0, 1]$, any classifier h over a distribution $\mathcal{D}$ is (ϵ_1, δ)-probably intersectionally fair with

$$\epsilon_1 = e^{-\gamma} \sup_{\mathcal{Y}} p_{\hat{Y}}^{1-d} \left(e^{\frac{\sqrt{2}s^*}{\sqrt{\delta}}} \prod_{k=1}^{d} \sup_{\mathcal{A}_k} p_{\hat{Y}|A_k} - e^{-\frac{\sqrt{2}s^*}{\sqrt{\delta}}} \prod_{k=1}^{d} \inf_{\mathcal{A}_k} p_{\hat{Y}|A_k} \right). \tag{21}$$

Intersectional Fairness when comparing to the population average: We consider the following definition of unfairness:

$$u^* = \sup_{y \in \mathcal{Y}} \sup_{a \in \mathcal{A}} u(y, a) \tag{22}$$

$$\text{with} \quad u(y, a) = \Big| \log \Big(\frac{\Pr(\hat{Y} = y \mid A = a)}{\Pr(\hat{Y} = y)} \Big) \Big|. \tag{23}$$

The new version of Theorem 3.2 is

Theorem. For $\delta \in (0, 1]$, any classifier h over a distribution $\mathcal{D}$ is (ϵ_1, δ)-probably intersectionally fair with

$$\epsilon_1 = \sqrt{2} \frac{s^*}{\sqrt{\delta}} + \sup_{\mathcal{Y}} \max\{\gamma + \sum_{k=1}^{d} \log \Big(\frac{p_{\hat{Y}}}{\inf_{\mathcal{A}_k} p_{\hat{Y}|A_k}} \Big), -\gamma + \sum_{k=1}^{d} \log \Big(\frac{\sup_{\mathcal{A}_k} p_{\hat{Y}|A_k}}{p_{\hat{Y}}} \Big)\} \tag{24}$$

The proof for both of these variants is exactly the same as for 3.2 until we obtain an upper and lower bound on $\sup p_{\hat{Y}|A}$ and $\inf p_{\hat{Y}|A}$. We then use the fact that for $E \subset \mathbb{R}$, $\sup_{(x,y) \in E^2} |x - y| \leq \sup_E x - \inf_E y$, and $\sup_{x \in E} |\log(x/y)| \leq \max\{\log(y/\inf_E x), \log(\sup_E x/y)\}$.

A.3 Comparison between other measures of fairness

In [21], a different fairness metric is used. Indeed, instead of simply measuring unfairness as the unweighted difference $|\Pr(\hat{Y} = 1 \mid A = a) - \Pr(\hat{Y} = 1)|$, they use the weighted difference $\Pr(A = a)|\Pr(\hat{Y} = 1 \mid A = a) - \Pr(\hat{Y} = 1)|$. For this subsection, we will use as a definition of unfairness $u^* = \sup_{a \in \mathcal{A}} u(a)$ with $u(a) = |\Pr(\hat{Y} = 1 \mid A = a) - \Pr(\hat{Y} = 1)|$. We define the weighted unfairness used in [21] as $w^* = \sup_{a \in \mathcal{A}} w(a)$ with $w(a) = p_A(a)u(a)$ the weighted version of u. This definition yields very useful statistical properties in terms of the unfairness estimation, and [21] shows with Theorem 2.11 that the error made using the empirical estimator is less than $\tilde{\mathcal{O}}(\sqrt{((1 + VCDIM(H)) \log(n) - \log(\delta))/n})$ with high probability $1 - \delta$. Unfortunately this notion of unfairness is hard to control as the meaning of $w^* \leq \epsilon$ may be difficult to use for a decision maker, and can lead to the discrimination of groups of small sizes compared to using u^*. This is already discussed and supported empirically in [14].

We will briefly give some inequalities relating these quantities. We have that

$$w^* = \sup_{\mathcal{A}} p_A u \leq \sum_{a \in \mathcal{A}} p_A(a) u(a) = \mathbb{E}[U] \leq u^*. \tag{25}$$

Through these equations we see that w^* cannot approach u^*.

The advantage of the notion of probable intersectional fairness compared to w^* is two-fold: we can be arbitrarily close to u^*, and through δ we explicitly control the size of the population that faces discrimination.

Additionally we will present an example, which shows that when the number of protected groups grows large, the notion of weighted unfairness can become inadequate for certain scenarios compared to probabilistic unfairness.

We will consider that we have 991 protected groups with three different protected groups sets of sizes 1, 495 and 495, for which we will denote any of their elements by a_1, a_2, and a_3 respectively. We will consider that $\Pr(A = a_1) = 0.01$, and that the remaining protected groups are distributed uniformly $\Pr(A = a_2) = \Pr(A = a_3) = 0.001$.

Suppose that $\Pr(\hat{Y} = 1 \mid A = a_1) = 1$ and $\Pr(\hat{Y} = 1 \mid A = a_2) = \Pr(\hat{Y} = 1 \mid A = a_3) = 1/2$. Then $\Pr(\hat{Y} = 1) = 1/100 + 99/200 = 101/200$. Thus $u(a_1) = 1 - 101/200 = 99/200$, $w(a_1) = 99/20000$, $u(a_2) = u(a_3) = 101/200 - 1/2 = 1/200$, and $w(a_2) = w(a_3) = 1/200000$. Which means that $w^* = 99/20000$ and the model is $(1/200, 0.99)$-probabilistically fair. Now suppose that $\Pr(\hat{Y} = 1 \mid A = a_1) = 1$, $\Pr(\hat{Y} = 1 \mid A = a_2) = 1$, and $\Pr(\hat{Y} = 1 \mid A = a_3) = 0$. Then we have $\Pr(\hat{Y} = 1) = 101/200$. Thus $u(a_1) = 99/200$, $w(a_1) = 99/20000$, $u(a_2) = 99/200$, $w(a_2) = 99/200000$, $u(a_3) = 101/200$, and $w(a_3) = 101/200000$. Which means that $w^* = 99/20000$ and the model is $(101/200, 0.99)$-probabilistically fair. Here we see from these two examples, that 99% population of their population saw their unfairness multiply by about a 100 times while w^* did not change. But probabilistic unfairness did manage to capture this change.

What we see is that when there is a high number of protected groups, relatively bigger groups tend to determine the weighted measure of unfairness w^*, but they can still consist of only a very small part of the total population overall.

We present here two simple inequalities relating (ϵ, δ) probabilistic fairness with w^* and $\mathbb{E}[U]$.

$$w^* \leq \max\{\sup_{\mathcal{A}} p_A \epsilon, \delta u^*\} \tag{26}$$

$$\mathbb{E}[U] \leq \epsilon + \delta u^* \tag{27}$$

Proof. Let $\mathcal{A}_\epsilon = \{a \in \mathcal{A} \mid u(a) \leq \epsilon\}$ and $\mathcal{A}_\epsilon^C$ its complementary set.

If $a^* = \arg\max w(a) \in \mathcal{A}_\epsilon$ then $w(a^*) = p_A(a^*)u(a^*) \leq \sup_{\mathcal{A}} p_A \epsilon$. Otherwise using that $\Pr(\mathcal{A}_\epsilon^C) \leq \delta$, we have $w(a) \leq \delta u^*$.

Now for $\mathbb{E}[U]$:

$$\begin{aligned}
\mathbb{E}[U] &= \sum_{a \in \mathcal{A}_\epsilon} p_A(a)u(a) + \sum_{a \in \mathcal{A}_\epsilon^C} p_A(a)u(a) \\
&\leq \sum_{a \in \mathcal{A}_\epsilon} p_A(a)\epsilon + \sum_{a \in \mathcal{A}_\epsilon^C} p_A(a)u^* \\
&= \epsilon \Pr(\mathcal{A}_\epsilon) + u^* \Pr(\mathcal{A}_\epsilon) \\
&\leq \epsilon + \delta u^*.
\end{aligned}$$

□

A.4 Intersectional Fairness and Continuous Protected Attributes

Here we will show that when A is continuous, even for reasonable distributions, we might end up with $u^* = \infty$. Whereas our definition of probabilistic unfairness still has finite values, and can therefore be used as an interpretable tool to compare unfairness across models.

Suppose that we have a random vector $(A_1, A_2, ..., A_d, \hat{Y})$ distributed according to a multivariate normal $\mathcal{N}(\mu, \Sigma)$ with μ and Σ the mean and covariance. Because $\hat{Y}$ is continuous, we will instead use the density $f_{\hat{Y}|A}$ in the definition of u^*. Because this vector is distributed according to a multivariate

normal, the conditional distribution is still normal and we can derive the exact parameters. The conditional distribution is

$$\hat{Y} \mid A = a \sim \mathcal{N}(\bar{\mu}, \bar{\Sigma}) \tag{28}$$

with $\bar{\mu}$ a linear form in a, and $\bar{\Sigma}$ that depends only in Σ. Basically, we can make the mean go to ∞ by making a go to ∞. Hence for a given $y \in \mathcal{Y}$, we have that $\inf_{a \in \mathcal{A}} f_{\hat{Y}|A}(y \mid a) = 0$ for all y, which means that the unfairness is always infinite.

Whereas our notion of probabilistic fairness, is finite and computationally tractable as we need to evaluate $\delta = \mathbb{E}[\mathbb{1}[U > \epsilon]]$. It goes to 1 as ϵ goes to ∞.

If we want to compare two machine learning models, and we do not want to compare for a specific point δ, then $\epsilon^*(\delta)$ can be seen as a function of ϵ, and we can compare these functions. If for two models h_1 and h_2 one function is always above the other, we could say that one is more fair than the other.

