Differential Fairness

James R. Foulds, Rashidul Islam, Kamrun Naher Keya, Shimei Pan
University of Maryland, Baltimore County, USA
{jfoulds, islam.rashidul, kkeya1, shimei}@umbc.edu

Abstract

We propose differential fairness, a multi-attribute definition of fairness in machine learning which is informed by the framework of intersectionality, a critical lens arising from the humanities literature, leveraging connections between differential privacy and legal notions of fairness. We show that our criterion behaves sensibly for any subset of the set of protected attributes, and we prove economic, privacy, and generalization guarantees. We provide a learning algorithm which respects our differential fairness criterion. Experiments on the COMPAS criminal recidivism dataset and census data demonstrate the utility of our methods.

1 Introduction and Motivation

The increasing impact of artificial intelligence and machine learning technologies on many facets of life, from commonplace movie recommendations to consequential criminal justice sentencing decisions, has prompted concerns that these systems may behave in an unfair or discriminatory manner [2, 24, 25]. A number of studies have subsequently demonstrated that bias and fairness issues in AI are both harmful and pervasive [1, 5, 4]. The AI community has responded by developing a broad array of mathematical formulations of fairness and learning algorithms which aim to satisfy them [11, 15, 3, 29]. Fairness, however, is not a purely technical construct, having social, political, philosophical and legal facets [6]. The necessity has now become clear for interdisciplinary analyses of fairness in AI and its relationship to society, to civil rights, and to the social goals which are to be achieved by mathematical fairness definitions, which have not always been made explicit [23]. In this work, we address the specific challenges of fairness in AI that are motivated by **intersectionality**, an analytical lens from the third-wave feminist movement which emphasizes that civil rights and feminism should be considered simultaneously rather than separately [10]. We propose an **intersectional AI fairness criterion** and perform a theoretical analysis of its properties relating to diverse fields including the **humanities**, **law**, **privacy**, **economics**, and **statistical machine learning**.

The principle of *intersectionality* emphasizes that systems of oppression built into society lead to *systematic disadvantages along intersecting dimensions*, which include not only gender, but also race, nationality, sexual orientation, disability status, and socioeconomic class [9, 8, 10, 16, 21, 27]. These systems are interlocking in their effects on individuals at *each intersection of the affected dimensions*. Intersectionality thus implies the use of multiple *protected attributes*, and has further implications. Many AI fairness definitions aim (implicitly or otherwise) to uphold the principle of *infra-marginality*, which states that differences between protected groups in the distributions of "merit" or "risk" (e.g. the probability of carrying contraband at a policy stop) should be taken into account when determining whether bias has occurred [26]. In short, the *infra-marginality* principle makes the implicit assumption that society is a fair, level playing field, and thus differences in "merit" or "risk" between groups in data and predictive algorithms are often to be considered legitimate. In contrast, *intersectionality* theory posits that these **distributions of merit and risk are often influenced by unfair societal processes**. In ideal intersectional fairness, since ability to succeed is affected by unfair processes, it is desired that this unfairness is corrected and individuals achieve their true

NeurIPS 2019 Workshop on Machine Learning with Guarantees, Vancouver, Canada.

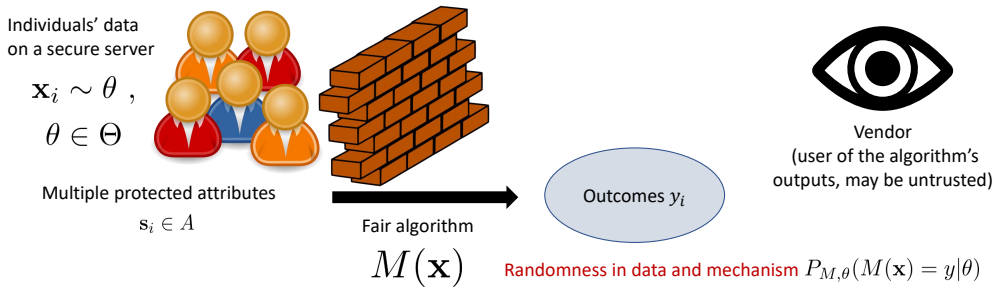


Figure 1: Diagram of the setting for the proposed differential fairness criterion.