As an example, we consider for $d = 10$ the couples $(A, \hat{Y}) \sim \mathcal{N}(0, \Sigma_w)$ with Σ_w generated through a Wishart distribution, and $(A, \hat{Y}) \sim \mathcal{N}(0, \Sigma_c)$ with $\Sigma_c = (\mathbf{1} + I_d)/2$ and $\mathbf{1}$ the constant matrix equal to 1. We then compute the probabilistic fairness on Figure 3 by computing the expectation of $\mathbb{1}[U > \epsilon]$.

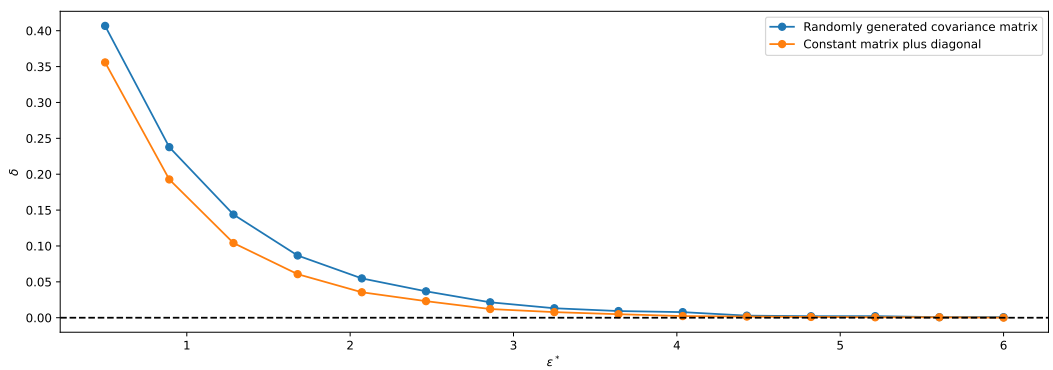

Figure 3: Probabilistic Unfairness for continuous protected attributes.

There are other fairness metrics specifically for continuous attributes, such as in [25] the HGR coefficient between A and $\hat{Y}$, but which may be less interpretable to decision makers.

B Missing proofs and elements of part 3

B.1 Counter example with independence of the sensitive attributes

Let us define the following probability distribution on $(A_1, A_2, \hat{Y})$, with A_1, A_2, and $\hat{Y}$ binary:

$$\begin{aligned}
\frac{3}{16} &= \Pr(A_1 = 0, A_2 = 0, \hat{Y} = 0) \\
&= \Pr(A_1 = 1, A_2 = 1, \hat{Y} = 0) \\
&= \Pr(A_1 = 0, A_2 = 1, \hat{Y} = 1) \\
&= \Pr(A_1 = 1, A_2 = 0, \hat{Y} = 1) \\
\text{and} \quad \frac{1}{16} &= \Pr(A_1 = 0, A_2 = 0, \hat{Y} = 1) \\
&= \Pr(A_1 = 0, A_2 = 1, \hat{Y} = 0) \\
&= \Pr(A_1 = 1, A_2 = 0, \hat{Y} = 0) \\
&= \Pr(A_1 = 1, A_2 = 1, \hat{Y} = 1)
\end{aligned}$$

We have $p_A = 1/4$, and $p_{A_1} = p_{A_2} = 1/2$. Therefore $A_1 \perp\!\!\!\perp A_2$. We have $p_{\hat{Y}} = 1/2$, hence $p_{A_1|\hat{Y}}(0 \mid 0) = p_{A_2|\hat{Y}}(0 \mid 0) = 1/2$ and $p_{A_1,A_2|\hat{Y}}(0,0,\mid 0) = 3/8 \neq p_{A_2|\hat{Y}}(0 \mid 0)p_{A_1|\hat{Y}}(0 \mid 0)$, therefore the A_k are not independent conditionally on $\hat{Y}$. Because $p_{A_1|\hat{Y}} = p_{A_2|\hat{Y}} = 1/2$, any form of marginal unfairness is 0, and $u^* = \log((3/4)/(1/4)) = \log(3) \neq 0$. In this example we have mutual independence of the A_k, independence between A_k and $\hat{Y}$, but still no meaningful relationship between intersectional and marginal fairness because we did not have independence conditionally on $\hat{Y}$.

B.2 Proof of Proposition 3.1

Using the assumed independence, for any a in $\mathcal{A}$ and y in $\mathcal{Y}$ we can rewrite $p_{\hat{Y}|A}$ with marginal quantities:

$$\Pr(\hat{Y} = y \mid A = a) = \frac{\Pr(A = a \mid \hat{Y} = y)\Pr(\hat{Y} = y)}{\Pr(A = a)} \tag{29}$$

$$= \Pr(\hat{Y} = y)\frac{\prod_{k=1}^{d}\Pr(A_i = a_i \mid \hat{Y} = y)}{\prod_{k=1}^{d}\Pr(A_i = a_i)} \tag{30}$$

$$= \Pr(\hat{Y} = y)\prod_{k=1}^{d}\frac{\Pr(\hat{Y} = y \mid A_i = a_i)}{\Pr(\hat{Y} = y)}. \tag{31}$$

Because the numerator is a product of independent variables (in the functional sense), taking the sup in $\mathcal{A}$ yields:

$$\sup_{a\in\mathcal{A}}\Pr(\hat{Y} = y \mid A = a) = \Pr(\hat{Y} = y)\prod_{k=1}^{d}\frac{\sup_{a_k\in\mathcal{A}_k}\Pr(\hat{Y} = y \mid A_k = a_k)}{\Pr(\hat{Y} = y)}. \tag{32}$$

We can do the same for $\inf$. Hence

$$\frac{\sup_{a\in\mathcal{A}}\Pr(\hat{Y} = y \mid A = a)}{\inf_{a\in\mathcal{A}}\Pr(\hat{Y} = y \mid A = a)} = \prod_{k=1}^{d}\frac{\sup_{a_k\in\mathcal{A}_k}\Pr(\hat{Y} = y \mid A_k = a_k)}{\inf_{a_k\in\mathcal{A}_k}\Pr(\hat{Y} = y \mid A_k = a_k)}, \tag{33}$$

and we obtain

$$u^* = \sup_{y\in\mathcal{Y}}\sup_{(a,a')\in\mathcal{A}^2}\Big|\log\Big(\frac{\Pr(\hat{Y} = y \mid A = a)}{\Pr(\hat{Y} = y \mid A = a')}\Big)\Big| = \sup_{y\in\mathcal{Y}}\sum_{k=1}^{d}\sup_{(a_k,a'_k)\in\mathcal{A}_k^2} u(y, a_k, a'_k). \tag{34}$$

The inequality is obtained by triangle inequality and because $\sup_{y\in\mathcal{Y}}\sup_{(a_k,a'_k)\in\mathcal{A}_k^2} u_k(y, a_k, a'_k) = u_k^*$ by definition.

B.3 Proof of Theorem 3.2 and Corollary 3.3

Theorem. For $\delta \in (0, 1]$, any classifier h over a distribution $\mathcal{D}$ is (ϵ_1, δ) and (ϵ_2, δ)-probably intersectionally fair with

$$\epsilon_1 = 2\sqrt{2}\frac{s^*}{\sqrt{\delta}} + \sup_{y\in\mathcal{Y}}\Big\{\sum_{k=1}^{d}\sup_{(a_k,a'_k)\in\mathcal{A}_k^2} u_k(y, a_k, a'_k)\Big\}$$

$$\epsilon_2 = \sqrt{2}\frac{s^*}{\sqrt{\delta}} + \gamma + \sup_{y\in\mathcal{Y}}\Big\{\sum_{k=1}^{d}\log\Big(\frac{p_{\hat{Y}}^{1-1/d}(y)}{\inf_{a_k\in\mathcal{A}_k} p_{\hat{Y}|A_k}(y\,|\,a_k)}\Big)\Big\}$$

where $\quad s^* = (\sigma^{2/3} + \sigma_y^{2/3})^{3/2} \quad$ and $\quad \gamma = C(A) - C(A\,|\,\hat{Y}) = \big(\sum_{k=1}^{d} I(A_k, \hat{Y})\big) - I(A, \hat{Y})$.

Proof. We want to show that our classifier is (ϵ, δ) probably fair for a given δ.

We will first bound in probability L and L_y, to be able to approach the joint densities through the product of marginal densities. We will denote by μ, μ_y, σ and σ_y the expectations and variances of L and L_y Let us apply Chebyshev's inequality to L. We obtain that

$$\Pr(|L-\mu| \geq \alpha_1) \leq \frac{\sigma^2}{\alpha_1^2}.$$

Using the fact that $\{|L-\mu| < \alpha_1\} \subset \{|L-\mu| \leq \alpha_1\}$ and taking the complementary event we can write that

$$\Pr(|L-\mu| \leq \alpha_1) \geq 1 - \Pr(|L-\mu| > \alpha_1) \geq 1 - \frac{\sigma^2}{\alpha_1^2}.$$

From this inequality we have

$$\begin{aligned} |L-\mu| \leq \alpha_1 \implies & \begin{cases} L-\mu \leq \alpha_1 \\ \mu - L \leq \alpha_1 \end{cases} \\ \implies & \begin{cases} L \leq \alpha_1 + \mu \\ L \geq \mu - \alpha_1 \end{cases} \\ \implies & \begin{cases} p_A(A) \leq e^{\alpha_1+\mu} \prod p_{A_k}(A_k) \\ p_A(A) \geq e^{\mu-\alpha_1} \prod p_{A_k}(A_k) \end{cases} . \end{aligned}$$

We can do the same for L_y with a parameter $\alpha_2 > 0$.