potential [28]. Assuming individuals' unbiased potential does not substantially differ across protected groups, this implies that parity between groups, and intersectional subgroups, is typically desirable.[1]

In the machine learning literature, the previous AI fairness definition most relevant to intersectionality is *statistical parity subgroup fairness (SF)* [18]. We adapt the notation of [19] to all definitions in this paper. Suppose $M(\mathbf{x})$ is a (possibly randomized) mechanism which takes an instance $\mathbf{x} \in \chi$ and produces an outcome y for the corresponding individual, $S_1, \ldots, S_p$ are discrete-valued protected attributes, $A = S_1 \times S_2 \times \ldots \times S_p$, and θ is the distribution which generates $\mathbf{x}$. For example, the mechanism $M(\mathbf{x})$ could be a deep learning model for a lending decision, A could be the applicant's possible gender and race, and θ the joint distribution of credit scores and protected attributes. The protected attributes are included in the attribute vector $\mathbf{x}$, although $M(\mathbf{x})$ is free to disregard them (e.g. if this is disallowed). The setting is illustrated in Figure 1.

Definition 1.1. *(Statistical Parity Subgroup Fairness [18]) Let $\mathcal{G}$ be a collection of protected group indicators $g : A \to \{0, 1\}$, where $g(\mathbf{s}) = 1$ designates that an individual with protected attributes $\mathbf{s}$ is in group g. Assume that the classification mechanism $M(\mathbf{x})$ is binary, i.e. $y \in \{0, 1\}$.*

Then $M(\mathbf{x})$ is γ-statistical parity subgroup fair *with respect to θ and $\mathcal{G}$ if for every $g \in \mathcal{G}$,*

$$|P_{M,\theta}(M(\mathbf{x}) = 1) - P_{M,\theta}(M(\mathbf{x}) = 1|g(\mathbf{s}) = 1)| \times P_\theta(g(\mathbf{s}) = 1) \leq \gamma \,. \tag{1}$$

From an intersectional perspective, one concern with SF is that it does not protect minority groups, often marginalized by society, and whose protection intersectionality emphasizes. The term $P_\theta(g(\mathbf{s}) = 1)$ weights the "per-group (un)fairness" for each group g, i.e. Equation 1 applied to g alone, by its proportion of the population, thereby *downweighting the consideration of minorities*.

2 Differential Fairness (DF) Measure

We propose an alternative fairness criterion which is more concordant with intersectionality, including its treatment of minorities and its other provable theoretical properties. We first motivate our criterion from a legal perspective. Consider the 80% rule, established in the Code of Federal Regulations [14] as a guideline for establishing disparate impact in violation of anti-discrimination laws such as Title VII of the Civil Rights Act of 1964. The 80% rule states that there is legal evidence of adverse impact if the ratio of probabilities of a particular favorable outcome, taken between a disadvantaged and an advantaged group, is less than 0.8:

$$P(M(\mathbf{x}) = 1|\text{group A})/P(M(\mathbf{x}) = 1|\text{group B}) < 0.8 \,. \tag{2}$$

Our proposed criterion, which we call **differential fairness (DF)**, extends the 80% rule to protect multi-dimensional intersectional categories, with respect to multiple output values. We similarly restrict ratios of outcome probabilities between groups, but instead of using a predetermined fairness threshold at 80%, we measure fairness on a sliding scale that can be interpreted similarly to that of *differential privacy*, a definition of privacy for data-driven algorithms [12]. Differential fairness measures the **fairness cost** of mechanism $M(\mathbf{x})$ with a parameter ϵ.

[1] Disparity could still be desirable if there are legitimate confounders which depend on protected groups, e.g. choice of department that individuals apply to in college admissions. We address this in Appendix E.1.

Definition 2.1. *A mechanism $M(\mathbf{x})$ is* ϵ-differentially fair (DF) *with respect to (A, Θ) if for all $\theta \in \Theta$ with $\mathbf{x} \sim \theta$, and $y \in Range(M)$,*

$$e^{-\epsilon} \leq \frac{P_{M,\theta}(M(\mathbf{x}) = y|\mathbf{s}_i, \theta)}{P_{M,\theta}(M(\mathbf{x}) = y|\mathbf{s}_j, \theta)} \leq e^{\epsilon}\,, \tag{3}$$

for all $(\mathbf{s}_i, \mathbf{s}_j) \in A \times A$ where $P(\mathbf{s}_i|\theta) > 0$, $P(\mathbf{s}_j|\theta) > 0$.