Now we want to consider a condition on the parameters $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$ so that the probability of the conjunction of the events $\{|L-\mu| \leq \alpha_1\}$ and $\{|L_y - \mu_y| \leq \alpha_2\}$ is greater than $1-\delta$. For $\delta > 0$, $\alpha_1 > 0$ and $\alpha_2 > 0$, a sufficient condition is that $\frac{\sigma^2}{\alpha^2} + \frac{\sigma_z^2}{\alpha_2^2} \leq \delta$. We can show this using complementary event and Boole's inequality:

$$\begin{aligned} & \Pr(\{|L-\mu| \leq \alpha_1\} \cap \{|L_y - \mu_y| \leq \alpha_2\}) \\ & \geq \Pr(\{|L-\mu| \leq \alpha_1\}) + \Pr(\{|L_y - \mu_y| \leq \alpha_2\}) - 1 \\ & \geq (1 - \frac{\sigma^2}{\alpha^2}) + (1 - \frac{\sigma_y^2}{\alpha_2^2}) - 1 \\ & \geq 1 - \delta. \end{aligned}$$

We define $g(\boldsymbol{\alpha}) = \frac{\sigma^2}{\alpha_1^2} + \frac{\sigma_z^2}{\alpha_2^2} - \delta$.
For any $\boldsymbol{\alpha}$ such that $g(\boldsymbol{\alpha}) \leq 0$ we have with probability at least $1-\delta$ that

$$\begin{aligned} p_{\hat{Y}|A}(\hat{Y} \mid A) &= \frac{p_{A|\hat{Y}}(A \mid \hat{Y}) p_{\hat{Y}}(\hat{Y})}{p_A(A)} && (35) \\ &\leq p_{\hat{Y}}(\hat{Y}) \frac{e^{\alpha_2+\mu_y}}{e^{\mu-\alpha_1}} \frac{\prod_{k=1}^d p_{A_k|\hat{Y}}(A_k \mid \hat{Y})}{\prod_{k=1}^d p_{A_k}(A_k)} && (36) \\ &\leq p_{\hat{Y}}(\hat{Y}) \frac{e^{\alpha_2+\mu_y}}{e^{\mu-\alpha_1}} \frac{\prod_{k=1}^d p_{\hat{Y}|A_k}(\hat{Y} \mid A_k)}{p_{\hat{Y}}(\hat{Y})^d} && (37) \\ &= p_{\hat{Y}}(\hat{Y}) \varphi(\mu_y, \mu) \psi(\boldsymbol{\alpha}) f(\hat{Y}, A), && (38) \end{aligned}$$

where $\varphi(\mu_y, \mu) = e^{\mu_y - \mu}$, $\psi(\boldsymbol{\alpha}) = e^{\alpha_1+\alpha_2}$, and $f(y,a) = \prod_{k=1}^d p_{\hat{Y}|A_k}(y \mid a_k)/p_{\hat{Y}}(y)^d$. Hence by taking the $\sup$ over a and $\inf$ over $\boldsymbol{\alpha}$ on the right hand-side, we obtain

$$p_{\hat{Y}|A}(A \mid \hat{Y}) \leq p_{\hat{Y}}(\hat{Y}) \varphi(\mu_y, \mu) \inf_{g(\boldsymbol{\alpha}) \leq 0} \psi(\boldsymbol{\alpha}) \sup_{a \in \mathcal{A}} f(y,a).$$

As it is a product of functions of independent variables, $\sup_{a\in\mathcal{A}} f(y,a)$ is just the product of the $\sup$ of each $p_{\hat{Y}|A_k}$.

We will now solve the constrained optimization problem for ψ. We can write $\inf e^{\alpha_1+\alpha_2} = e^{\inf(\alpha_1+\alpha_2)}$, so we will just need to solve the simpler problem $\inf_{g(\alpha)\leq 0} s(\alpha)$, with $s(\alpha) = \alpha_1 + \alpha_2$. Let us compute the gradients of s and g:

$$\nabla g = (-2\sigma^2\alpha_1^{-3}, -2\sigma_y^2\alpha_2^{-3})^\top,$$
$$\nabla s = (1,1)^\top.$$

We will now show that this is a convex problem. The function s is linear thus convex, and we will now compute the hessian of g:

$$H_g = 6 \cdot \begin{pmatrix} \sigma^2\alpha_1^{-4} & 0 \\ 0 & \sigma_y^2\alpha_2^{-4} \end{pmatrix}.$$

Clearly we have that the determinant of H_g, $Det(H_g)$ is strictly positive. Therefore H_g is definite positive, and g is convex. And for $\delta > 0$ there is a feasible interior point by taking α_1 and α_2 big enough, which means that Slater's conditions hold (e.g. a convex constraint with a feasible interior point). We will now analyze the KKT conditions for minimization with the dual parameter $c \geq 0$:

$$\begin{cases} \nabla s + c\nabla g = 0 \\ cg(\boldsymbol{\alpha}) = 0 \end{cases} \Leftrightarrow \begin{cases} 1 = 2c\sigma^2\alpha_1^{-3} \\ 1 = 2c\sigma_y^2\alpha_2^{-3} \\ cg(\alpha_1,\alpha_2) = 0 \end{cases}.$$

We obtain that $\alpha_1 = \sqrt[3]{2c\sigma^2}$ and $\alpha_2 = \sqrt[3]{2c\sigma_y^2}$
Clearly $c > 0$ otherwise the first two lines cannot be 1, hence using the last equation we have $g(\alpha_1,\alpha_2) = 0$. We now develop this last equality to obtain c:

$$g(\boldsymbol{\alpha}) = 0 \implies \frac{\sigma^2}{(2c\sigma^2)^{2/3}} + \frac{\sigma_y^2}{(2c\sigma_y^2)^{2/3}} = \delta$$
$$\implies c = \frac{1}{2}\left(\frac{\sigma^{2/3}+\sigma_y^{2/3}}{\delta}\right)^{3/2}.$$

Plugging c in the previous expressions we have

$$\implies \begin{cases} \alpha_1^* = \sqrt{\left(\frac{\sigma^{2/3}+\sigma_y^{2/3}}{\delta}\right)}\sqrt[3]{\sigma^2} = \frac{s_1^*}{\sqrt{\delta}} \\ \alpha_2^* = \sqrt{\left(\frac{\sigma^{2/3}+\sigma_y^{2/3}}{\delta}\right)}\sqrt[3]{\sigma_y^2} = \frac{s_2^*}{\sqrt{\delta}} \end{cases}$$
$$\text{with} \quad \begin{cases} s_1^* = \sqrt{\sigma^{2/3}+\sigma_y^{2/3}}\sigma^{2/3} \\ s_2^* = \sqrt{\sigma^{2/3}+\sigma_y^{2/3}}\sigma_y^{2/3} \end{cases}.$$

Finally the minimum is

$$\inf_{g(\boldsymbol{\alpha})\leq 0} s = \frac{s_1^*+s_2^*}{\sqrt{\delta}} = \frac{\left(\sigma^{2/3}+\sigma_y^{2/3}\right)^{3/2}}{\sqrt{\delta}} = \frac{s^*}{\sqrt{\delta}}$$
$$\text{with} \quad s^* = \left(\sigma^{2/3}+\sigma_y^{2/3}\right)^{3/2}.$$

We will now do the same in order to lower bound $p_A(\hat{Y}|A)$.

$$
\begin{aligned}
p_{\hat{Y}|A}(\hat{Y} \mid A) &= \frac{p_{A|\hat{Y}}(A \mid \hat{Y}) p_{\hat{Y}}(\hat{Y})}{p_A(A)} && (39)\\
&\geq p_{\hat{Y}}(\hat{Y}) \frac{e^{-\alpha_2+\mu_y}}{e^{\mu+\alpha_1}} \frac{\prod_{k=1}^d p_{A_k|\hat{Y}}(A_k \mid \hat{Y})}{\prod_{k=1}^d p_{A_k}(A_k)} && (40)\\
&\geq p_{\hat{Y}}(\hat{Y}) \frac{e^{-\alpha_2+\mu_y}}{e^{\mu+\alpha_1}} \frac{\prod_{k=1}^d p_{\hat{Y}|A_k}(\hat{Y} \mid A_k)}{p_{\hat{Y}}(\hat{Y})^d} && (41)\\
&= p_{\hat{Y}}(\hat{Y}) \varphi(\mu_y, \mu) \psi(\boldsymbol{\alpha})^{-1} f(\hat{Y}, A), && (42)
\end{aligned}
$$

Here we take the $\sup$ over $\boldsymbol{\alpha}$ and $\inf$ over a instead. We have that $\sup \psi(\boldsymbol{\alpha})^{-1} = (\inf \psi(\boldsymbol{\alpha}))^{-1} = \exp(-s^*/\sqrt{\delta})$.

Because U involves the two variables A and A', we need to bound L' and L'_y the variables L and L_y that are taken as a function of A' instead of A. Because $(A', \hat{Y}) \sim (A, \hat{Y})$, all the computations above still apply, and we have

$$
\Pr(\{|L-\mu| \leq \alpha_1^*\} \cap \{|L_y - \mu_y| \leq \alpha_2^*\} \cap \{|L' - \mu| \leq \alpha_1^*\} \cap \{|L'_y - \mu_y| \leq \alpha_2^*\}) \geq 1 - 2\delta.
$$

Hence we only need to replace δ by $\delta/2$ in the above inequalities.