In Equation 3, $\mathbf{s}_i$, $\mathbf{s}_j \in A$ are tuples of *all* protected attribute values, e.g. gender, race, and nationality, and Θ is a set of distributions θ which could plausibly generate each instance $\mathbf{x}$.[2] For example, Θ could be the set of Gaussian distributions over credit scores per value of the protected attributes, with mean and standard deviation in a certain range.

This is an intuitive **intersectional definition of fairness**: *regardless of the combination of protected attributes, the probabilities of the outcomes will be similar*, as measured by the ratios versus other possible values of those variables, for small values of ϵ. For example, the probability of being given a loan would be similar regardless of a protected group's intersecting combination of gender, race, and nationality, marginalizing over the remaining attributes in $\mathbf{x}$. If the probabilities are always equal, then $\epsilon = 0$, otherwise $\epsilon > 0$. We have arrived at our criterion based on the 80% rule, but it can also be derived as a special case of *pufferfish* [19], a generalization of differential privacy [13] which uses a variation of Equation 3 to hide the values of an arbitrary set of secrets. If $P_{M,\theta}$ is unknown, it can be estimated using the empirical distribution, or via a probabilistic model of the data.

We can adapt DF to measure fairness in data, i.e. outcomes assigned by a black-box algorithm or social process, by using (a model of) the data's generative process as the mechanism.

Definition 2.2. *A labeled dataset $D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$ is* ϵ-differentially fair (DF) *in A with respect to model $P_{Model}(\mathbf{x}, y)$ if mechanism $M(\mathbf{x}) = y \sim P_{Model}(y|\mathbf{x})$ is* ϵ-differentially fair *with respect to $(A, \{P_{Model}(\mathbf{x})\})$, for P_{Model} trained on the dataset.*

In the Appendix, we consider extensions of DF to handle confounder variables, and to measure the amplification of bias due to an algorithm. We also provide worked examples and estimation strategies.

3 Properties of Differential Fairness (Proofs Given in Appendix)

Intersectionality: Differential fairness explicitly encodes protection of intersectional groups. For DF, we prove that this automatically implies fairness for *each of the protected attributes individually*, and indeed, *any subset* of the protected attributes. In other words, by ensuring fairness at the intersection of gender, race, and nationality under our criterion, we also ensure the same degree of fairness between genders overall, and between gender/nationality pairs overall, and so on.

Theorem 3.1. ***(Intersectionality Property)*** *Let M be an* ϵ-differentially fair *mechanism in (A, Θ), $A = S_1 \times S_2 \times \ldots \times S_p$, and let $D = S_a \times \ldots \times S_k$ be the Cartesian product of a nonempty proper subset of the protected attributes included in A. Then M is* ϵ-differentially fair *in (D, Θ).*

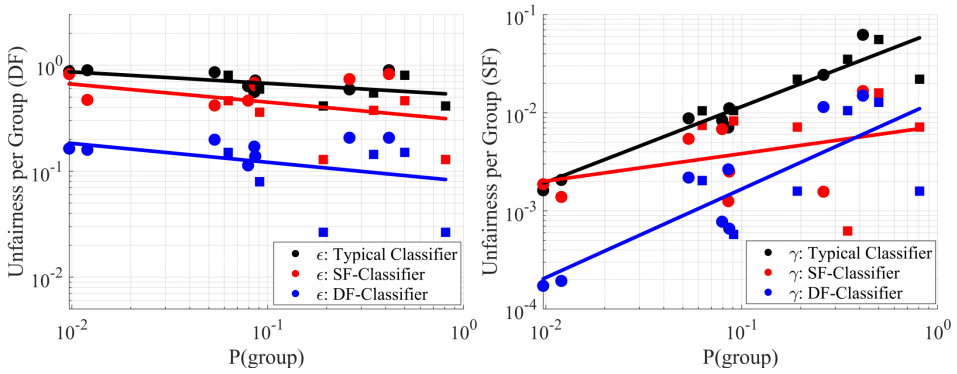
Privacy: The differential fairness definition, and the resulting level of fairness obtained at any particular measured fairness parameter ϵ, can be interpreted by viewing the definition through the lens of privacy. Differential fairness ensures that given the outcome, *an untrusted vendor/adversary can learn very little about the protected attributes of the individual*, relative to their prior beliefs, assuming their prior beliefs are in Θ:

$$e^{-\epsilon}\frac{P(\mathbf{s}_i|\theta)}{P(\mathbf{s}_j|\theta)} \leq \frac{P(\mathbf{s}_i|M(\mathbf{x}) = y, \theta)}{P(\mathbf{s}_j|M(\mathbf{x}) = y, \theta)} \leq e^{\epsilon}\frac{P(\mathbf{s}_i|\theta)}{P(\mathbf{s}_j|\theta)}\,. \tag{4}$$

The privacy guarantee only holds if $\theta \in \Theta$, which may not always be the case. Regardless, the value of ϵ may typically be interpreted as a privacy guarantee against a "reasonable adversary."

Utility: An ϵ-differentially fair mechanism admits a disparity in expected utility of as much as a factor of $\exp(\epsilon) \approx 1 + \epsilon$ (for small values of ϵ) between pairs of protected groups with $\mathbf{s}_i \in A$, $\mathbf{s}_j \in A$, for any utility function. The proof follows the case of differential privacy [13], see Appendix.

[2] The possibility of multiple $\theta \in \Theta$ is valuable from a privacy perspective, where Θ is the set of *possible beliefs* that an adversary may have about the data, and is motivated by the work of [19]. Continuous protected attributes are also possible, in which case sums are replaced by integrals in our proofs.



(a) Improvement in DF measures (b) Improvement in SF measures

Figure 2: Per-group measurements of (a) ϵ-DF and (b) γ-SF of the classifiers vs group size (probability), COMPAS dataset, calculated using Equations 1 and 3 with the group held fixed. Circles: intersectional subgroups. Squares: top-level groups. The methods improve fairness, both per group and overall, but SF-Classifier is seen to ignore minority groups in the overall γ-SF measurement, calculated as a worst-case over all groups.

Generalization: To ensure that an algorithm is truly fair, the fairness properties obtained on a dataset must extend to the underlying population. We prove a generalization guarantee for estimating DF although it is weaker than for subgroup fairness [18] – the price of protecting minority subgroups:

Theorem 3.2. ***(Generalization Property)*** *Fix a class of functions $\mathcal{H}$, which without loss of generality aim to discriminate the outcome $y = 1$ from any other value, denoted here as $y = 0$. For any conditional distribution $P(y, \mathbf{x}|\mathbf{s})$ given a group $\mathbf{s}$, let $S \sim P^m$ be a dataset consisting of m examples $(\mathbf{x}_i, y_i)$ sampled i.i.d. from $P(y, \mathbf{x}|\mathbf{s})$. Then for any $0 < \delta < 1$, with probability $1 - \delta$, for every $h \in \mathcal{H}$, we have:*

$$|P(y = 1|\mathbf{s}, h) - P_S(y = 1|\mathbf{s}, h)| \leq \tilde{O}\left(\sqrt{\frac{VCDIM(\mathcal{H}) \log m + \log(1/\delta)}{m}}\right). \tag{5}$$

4 Experiments and Conclusion

We trained deep neural network classifiers with a penalty term for differential fairness (*DF-Classifier*) via adaptive gradient descent (Adam) (see Appendix). The same approach was used to train a subgroup fair classifier (*SF-Classifier*). We performed experiments on the COMPAS dataset regarding a system that is used to predict criminal recidivism [1] (protected attributes: *race* and *gender*). Further experiments were performed on the Adult 1994 U.S. census income data from the UCI repository [20] (protected attributes: *race*, *gender*, USA vs non-USA *nationality*), see the Appendix.[3]