Combining everything, we can conclude that when the event $\{|L-\mu| \leq \alpha_1^*\} \cap \{|L_y - \mu_y| \leq \alpha_2^*\} \cap \{|L' - \mu| \leq \alpha_1^*\} \cap \{|L'_y - \mu_y| \leq \alpha_2^*\}$ occurs, we have

$$
\begin{aligned}
U &= \left| \log \left(\frac{p_{\hat{Y}|A}(\hat{Y} \mid A)}{p_{\hat{Y}|A'}(\hat{Y} \mid A')} \right) \right| \\
&\leq \log \left(\frac{p_{\hat{Y}}(\hat{Y}) \varphi(\mu_y, \mu) \exp(\sqrt{2} s^*/\sqrt{\delta}) \sup_{\mathcal{A}} f(\hat{Y}, a)}{p_{\hat{Y}}(\hat{Y}) \varphi(\mu_y, \mu) \exp(-\sqrt{2} s^*/\sqrt{\delta}) \inf_{\mathcal{A}} f(\hat{Y}, a)} \right) \\
&= 2\sqrt{2} \frac{s^*}{\sqrt{\delta}} + \log \left(\frac{\prod_{k=1}^d \sup_{a_k \in \mathcal{A}_k} p_{\hat{Y}|A_k}(a_k)}{\prod_{k=1}^d \inf_{a_k \in \mathcal{A}_k} p_{\hat{Y}|A_k}(a_k)} \right) \\
&\leq 2\sqrt{2} \frac{s^*}{\sqrt{\delta}} + \sup_{y \in \mathcal{Y}} \sum_{k=1}^d \sup_{(a,a') \in \mathcal{A}^2} u_k(y, a, a').
\end{aligned}
$$

We can conclude that

$$
\begin{aligned}
&\Pr(U \leq \epsilon_1) \geq 1 - \delta \\
&\text{with} \quad \epsilon_1 = 2\sqrt{2} \frac{s^*}{\sqrt{\delta}} + \sup_{y \in \mathcal{Y}} \sum_{k=1}^d \sup_{(a,a') \in \mathcal{A}^2} u_k(y, a, a'),
\end{aligned}
$$

which means that our classifier is (ϵ_1, δ)-probably intersectionally fair. Note that ϵ_1 is a function of $(\delta, \sigma, \sigma_y)$.

In order to derive the proof for ϵ_2, we simply remark that $p_{\hat{Y}|A} \leq 1$ which can be used to upper bound the numerator. Therefore when the event $\{|L-\mu| \leq \alpha_1^*\} \cap \{|L_y - \mu_y| \leq \alpha_2^*\} \cap \{|L' - \mu| \leq$

$\alpha_1^*\} \cap \{|L'_y - \mu_y| \leq \alpha_2^*\}$ occurs, we have

$$
\begin{aligned}
U &= \left| \log \left(\frac{p_{\hat{Y}|A}(\hat{Y} \mid A)}{p_{\hat{Y}|A'}(\hat{Y} \mid A')} \right) \right| \\
&\leq \log \left(\frac{1}{p_{\hat{Y}}(\hat{Y}) \varphi(\mu_y, \mu) \exp(-\sqrt{2} s^* / \sqrt{\delta}) \inf_{\mathcal{A}} f(\hat{Y}, a)} \right) \\
&= \sqrt{2} \frac{s^*}{\sqrt{\delta}} + \gamma + \sum_{k=1}^{d} \log \Big(\frac{p_{\hat{Y}}^{1-1/d}(\hat{Y})}{\inf_{a_k \in \mathcal{A}_k} p_{\hat{Y}|A_k}(\hat{Y} \,|\, a_k)} \Big) \\
&\leq \sqrt{2} \frac{s^*}{\sqrt{\delta}} + \gamma + \sup_{y \in \mathcal{Y}} \Big\{ \sum_{k=1}^{d} \log \Big(\frac{p_{\hat{Y}}^{1-1/d}(y)}{\inf_{a_k \in \mathcal{A}_k} p_{\hat{Y}|A_k}(y \,|\, a_k)} \Big) \Big\} \\
&= \epsilon_2.
\end{aligned}
$$

We can conclude in the same way as for ϵ_1.

Looking at the proof, it also holds true that the model will also be $(\min\{\epsilon_1, \epsilon_2\}, \delta)$-probably intersectionally fair. □

Now let us prove Corollary 3.3. Recall we now suppose that we are interested in only one outcome $y \in \mathcal{Y}$, and we redefine our notions of unfairness and probabilistic unfairness, as well as s^* and γ, as done in (15).

Corollary. Denoting $(\Omega, \mathcal{T}, \Pr)$ the probability space on which (A, A') is defined, there exists an event F so that for $f(a) = \prod_{k=1}^{d} p_{\hat{Y}=y|A_k}(a) / p_{\hat{Y}=y}^{d}$ we have

$$
\sup_{\omega \in F} p_{\hat{Y}=y|A}(A(\omega)) \in [p_{\hat{Y}}(y) e^{\frac{-2\sqrt{2}s^*}{\sqrt{\delta}} - \gamma} \sup_{\omega \in F} f(A)(\omega), p_{\hat{Y}}(y) e^{\frac{2\sqrt{2}s^*}{\sqrt{\delta}} - \gamma} \sup_{\omega \in F} f(A(\omega))], \tag{43}
$$

$$
\inf_{\omega \in F} p_{\hat{Y}=y|A}(A(\omega)) \in [p_{\hat{Y}}(y) e^{\frac{-2\sqrt{2}s^*}{\sqrt{\delta}} - \gamma} \inf_{\omega \in F} f(A)(\omega), p_{\hat{Y}}(y) e^{\frac{2\sqrt{2}s^*}{\sqrt{\delta}} - \gamma} \inf_{\omega \in F} f(A(\omega))], \tag{44}
$$

$$
\text{and} \quad \Pr(F \mid \hat{Y} = y) \geq 1 - \delta, \tag{45}
$$

and the same inequalities hold for $\sup_{\omega \in F} p_{\hat{Y}=y|A}(A')(\omega)$ for the same event F.

Proof. Let F be the event defined as follows:

$$
F = \{\omega \in \Omega \mid p_{\hat{Y}|A}(y \mid A(\omega)) \leq p_{\hat{Y}}(y) \exp\big(\frac{2\sqrt{2}s^*}{\sqrt{\delta}} - \gamma\big) f(A)(\omega), \tag{46}
$$

$$
p_{\hat{Y}|A'}(y \mid A'(\omega)) \leq p_{\hat{Y}}(y) \exp\big(\frac{2\sqrt{2}s^*}{\sqrt{\delta}} - \gamma\big) f(A')(\omega), \tag{47}
$$

$$
p_{\hat{Y}|A}(y \mid A(\omega)) \geq p_{\hat{Y}}(y) \exp\big(\frac{-2\sqrt{2}s^*}{\sqrt{\delta}} - \gamma\big) f(A)(\omega), \tag{48}
$$

$$
p_{\hat{Y}|A'}(y \mid A'(\omega)) \geq p_{\hat{Y}}(y) \exp\big(\frac{-2\sqrt{2}s^*}{\sqrt{\delta}} - \gamma\big) f(A')(\omega)\}. \tag{49}
$$

What we have shown in the proof of the theorem, is that the event $\{|L - \mu| \leq \alpha_1^*\} \cap \{|L_y - \mu_y| \leq \alpha_2^*\} \cap \{|L' - \mu| \leq \alpha_1^*\} \cap \{|L'_y - \mu_y| \leq \alpha_2^*\}$ is included in F, and hence $\Pr(F \mid \hat{Y} = y) \geq 1 - \delta$. Now we just need to take the $\sup$ and $\inf$ in ω over F for these 4 inequalities in the definition of F to directly obtain the corollary. Note that we did not obtain a direct statement on $U(\omega)$ because lower bounding the $\sup$ of a ratio of two functions is not easy.

□

B.4 Additional Bounds on Probable Intersectional Fairness

Looking at how we proved Theorem 3.2, we can derive more bounds by using other concentration inequalities. Let $\kappa(t) = \log(\mathbb{E}[e^{tL}])$ and $\kappa_y(t) = \log(\mathbb{E}[e^{tL_y}])$ be the cumulant generating-function of L and L_y. We define the α-Renyi Divergence between two discrete distributions P and Q of size S for $\alpha > 0$ as

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1}\log(\sum_{k=1}^{S}\frac{p_k^\alpha}{q_k^{\alpha-1}}). \tag{50}$$

These moments generating functions can be expressed as functions of Renyi Divergences, indeed

$$\kappa(t) = tD_{t+1}(p_A\|\bigotimes_{k=1}^{d} p_{A_k}) \tag{51}$$

$$\text{and} \quad \kappa_y(t) = \log(\sum_{y\in\mathcal{Y}} p_{\hat{Y}}(y)\exp(tD_{t+1}(p_{A|\hat{Y}}(\cdot \mid y)\|\bigotimes_{k=1}^{d} p_{A_k|\hat{Y}}(\cdot \mid y)))). \tag{52}$$

While $\kappa(t)$ can therefore be estimated using techniques of [15] for instance, the estimation of κ_y is less straightforward.