An important goal of this work was to consider the impact of the fairness methods on minority groups. In Figure 2, we report the "per-group unfairness," defined as Equations 1 and 3 with one group held fixed, versus the group's probability (i.e. size) on the COMPAS dataset. Both methods improve their corresponding per-group unfairness measures over the typical classifier. On the other hand, the γ-SF metric only assigns high per-group unfairness values to large groups in its measurement, so **minority groups are not able to influence the overall γ-SF unfairness**. This was not the case **for ϵ-DF metric, where groups of various sizes had similarly high per-group ϵ values**. Furthermore, the **DF-Classifier improved the per-group fairness under both metrics for groups of all sizes, while the SF-classifier did not improve the per-group γ-SF for small groups**. Further experiments, given in the Appendix, show that DF-Classifier and SF-Classifier behave similarly in terms of accuracy, and that they can be tuned to improve fairness with little loss in accuracy. Our overall conclusion is that the *DF-Classifier is able to achieve intersectionally fair classification with minor loss in performance, while providing greater protection to minority groups than when enforcing subgroup fairness*.

[3]Predicted income, used for consequential decisions like housing approval, may result in *digital redlining* [2].

Acknowledgments

This work was performed under the following financial assistance award: 60NANB18D227 from U.S. Department of Commerce, National Institute of Standards and Technology.

This material is based upon work supported by the National Science Foundation under Grant No. IIS 1850023. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

We thank Rosie Kar for valuable advice and feedback regarding intersectional feminism.

References

[1] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica, May*, 23, 2016.

[2] S. Barocas and A.D. Selbst. Big data's disparate impact. *Cal. L. Rev.*, 104:671, 2016.

[3] R. Berk, H. Heidari, S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, S. Neel, and A. Roth. A convex framework for fair regression. *FAT/ML Workshop*, 2017.

[4] T. Bolukbasi, K.-W. Chang, J.Y. Zou, V. Saligrama, and A.T. Kalai. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in NeurIPS*, 2016.

[5] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *FAT**, pages 77–91, 2018.

[6] A. Campolo, M. Sanfilippo, M. Whittaker, A. Selbst K. Crawford, and S. Barocas. *AI Now 2017 Symposium Report*. AI Now, 2017.

[7] C.R. Charig, D.R. Webb, S.R. Payne, and J.E. Wickham. Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy. *British Medical Journal (BMJ) (Clin Res Ed)*, 292(6524):879–882, 1986.

[8] P.H. Collins. *Black feminist thought: Knowledge, consciousness, and the politics of empowerment (2nd ed.)*. Routledge, 2002 [1990].

[9] Combahee River Collective. A black feminist statement. In Z. Eisenstein, editor, *Capitalist Patriarchy and the Case for Socialist Feminism*. Monthly Review Press, New York, 1978.

[10] K. Crenshaw. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *U. Chi. Legal F.*, pages 139–167, 1989.

[11] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of ITCS*, pages 214–226. ACM, 2012.

[12] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Th. of Cryptography*, pages 265–284, 2006.

[13] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Theoretical Computer Science*, 9(3-4):211–407, 2013.

[14] Equal Employment Opportunity Commission. Guidelines on employee selection procedures. *C.F.R.*, 29.1607, 1978.

[15] M. Hardt, E. Price, N. Srebro, et al. Equality of opportunity in supervised learning. In *Advances in NeurIPS*, pages 3315–3323, 2016.

[16] b. hooks. *Ain't I a Woman: Black Women and Feminism*. South End Press, 1981.

[17] S.A. Julious and M.A. Mullee. Confounding and simpson's paradox. *British Medical Journal (BMJ)*, 309(6967):1480–1481, 1994.

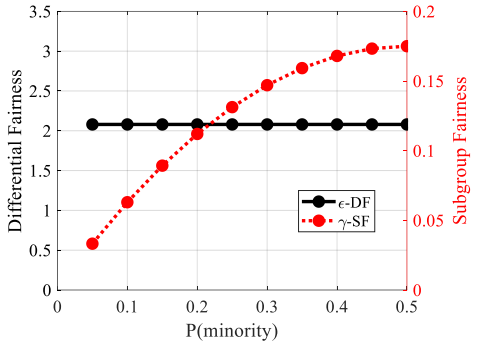


Figure 3: Toy example: probability of the "positive" class is 0.8 for a majority group, 0.1 for a minority group, varying $P(\text{minority})$.

[18] M. Kearns, S. Neel, A. Roth, and Z.S. Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In J. Dy and A. Krause, editors, *Proc. of ICML, PMLR 80*, pages 2569–2577, 2018.

[19] D. Kifer and A. Machanavajjhala. Pufferfish: A framework for mathematical privacy definitions. *TODS*, 39(1):3, 2014.

[20] R. Kohavi. Scaling up the accuracy of naive-Bayes classifiers: a decision-tree hybrid. In *Proceedings of SIGKDD*, pages 202–207, 1996.

[21] A. Lorde. Age, race, class, and sex: Women redefining difference. In *Sister Outsider*, pages 114–124. Ten Speed Press, 1984.

[22] Max O Lorenz. Methods of measuring the concentration of wealth. *Publications of the American statistical association*, 9(70):209–219, 1905.

[23] S. Mitchell, E. Potash, and S. Barocas. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint arXiv:1811.07867*, 2018.

[24] C. Munoz, M. Smith, and D.J. Patil. *Big data: A report on algorithmic systems, opportunity, and civil rights*. Exec. Office of the President, 2016.

[25] S.U. Noble. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press, 2018.

[26] C. Simoiu, S. Corbett-Davies, S. Goel, et al. The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics*, 11(3):1193–1216, 2017.

[27] S. Truth. Ain't I a woman?, 1851. Speech delivered at Women's Rights Convention, Akron, Ohio.

[28] C. Verschelden. *Bandwidth Recovery: Helping Students Reclaim Cognitive Resources Lost to Poverty, Racism, and Social Marginalization*. Stylus, 2017.

[29] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of EMNLP*, 2017.

A A Toy Example Illustrating Minority Bias of Subgroup Fairness

In Figure 3, we show an example where varying the size of a minority group $P(\text{minority})$ drastically alters γ-subgroup fairness, which finds that **a rather extreme scenario is more acceptable when the minority group is small**. Our proposed criterion, ϵ-DF, is constant in $P(\text{minority})$.

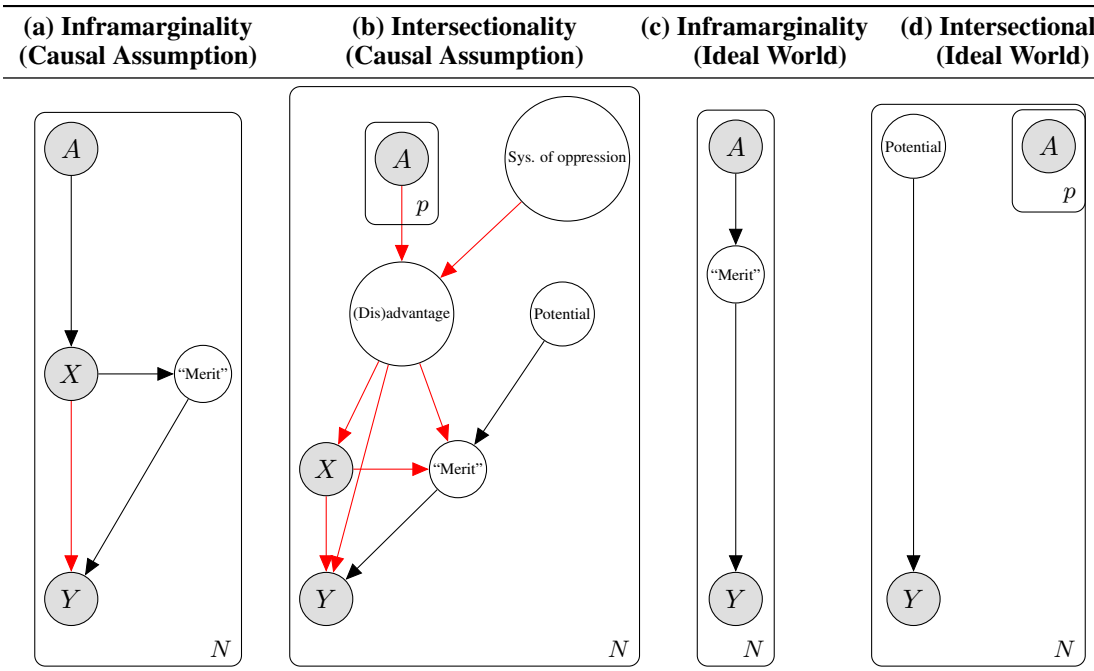


Figure 4: Implicit causal assumptions (a,b) and values-driven ideal world scenarios (c,d) for inframarginality and intersectionality notions of fairness. Here, A denotes protected attributes, X observed attributes, Y outcomes, N individuals, p number of protected attributes. Red arrows denote potentially unfair causal pathways, which are removed to obtain the ideal world scenarios (c,d). The above summarizes broad strands of research; individual works may differ.