For λ^+ and λ^- in $\mathbb{R}$, we define $I_y^+(\lambda^+) = \sup_{t\in\mathbb{R}^+}\{t\lambda^+ - \kappa_y(t)\}$ and $I^-(\lambda^-) = \sup_{t\in\mathbb{R}^-}\{t\lambda^- - \kappa(t)\}$. We apply the generic Chernoff bounds to L and L_y:

$$\Pr(L_y \geq \lambda^+) \leq e^{-I_y^+(\lambda^+)} \tag{53}$$

$$\text{and} \quad \Pr(L \leq \lambda^-) \leq e^{-I^-(\lambda^-)}. \tag{54}$$

We will apply the same reasoning used in Appendix $B.3$ and will only highlight the differences.

$$\Pr([L_y \geq \lambda^+] \cap [L \leq \lambda^-]) \geq 1 - (e^{-I_y^+(\lambda^+)} + e^{-I^-(\lambda^-)})$$

We want to ensure that the right hand-side is greater than $1 - \delta$. Because we need to make sure that it also holds for L' and L'_y, we need 2δ instead of δ. We define the constraint

$$g(\lambda^+, \lambda^-) = e^{-I_y^+(\lambda^+)} + e^{-I^-(\lambda^-)} - 2\delta, \tag{55}$$

we need to have $g(\lambda^+, \lambda^-) \leq 0$. Hence using feasible values of λ^+ and λ^-, the event $[L_y \geq \lambda^+] \cap [L \leq \lambda^-]$ implies

$$U \leq (\lambda^+ - \lambda^-) + \sup_{\mathcal{Y}}\sum_{k=1}^{d}\log(\frac{p_{\hat{Y}}^{1-1/d}}{\inf_{\mathcal{A}_k} p_{\hat{Y}|A_k}}) \tag{56}$$

$$\leq \inf_{g(\boldsymbol{\lambda})\leq 0}(\lambda^+ - \lambda^-) + \sup_{\mathcal{Y}}\sum_{k=1}^{d}\log(\frac{p_{\hat{Y}}^{1-1/d}}{\inf_{\mathcal{A}_k} p_{\hat{Y}|A_k}}) \tag{57}$$

We therefore have the following Theorem:

Theorem B.1. *For $\delta \in (0, 1]$, any classifier h over a distribution $\mathcal{D}$ is (ϵ_3, δ)-probably intersectionally fair, with*

$$\epsilon_3 = \inf_{g(\boldsymbol{\lambda})\leq 0}(\lambda^+ - \lambda^-) + \sup_{\mathcal{Y}}\sum_{k=1}^{d}\log(\frac{p_{\hat{Y}}^{1-1/d}}{\inf_{\mathcal{A}_k} p_{\hat{Y}|A_k}}) \tag{58}$$

Compared to using Chebyshev, this should be both tighter as we are using more information than the first and second moment, and this bound should also be more efficient in terms of probability, as we are using one sided concentration inequalities.

B.5 Properties of the cumulant generating-function

We now want to be able to say a bit more on the properties of this constrained minimization problem. We will show by first recalling and developing useful properties on κ that the problem is non convex and differentiable almost everywhere.

We will list the properties about κ that will be useful to us developed in [9], with [23] being a summary containing all the information needed.

Lemma B.2. *We have that κ is strictly convex and infinitely many times differentiable. This means that κ' is strictly increasing and that it can be inverted on $\mathrm{Im}(\kappa')$. We will write $\kappa'^{(-1)} = \eta$, and for all x so that $\kappa''(\eta(x)) \neq 0$ we have $\eta' = 1/\kappa''(\eta)$.*

We also have that $\kappa(0) = 0$, $\kappa'(0) = \mu$, and $\kappa''(0) = \sigma^2$.

From these properties we can conclude that $\eta(\mu) = 0$, and that $\eta'(\mu) = 1/\kappa''(\eta(\mu)) = 1/\sigma^2$.

We will recall the definition of the convex conjugate of a function.

Definition B.3. Let $\mathcal{E}$ be a Euclidean vector space with scalar product $\langle \cdot, \cdot \rangle$, we define the convex-conjugate of a function $f : \mathcal{E} \to \mathbb{R}$ for $x \in \mathcal{E}$ by

$$f^\star(x) = \sup_{t \in \mathcal{E}} \{\langle x, t \rangle - f(t)\}. \tag{59}$$

We will now list the useful properties about $I = \kappa^\star$ also developed in [23].

Lemma B.4. *The function I is infinitely many times differentiable on $\mathrm{Im}(\kappa')$, $I(\mu) = 0$, and for every $x \in \mathrm{Im}(\kappa')$ we can rewrite I as*

$$I(x) = x\eta(x) - \kappa(\eta(x)). \tag{60}$$

Proposition B.5. *The function I^+ is continuously differentiable on $(-\infty, \sup \kappa')$.*

Proof. We define $\tilde{I}^+(x) = I(x)$ if $x \geq \mu$, and $\tilde{I}^+(x) = 0$ otherwise. We will first show that $\tilde{I}^+ = I^+$.

We have $\mu \in \mathrm{Im}(\kappa')$ because $\kappa'(0) = \mu$. Let $f_x(t) = tx - \kappa(t)$. If $x < \mu$, then because κ' is increasing we have for any $t \geq 0$

$$\begin{aligned} f_x'(t) &= x - \kappa'(t) \\ &\leq \mu - \kappa'(0) \\ &= 0. \end{aligned}$$

This means that the $\max$ on $\mathbb{R}^+$ is at 0, and therefore $I^+(x) = f_x(0) = 0x - \kappa(0) = 0 = \tilde{I}^+(x)$.

If $x \geq \mu$, then for any $t \leq 0$ we have

$$\begin{aligned} f_x'(t) &= x - \kappa'(t) \\ &\geq \mu - \kappa'(0) \\ &= 0. \end{aligned}$$

Hence for all $t \leq 0$ we have $f_x(t) \leq f_x(0)$ therefore the sup of f_x is not on $\mathbb{R}^-$. Consequently when $x \geq \mu$ we have $I^+(x) = \tilde{I}^+(x)$. All in all we can conclude that $\tilde{I}^+ = I^+$ on $\mathrm{Im}(\kappa')$.

Let us analyze the potential discontinuity at μ. We have $I(\mu) = 0$ and $I^+ = 0$ on $(-\infty, \mu)$, so the function is continuous on $(-\infty, \sup \kappa')$. Let us compute I':

$$\begin{aligned} I'(x) &= \eta(x) + x\eta'(x) - \eta'(x)\kappa'(\eta(x)) \\ &= \eta(x) + x\eta'(x) - \eta'(x)x \\ &= \eta(x), \end{aligned}$$

and we know that $\eta(\mu) = 0$. Hence we have that I^+ is continuously differentiable on $(-\infty, \sup \kappa')$. □

Proposition B.6. *The function e^{-I^+} is non-convex at μ.*

Proof. We will simply look at the second derivative of $h(x) = e^{-I^+(x)}$ for $x \geq \mu$:

$$h''(x) = (I^{+'}(x)^2 - I^{+''}(x))e^{-I(x)}$$
$$= (\eta(x)^2 - \frac{1}{\kappa''(\eta(x))})e^{-I(x)}.$$

Therefore $h''(\mu) = -1/\sigma^2 < 0$ for $\sigma \neq 0$, which means that it is non-convex at μ. □

The same proposition applies to I_y^+ and I^-, with the relevant κ or κ_y. This means that $g''((\mu_y, \mu)) = -(1/\sigma^2 + 1/\sigma_y^2) < 0$ hence the following corollary

Corollary B.7. *The constraint g is non convex at (μ_y, μ).*

Finally we remark that when $\mathcal{A}$ is finite, the sup of κ' is bounded and therefore there are finite values of λ for which $I(\lambda) = \infty$. Hence there are feasible points for any $\delta \in [0, 1]$.

B.6 Errors bounds on Q

Let $P = (p_1, ..., p_S)$ be a discrete distribution of size $|P| = S$, we want to estimate the quantity $Q(P) = \sum_{k=1}^{S} p_k \log^2(p_k)$ with n i.i.d. realizations of P. We denote N_k, the number of realizations for category k.

In order to bound the L_2 error of $\hat{Q} = \sum_{k=1}^{S}(N_k/n)\log^2(N_k/n)$, we will use the bias variance decomposition of $\hat{Q}$:

$$\mathbb{E}[(\hat{Q} - Q)^2] = b(\hat{Q})^2 + \mathrm{Var}(\hat{Q}) \tag{61}$$

$$\text{where} \quad b(\hat{Q}) = \mathbb{E}[\hat{Q}] - Q, \mathrm{Var}(\hat{Q}) = \mathbb{E}[(\hat{Q} - \mathbb{E}[\hat{Q}])^2] \tag{62}$$

The analysis of these error terms is completely derived from [20]. In particular, the method they use for entropy is close to this problem. They show that the bias term can be bounded by deriving smoothness modulus for the function $x \mapsto x\log(x)$, and that the variance term can be bounded using an Efron-Stein inequality. Here, we need to analyse $x \mapsto x\log^2(x)$, which is technically more difficult as some nice properties such as the convexity of $x\log(x)$ is lost. Still we can show the following two lemmas with the proof further down:

Lemma B.8.

$$\mathrm{Var}(Q(\hat{P}_n)) = \mathcal{O}\Big(\frac{\log^4(n)}{n}\Big). \tag{63}$$

Lemma B.9.

$$b(\hat{Q})^2 = \mathcal{O}\Big(\Big(\frac{|P|\log(n)}{n}\Big)^2\Big). \tag{64}$$

Using these two lemmas, we can directly conclude that $\mathbb{E}[(\hat{Q} - Q)^2] = \mathcal{O}(\log^4(n)/n)$.

Note that we will directly use some elements already derived in [20], and will only show here the parts where special care is needed.