B Causal Assumptions of Intersectionality and Inframarginality

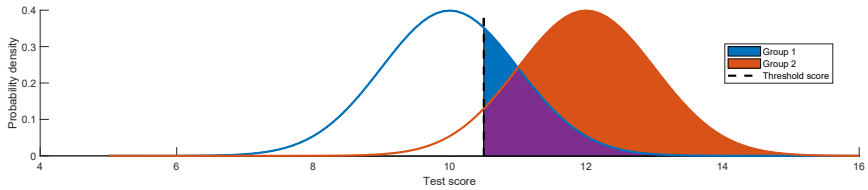
In Figure 4 we summarize the causal assumptions regarding society and data, and the idealized "perfect world" scenarios implicit in the two approaches to fairness. Inframarginality *(a)* emphasizes that the distribution over relevant attributes X varies across protected groups A, which leads to potential differences in so-called "merit" or "risk" between groups, typically presumed to correspond to latent ability and thus "deservedness" of outcomes Y [26]. Intersectionality *(b)* emphasizes that we must also account for systems of oppression which lead to (dis)advantage at the intersection of multiple protected groups, impacting all aspects of the system including the ability of individuals to succeed ("merit") to their potential, had they not been impacted by (dis)advantage [10].

In the ideal world that an algorithmic (or other) intervention aims to achieve, inframarginality-based fairness desires that individual "merit" is the sole determiner of outcomes *(c)* [26, 15], which can lead to disparity between groups [11]. In the ideal world that intersectionality-based fairness aims to achieve, unfair (dis)advantages between groups are removed so that individuals can achieve their potential. This implies parity of outcomes between groups, assuming that there are no legitimate confounder variables (e.g. the choice of department to apply to in a university admissions scenario).

C Estimating Differential Fairness from Data

Assuming discrete outcomes, $P_{Data}(y|\mathbf{s}) = \frac{N_{y,\mathbf{s}}}{N_{\mathbf{s}}}$, where $N_{y,\mathbf{s}}$ and $N_{\mathbf{s}}$ are empirical counts of their subscripted values in the dataset D. **Empirical differential fairness (EDF)** corresponds to verifying that for any y, $\mathbf{s}_i$, $\mathbf{s}_j$, we have

$$e^{-\epsilon} \leq \frac{N_{y,\mathbf{s}_i}}{N_{\mathbf{s}_i}} \frac{N_{\mathbf{s}_j}}{N_{y,\mathbf{s}_j}} \leq e^{\epsilon} \,, \tag{6}$$



Probability of Hiring Outcome Given Group			
		Group	
		1	2
Outcome	yes	0.3085	0.9332
	no	0.6915	0.0668

Log Ratios of Probabilities			
y	$\mathbf{s}_i$	$\mathbf{s}_j$	$\log \frac{P_{M,\theta}(M(\mathbf{x})=y\|\mathbf{s}_i,\theta)}{P_{M,\theta}(M(\mathbf{x})=y\|\mathbf{s}_j,\theta)}$
no	1	2	2.337
	2	1	-2.337
yes	1	2	-1.107
	2	1	1.107

Figure 5: Worked example of differential fairness. The calculations above show that $\epsilon = 2.337$.

Alternatively, if we estimate ϵ-DF via the posterior predictive distribution of a Dirichlet-multinomial model, the criterion for any y, $\mathbf{s}_i$, $\mathbf{s}_j$ becomes

$$e^{-\epsilon} \leq \frac{N_{y,\mathbf{s}_i} + \alpha}{N_{\mathbf{s}_i} + |\mathcal{Y}|\alpha} \frac{N_{\mathbf{s}_j} + |\mathcal{Y}|\alpha}{N_{y,\mathbf{s}_j} + \alpha} \leq e^{\epsilon} , \tag{7}$$

where scalar α is each entry of the parameter of a symmetric Dirichlet prior with concentration parameter $|\mathcal{Y}|\alpha$, $\mathcal{Y} = \text{Range}(M)$. We refer to this as **smoothed EDF**.