Proof of Lemma B.8. Let $f : x \mapsto x\log^2(x)$. [20] analyze the statistic of the form $F(P) = \sum_{k=1}^{S} f(p_k)$. We apply Lemma 13 of [20] for discrete functionals of P, which is derived from a corollary of the Efron-Stein inequality, on Q:

$$\mathrm{Var}(\hat{Q}) \leq n \max_{0 \leq j \leq n} (f(\frac{j+1}{n}) - f(\frac{j}{n}))^2. \tag{65}$$

We will look for n in $\mathbb{N}^*$ at the function $g : x \mapsto \frac{x+1}{n}\log^2(\frac{x+1}{n}) - \frac{x}{n}\log^2(\frac{x}{n})$ for $x \in [0, n]$. We have

$$g(x) = \frac{x}{n}\log(\frac{x+1}{x})(\log(x(x+1)) - \log(n^2)) + \frac{1}{n}\log^2(\frac{x+1}{n}) \tag{66}$$

$$= \frac{x}{n}\log(1+\frac{1}{x})(\log(x(x+1)) - \log(n^2)) + \frac{1}{n}(\log^2(x+1) + \log^2(n) - 2\log(x+1)\log(n)). \tag{67}$$

We first look at the term in $1/n \cdot h(x)$. We have that $h'(x) = 2(\log(x+1) - \log(n))/(x+1)$ which is 0 for $x = n-1$. Thus $\arg\max_{x\in[0,n]} |h(x)| \in \{0, n-1, n\}$. We evaluate h at these points: $h(n-1) = 0$, $h(n) = \log^2(1+1/n) \sim 1/n^2$, and $h(0) = \log(n)^2$. Hence the max of $|h(x)|$ over $[0,n]$ is $\log^2(n)$ Using on (67) the inequality $\log(1+1/x) \leq 1/x$, and because $x \mapsto \log(x(x+1))$ is increasing over $\mathbb{R}^+$, we obtain

$$|g(x)| \leq \frac{1}{n}(\log(n(n+1)) + \log(n^2)) + \frac{1}{n}\log^2(n) = \mathcal{O}\Big(\frac{\log^2(n)}{n}\Big) \tag{68}$$

Finally

$$\mathrm{Var}(\hat{Q}) = n\mathcal{O}\Big(\frac{\log^4(n)}{n^2}\Big) = \mathcal{O}\Big(\frac{\log^4(n)}{n}\Big). \tag{69}$$

□

Proof of Lemma B.9. In order to bound the bias [20] use the fact that for any f,

$$\mathbb{E}[\hat{Q}] - Q = \sum_{k=1}^{S}(B_n(f)(p_k) - f(p_k)) \tag{70}$$

$$\text{with} \quad B_n(f)(x) = \sum_{i=1}^{n} f(\frac{i}{n})\binom{i}{n}x^i(1-x)^{n-i}. \tag{71}$$

The function $B_n(f)(x)$ is the Bernstein polynomial of $f(x)$. Lemma 5. of [20] shows that for $\varphi(x) = \sqrt{x(1-x)}$:

$$|\mathbb{E}[B_n(f)(x)] - f(x)| \leq \frac{5}{2}\omega_\varphi^2(f, n^{-1/2}) \tag{72}$$

$$\text{with} \quad \omega_\varphi^2(f,t) = \sup\{|f(u)+f(v)-2f(\frac{u+v}{2})|, (u,v) \in [0,1]^2, |u-v| \leq 2t\varphi(\frac{u+v}{2})\}. \tag{73}$$

The quantity ω_φ^2 is the second-order Ditzian-Totik modulus of smoothness of f. It is shown in Lemma 8 of [20] that for the function $x \mapsto x\log(x)$ (which corresponds the the entropy), $\omega_\varphi^2(f,t) = t^2\log(4)/(1+t^2)$. We will use a proof similar to Lemma 8 to derive the modulus of smoothness for $x \mapsto x^2\log(x)$.

In addition, we remark using the triangle inequality that for any f and g continuous functions on $[0,1]$, then

$$\omega_\varphi^2(f+g,t) \leq \omega_\varphi^2(f,t) + \omega_\varphi^2(g,t). \tag{74}$$

Let $f : x \mapsto x\log^2(x)$ for x in $[0,1]$. By expanding $g(x) = x\log^2(x/e) = x\log^2(x) + x - 2x\log(x)$ we see that we can rewrite f as $f(x) = x\log^2(x/e) + 2x\log(x) - x$. Therefore using the previous remark and because $\omega_\varphi^2(x \mapsto x, t) = 0$ we have

$$\omega_\varphi^2(f,t) \leq \omega_\varphi^2(g,t) + 2\omega_\varphi^2(f,t) + \omega_\varphi^2(x \mapsto x, t) \leq \omega_\varphi^2(f,t) + \frac{2t^2\log(4)}{1+t^2}. \tag{75}$$

It remains to upper bound $\omega_\varphi^2(g,t)$.

First we will show that g is concave on $(0,1]$. We compute the first and second derivative of g for x in $(0,1]$:

$$g'(x) = 2\log(\frac{x}{e}) + \log^2(\frac{x}{e}) = \log^2(x) - 1, \tag{76}$$

$$\text{thus} \quad g''(x) = 2\frac{\log(x)}{x} < 0. \tag{77}$$

Hence g is strictly concave over $(0,1]$.

Now we will upper bound $\omega_\varphi^2(g,t)$. Let t in $[0,1/2]$. *To be clear, we will use the same language and logic developed in [20] for the proof of Lemma 8 to make the comparison easier.* Defining $M = (u+v)/2 \in [0,1]$, then the computation of the second order modulus is an optimization over the regime $|u-v| \leq 2t\sqrt{M(1-M)}$. Equivalently, it is in the interval $[M(1-\Delta), M(1+$

$\Delta)] \cap [0,1]$, where $\Delta = t\sqrt{(1-M)/M}$. Because g is strictly concave on $[0,1]$, the maximum of $|g(x)+g(y)-2g((x+y)/2)|$ is reached at the boundaries of the above feasible interval. We have

$$M(1-\Delta) \geq 0 \Leftrightarrow M \geq \frac{t^2}{1+t^2}, \tag{78}$$

$$M(1+\Delta) \leq 1 \Leftrightarrow M \leq \frac{1}{1+t^2}. \tag{79}$$

Therefore the optimization problem defined by the second order modulus of smoothness is equivalent to the maximization of $h(u,v) = |g(u)+g(v)-2g((u+v)/2)|$ over three different regimes:

$$\text{Regime A:} \quad u=0, v=2M, 0 \leq M \leq \frac{t^2}{1+t^2} \tag{80}$$

$$\text{Regime B:} \quad u=2M-1, v=1, 1 \geq M \geq \frac{1}{1+t^2} \tag{81}$$

$$\text{Regime C:} \quad u=M(1+\Delta), v=M(1-\Delta), M \in [\frac{t^2}{1+t^2}, \frac{1}{1+t^2}]. \tag{82}$$

Over the regime A:

$$\begin{aligned} h(u,v) &= |2M\log^2(\frac{2M}{e}) - 2M\log^2(\frac{M}{e})| \\ &= 2M|(\log(\frac{2M}{e}) - \log(\frac{M}{e}))(\log(\frac{2M}{e}) + \log(\frac{M}{e}))| \\ &= 2M\log(2)\log(\frac{1}{2}(\frac{e}{M})^2) \\ &= 2M\log(2)(2-\log(2)-2\log(M)). \end{aligned}$$

The function $M \mapsto M$ reaches its max over regime A at $t^2/(1+t^2)$. The function $x \mapsto -x\log(x)$ is positive increasing until $x = 1/e$. Hence because $t^2/(1+t^2) \leq 1/e$ for $t \in [0,1/2]$, $M \mapsto -M\log(M)$ reaches its max over regime A also at $t^2/(1+t^2)$. Thus over regime A

$$h(u,v) \leq \frac{2t^2}{1+t^2}\log(2)(2-\log(2)+2\log(1+\frac{1}{t^2})) =_{t\to 0} \frac{-8\log(2)t^2\log(t)}{1+t^2} + o(-t^2\log(t)) \tag{83}$$

Over the regime B:

$$\begin{aligned} &h(u,v) = |(2M-1)\log^2(\frac{2M-1}{e}) + 1 - 2M\log^2(\frac{M}{e})| \\ &= |(2M-1)(\log^2(2M-1) + 1 - 2\log(2M-1)) + 1 - 2M(\log^2(M) + 1 - 2\log(M))| \\ &= |2M(\log^2(2M-1) - \log^2(M)) + 4M(\log(M) - \log(2M-1)) + 2\log(2M-1) - \log^2(2M-1)| \\ &\leq |2M(\log^2(2M-1) - \log^2(M))| + |4M(\log(M) - \log(2M-1))| + |2\log(2M-1) - \log^2(2M-1)| \\ &= |2M\log(\frac{2M-1}{M})\log(M(2M-1))| + |4M(\log(\frac{M}{2M-1})| + |\log(2M-1)(2-\log(2M-1))| \end{aligned}$$

We will upper bound each of those three terms. First note that as $t \in [0,1/2]$, $t^2/(1-t^2) \leq 1/3$ and $(1+t^2)^2/(1-t^2) \leq 25/12$ Because $-\log$ is decreasing and $2M-1 \leq 1$, we have that

$$|\log(2M-1)| = -\log(2M-1) \leq \log(\frac{1+t^2}{1-t^2}) = \log(1+\frac{2t^2}{1-t^2}) \leq \frac{2t^2}{1-t^2}.$$

Hence

$$|\log(2M-1)(2-\log(2M-1))| \leq \frac{t^2}{1-t^2}\frac{16}{3} \tag{84}$$

For $M \leq 1$, we have that $(2M-1)/M \leq 1$. Hence

$$|4M\log(\frac{M}{2M-1})| = 4M\log(\frac{M}{2M-1}) \leq 4\log(\frac{M}{2M-1}) \leq 4\log(1+\frac{t^2}{1-t^2}) \leq \frac{4t^2}{1-t^2}.$$

The functions $M \mapsto 1/(M(2M-1))$ and $M \mapsto M/(2M-1)$ are decreasing in M over $(1/2, 1]$ and bigger than 1. Therefore

$$|\log(M(2M-1))| = \log(\frac{1}{M(2M-1)}) \leq \log(\frac{(1+t^2)^2}{1-t^2}) \leq \log(\frac{25}{12})$$
$$\text{and} \quad |\log(\frac{2M-1}{M})| = \log(\frac{M}{2M-1}) \leq \log(\frac{1}{1-t^2}) \leq \frac{t^2}{1-t^2}$$

Combining everything we have over regime B using that $M \leq 1$

$$h(u,v) \leq \frac{(4+16/3+\log(25/12))t^2}{1-t^2} =_{t\to 0} o(-t^2\log(t)) \tag{85}$$

Over the regime C:

$$\begin{aligned}
h(u,v) &= M|(1-\Delta)\log^2(\frac{M}{e}(1-\Delta)) + (1+\Delta)\log^2(\frac{M}{e}(1+\Delta)) - 2\log^2(\frac{M}{e})| \\
&= M|(1-\Delta)\log^2(1-\Delta)+(1+\Delta)\log^2(1+\Delta)-2\log(\frac{M}{e})((1-\Delta)\log(1-\Delta)+(1+\Delta)\log(1+\Delta))| \\
&= \frac{t^2}{t^2+\Delta^2}|(1-\Delta)\log^2(1-\Delta)+(1+\Delta)\log^2(1+\Delta)-2\log(\frac{M}{e})((1-\Delta)\log(1-\Delta)+(1+\Delta)\log(1+\Delta))| \\
&\leq \frac{t^2}{\Delta^2}|(1-\Delta)\log^2(1-\Delta)+(1+\Delta)\log^2(1+\Delta)| + \frac{2(1-\log(M))t^2}{\Delta^2}|(1-\Delta)\log(1-\Delta)+(1+\Delta)\log(1+\Delta)| \\
&\leq \frac{t^2}{\Delta^2}|(1-\Delta)\log^2(1-\Delta)+(1+\Delta)\log^2(1+\Delta)| + \frac{2(1+\log(1+\frac{1}{t^2}))t^2}{\Delta^2}|(1-\Delta)\log(1-\Delta)+(1+\Delta)\log(1+\Delta)|
\end{aligned}$$

The functions $g_1 : \Delta \mapsto ((1+\Delta)\log(1+\Delta) + (1-\Delta)\log(1-\Delta))/\Delta^2$ and $g_2 : \Delta \mapsto ((1+\Delta)\log^2(1+\Delta) + (1-\Delta)\log^2(1-\Delta))/\Delta^2$ are both continuous over $[0,1]$ hence bounded with a max reached respectively (can be seen graphically, or by looking at the derivative) at 1 and 0. With $g_1(1) = 2\log(2)$, and for $\Delta \to 0$:

$$g_2(\Delta) = \frac{(1-\Delta)(\Delta^2+\Delta^3+o(\Delta^3)) + (1+\Delta)(\Delta^2-\Delta^3+o(\Delta^3))}{\Delta^2} \longrightarrow_{\Delta\to 0} 2. \tag{86}$$

Finally

$$h(u,v) \leq_{t\to 0} -8\log(t)t^2 + o(-\log(t)t^2). \tag{87}$$

Hence using these bounds over regime A, B, and C, and remarking that it is reached for regime A on $t^2/(1+t^2)$, we obtain

$$\omega_\varphi^2(g,t) =_{t\to 0} \frac{-8\log(2)t^2\log(t)}{1+t^2} + o(-t^2\log(t)). \tag{88}$$

By applying on Equation (75) the upper bounds we derived and taking $t = n^{-1/2}$, we can conclude

$$|b(\hat{Q})| \leq \sum_{k=1}^{S} |\mathbb{E}[B_n(f)(p_k)] - f(p_k)| \tag{89}$$
$$\leq S\frac{5}{2}\omega_\varphi^2(f, n^{-1/2}) \tag{90}$$
$$S \leq_{n\to\infty} 10\log(2)\frac{\log(n)}{n} + o(\frac{\log(n)}{n}) \tag{91}$$
$$=_{n\to\infty} \mathcal{O}\Big(\frac{S\log(n)}{n}\Big) \tag{92}$$

□

C Using partitions of protected attributes

C.1 Consistency of $u_I^{(q^*)}$

The main idea of this proof, is that when the number of samples n increases, the probability that $\min_{\mathcal{A},\mathcal{Y}} N_{a,y} < \tau$ goes to 0 as $n \to \infty$. And when $\min_{\mathcal{A},\mathcal{Y}} N_{a,y} \geq \tau$ then by definition $u_I^{(q^*)} = \hat{u}^*$ which is consistent.

We define for a in $\mathcal{A}$ the modified empirical estimator $\hat{p}_A(A = a)$, with $\hat{p}_A(A = a) = N_a/n$ if $N_a > 0$ and 1 otherwise. Using Chebyshev's inequality and because $N_{a,y} \sim \mathcal{B}(p_{A,\hat{Y}}(a, y), n)$ we have for $\epsilon > 0$:

$$\Pr(|\hat{p}_A(a) - p_A(a)| \geq \epsilon) = \Pr(|\hat{p}_A(a) - p_A(a)| \geq \epsilon, N_a = 0) + \Pr(|\hat{p}_A(a) - p_A(a)| \geq \epsilon, N_a > 0) \tag{93}$$

$$\leq \Pr(N_a = 0) + \Pr(|\frac{N_a}{n} - p_A(a)| \geq \epsilon, N_a > 0) \tag{94}$$

$$\leq (1 - p_A(a))^n + \Pr(|N_a - np_A(a)| \geq n\epsilon) \tag{95}$$

$$\leq (1 - p_A(a))^n + \frac{p_A(a)(1 - p_A(a))}{n\epsilon^2} \tag{96}$$

$$\longrightarrow_{n\to\infty} 0, \tag{97}$$

which means that $\hat{p}_A$ is a consistent estimator of p_A. By Slutsky's Theorem and because $p_A(a) > 0$, $\hat{p}_{A,\hat{Y}}(a, y)/\hat{p}(a)$ is a consistent estimator of $p_{\hat{Y}|A}(y \mid a)$. Hence by the Continuous Mapping Theorem using the continuous functions $\max$, $\min$, $\log$ and $|\cdot|$, we have that $\hat{u^*}$ the estimator using the modified empirical probabilities, is a consistent estimator of u^*.

Now for the consistency of $u_I^{(q^*)}$, for $\tau > 0$ and $\epsilon > 0$, we have

$$\Pr(|u_I^{(q^*)} - u^*| > \epsilon) = 1 - \Pr(|u_I^{(q^*)} - u^*| \leq \epsilon) \tag{98}$$

$$\leq 1 - \Pr(|u_I^{(q^*)} - u^*| \leq \epsilon, \min_{\mathcal{A},\mathcal{Y}} N_{a,y} > \tau) \tag{99}$$

$$= 1 - \Pr(|\hat{u} - u^*| \leq \epsilon, \min_{\mathcal{A},\mathcal{Y}} N_{a,y} > \tau) \tag{100}$$

$$= \Pr([|\hat{u} - u^* >\leq \epsilon] \cup [\min_{\mathcal{A},\mathcal{Y}} N_{a,y} \leq \tau]) \tag{101}$$

$$\leq \Pr(|\hat{u} - u^*| > \epsilon) + \Pr(\min_{\mathcal{A},\mathcal{Y}} N_{a,y} \leq \tau) \tag{102}$$

$$\leq \Pr(|\hat{u} - u^*| > \epsilon) + \Pr(\exists (a, y) \in \mathcal{A} \times \mathcal{Y}, N_{a,y} \leq \tau) \tag{103}$$

$$\leq \Pr(|\hat{u} - u^*| > \epsilon) + \sum_{(a,y)\in\mathcal{A}\times\mathcal{Y}} \Pr(N_{a,y} \leq \tau). \tag{104}$$

The first term goes to 0 by the consistency of $\hat{u}$. We will show that the second term also goes to zero as $n \to \infty$. For $\tau > 0$ we apply Hoeffding's inequality on $N_{a,y} \sim \mathcal{B}(p_{A,\hat{Y}}(a, y), n)$ to obtain the following concentration inequality:

$$\Pr(N_{a,y} \leq \tau) \leq \exp(-2n(p_{A,\hat{Y}}(a, y) - \frac{\tau}{n})) \longrightarrow_{n\to\infty} 0. \tag{105}$$

Therefore because $|\mathcal{A}||\mathcal{Y}|$ is finite we have the consistency of $u_I^{(q^*)}$.

C.2 Some intuition on the impact of grouping protected attributes

Let q a partition and ρ a partition coarser than q, which means that every element of q is a subset of some element of ρ. We want to somewhat relate the approximations and inequalities obtained using ρ or q. Using the fact that ρ is coarser than q, we can define for any $r \in \rho$ the partition $q_r = \{t \in q \mid t \subset r\}$ of r. This is a partition because the t are disjoints, cover the whole set so r as well in particular, and any $t \in q$ can either be a subset of r or disjoint as ρ is coarser than q.

We redefine the random variable L for these partitions. We define $L^{(q)} = \log(p_A/\prod_{t\in q} p_{A_t}) \circ A$ and $L^{(q,r)} = \log(p_{A_r}/\prod_{t\in q_r} p_{A_t}) \circ A$.

Proposition C.1. *We have*

$$L^{(q)} = L^{(\rho)} + \sum_{r \in \rho \setminus q} L^{(q,r)}. \tag{106}$$

Therefore

$$\mathbb{E}[L^{(q)}] = \mathbb{E}[L^{(\rho)}] + \sum_{r \in \rho \setminus q} \mathbb{E}[L^{(q,r)}] \tag{107}$$

and

$$\begin{aligned} \mathrm{Var}(L^{(q)}) = \mathrm{Var}(L^{(\rho)}) + \sum_{r \in \rho \setminus q} \mathrm{Var}(L^{(q,r)}) + \\ 2 \sum_{r \in \rho \setminus q} \mathrm{Cov}(L^{(\rho)}, L^{(q,r)}) + \sum_{\substack{(r_1,r_2) \in (\rho \setminus q)^2 \\ r_1 \neq r_2}} \mathrm{Cov}(L^{(q,r_1)}, L^{(q,r_2)}). \end{aligned} \tag{108}$$

Proof. Using the partitions q_r we can group together the p_{A_t} terms:

$$\begin{aligned} \frac{p_A(a)}{\prod_{t \in q} p_{A_t}(a_t)} &= \frac{p_A(a)}{\prod_{t \in q} p_{A_t}(a_t)} \frac{\prod_{r \in \rho} p_{A_r}(a_r)}{\prod_{r \in \rho} p_{A_r}(a_r)} \\ &= \frac{p_A(a)}{\prod_{r \in \rho} p_{A_r}(a_r)} \prod_{r \in \rho \setminus q} \frac{p_{A_r}(a_r)}{\prod_{t \in q_r} p_{A_t}(a_t)}. \end{aligned}$$

Then by taking the $\log$, we obtain the proposition. □

Looking at (107), we have the interesting property that if ρ is a coarser partition than q, then $C(A^{(\rho)}) \leq C(A^{(q)})$. This corresponds to the intuition that taking coarser partition decreases some measure of independence, which is here the total correlation.

We define $\ell_t = \log(p_{A_t} / \prod_{k \in t} p_{A_k})$ and $L_t = \ell_t(A_t)$. By applying the previous proposition, and by remarking that any partition is coarser than the set of all singletons we obtain the following corollary.

Corollary C.2. *For any partition $q \in \mathcal{Q}$ we have*

$$L = L^{(q)} + \sum_{\substack{t \in q \\ |t| > 1}} L_t \tag{109}$$

Which is why when using a partition q to group together the protected attributes, we may be able to reduce the original variance of L and L_y, hence reduce s^* which is an increasing function of the variances. The decrease in s^* is not guaranteed when using a coarser partition because of the covariance terms, but empirically this is often the case.

D Additional Experiments and Plots

In this section we will present additional plots from the experiments conducted in Section 5.

The experiments were conducted on a machine with a i7 7700HQ CPU, and 8gb of ram. Running all the experiments took about 1 full day.

The main dataset used is a publicly available sample from the 1990 US census. The US census is legally mandated, hence every citizen has to give its information to the US government. No identifiable information is available, and the samples were randomly chosen from the original full dataset. Full information is available at the UCI archive link.

We reproduce here Figure 1 and Figure 2 on Figure 4 and Figure 5 adding the 1st and 10th decile but only using $\alpha = 1$ for u_B and $\tau = 10$ to make it readable.

We recall that we always take $\delta = 0.1$. We present on Figure 6 a comparison of the relative error rate between u_B, s^*, and $\inf_{g(\boldsymbol{\lambda})} \lambda^+ - \lambda^-$ where g is estimated through the empirical distribution, and the optimization problem is solved numerically. We see that while it is easier to estimate than u_B, it

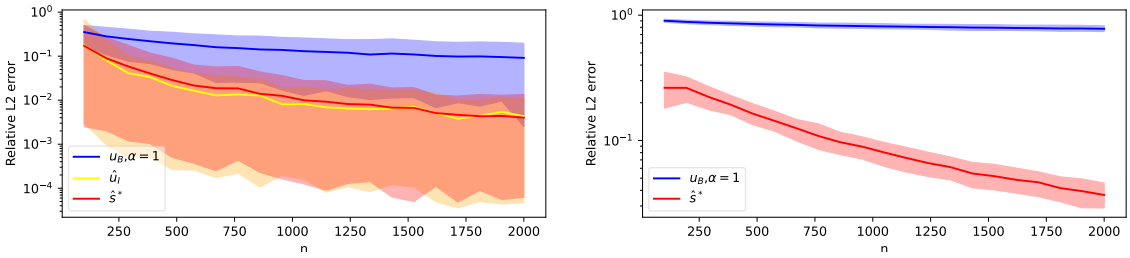

Figure 4: Average L_2^r convergence rate, on real data for the left one, and synthetic data for the right one. In all these graphs the intervals represent the 1st and 10th decile.

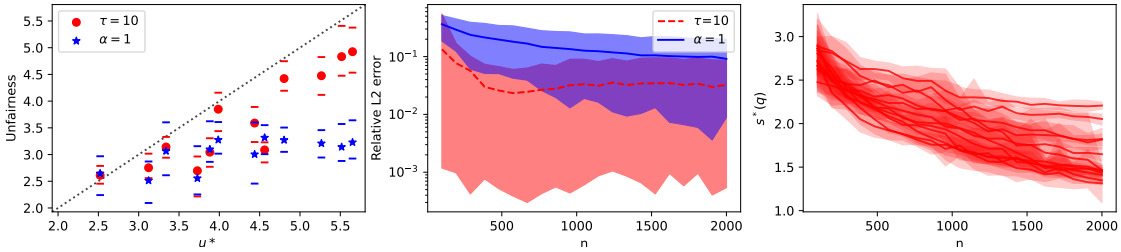

Figure 5: In the left-most plot each point with same shape and color correspond a different D_i, with the estimators values taken at $n = 2000$. The middle plot is the average L_2^r error rate on the real data. The rightmost plot is the average evolution of $s^*(q^*)$ for $\tau = 10$ as n increases. In all these graphs the intervals represent the 1st and 10th decile.

is harder than s^*. Note that numerically solving the minimization problem may lead to numerical errors for too low number of samples.

We also compare some of the bounds presented throughout this paper. Here we do not care about their estimation, but only their asymptotic value. In addition, we want to evaluate the impact of using partitions on these bounds. In order to do so we take a sample of size $n = 2000$ of each of our D_i, and compute q^*. Then we use the full dataset to compute $s^*(q^*)$ for $\tau = 10$, $\gamma(q^*)$ and $\inf_{g(\boldsymbol{\lambda})} \lambda^+ - \lambda^-$. We also compute the exact unfairness quantile $\epsilon^*(\delta)$. We obtain Figure 7. We see that using partitions seem to always yield tighter bounds, and that most of the time $\epsilon_1 \geq \epsilon_2 \geq \epsilon_3$. Even the improved bounds are still far from ϵ^* (the optimal bound in probability), but it shows that these bounds can be improved. We conjecture that if we want to find reliable information on u^* when d becomes very large, these bounds can be useful in practice. Conversely, these bounds and approximations should not be used if sufficient information is available to directly use u_B (for instance at least 1 sample by protected group).

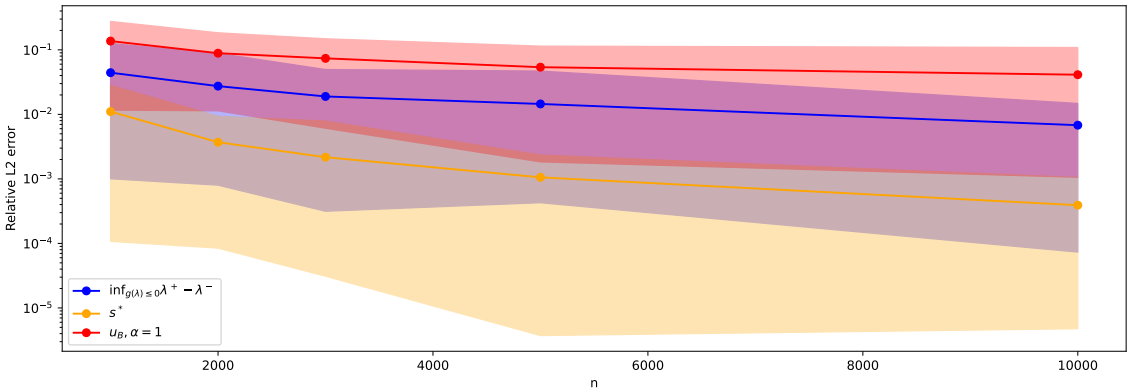

Figure 6: Average L_2^r convergence rate, on real data. The intervals represent the 1st and 10th decile.

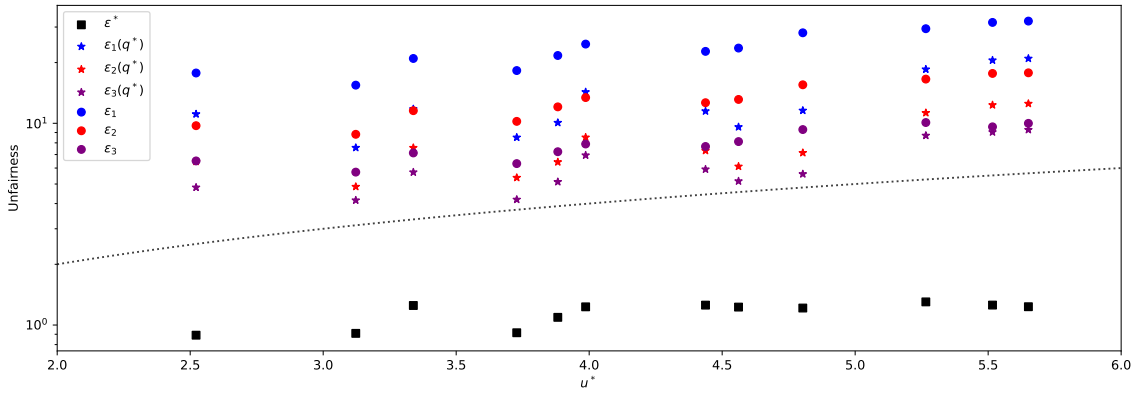

Figure 7: Exact computation of the various bounds for each of the 12 D_i selected.