Note that EDF and smoothed EDF methods can sometimes be unstable in extreme cases when nearly all instances are assigned to the same class. To address this issue, instead of using empirical hard counts per group $N_{y,s}$, we can also use *soft counts* for (smoothed) EDF, based on a probabilistic classifier's predicted $P(y|\mathbf{x})$, as follows:

$$e^{-\epsilon} \leq \frac{\sum_{\mathbf{x}\in D:A=\mathbf{s}_i} P(y|\mathbf{x}) + \alpha}{N_{\mathbf{s}_i} + |\mathcal{Y}|\alpha} \frac{N_{\mathbf{s}_j} + |\mathcal{Y}|\alpha}{\sum_{\mathbf{x}\in D:A=\mathbf{s}_j} P(y|\mathbf{x}) + \alpha} \leq e^{\epsilon} . \tag{8}$$

D Illustrative Worked Examples

A simple worked example of differential fairness is given in Figure 5. In the example, given an applicant's score x on a standardized test, the mechanism $M(x) = x \geq t$ approves the hiring of a job applicant if their test score $x \geq t$, with $t = 10.5$. The scores are distributed according to θ, which corresponds to the following process. The applicant's protected group is 1 or 2 with probability 0.5. Test scores for group 1 are normally distributed $N(x; \mu_1 = 10, \sigma = 1)$, and for group 2 are distributed $N(x; \mu_2 = 12, \sigma = 1)$. In the figure, the group-conditional densities are plotted on the top, along with the threshold for the hiring outcome being *yes* (i.e. $M(x) = 1$). Shaded areas indicate the probability of a *yes* hiring decision for each group (overlap in purple). On the bottom, the calculations show that $M(x)$ is ϵ-differentially fair for $\epsilon = 2.337$. This means that the probability ratios are bounded within the range $(e^{-\epsilon}, e^{\epsilon}) = (0.0966, 10.35)$, i.e. one group has around 10 times the probability of some particular hiring outcome than the other ($y = no$). Under the presumption that the two groups are roughly equally capable of performing the job overall, this is clearly unsatisfactory in terms of fairness.

The *intersectional* setting, in which there are multiple protected variables, is specifically addressed by differential fairness, by considering the probabilities of outcomes for each intersection of the set of protected variables. We illustrate this setting with an example on admissions of prospective students to a particular University X. In the scenario, the protected attributes are gender and race, and the mechanism is the admissions process, with a binary outcome. Our data, shown in Table 1, is adapted from a real-world scenario involving treatments for kidney stones, often used to demonstrate Simpson's paradox [7, 17]. Here, the "paradox" is that for race 1, individuals of gender A are more likely to be admitted than those of gender B, and for race 2, those of gender A are also more likely to be admitted than those of gender B, yet counter-intuitively, gender B is more likely to be admitted overall.

Probability of Being Admitted to University X				
		Gender		
		A	B	Overall
Race	1	$\frac{81}{87}$ (0.931)	$\frac{234}{270}$ (0.867)	$\frac{315}{357}$ (0.882)
	2	$\frac{192}{263}$ (0.730)	$\frac{55}{80}$ (0.688)	$\frac{247}{343}$ (0.720)
	Overall	$\frac{273}{350}$ (0.780)	$\frac{289}{350}$ (0.826)	

Table 1: Intersectional example: Simpson's paradox.

Since the admissions process is a black box, we model it using Equation 6, empirical differential fairness (EDF). By calculating the log probability ratios of $(Gender, Race)$ pairs from Table 1, as well as for the pairs of probabilities for the declined admission outcome $(1 - P(\text{admit}))$, and plugging them into Equation 6, we see that the mechanism is $\epsilon = 1.511$-DF with $A = Gender \times Race$. By calculating ϵ using the admission probabilities in the *Overall* row $(Gender)$ and the *Overall* column $(Race)$, we find that $\epsilon = 0.2329$ for $A = Gender$, and $\epsilon = 0.8667$ for $A = Race$. We will prove in Theorem 3.1 that ϵ with $A = Gender \times Race$ is an upper bound on ϵ-DF for $A = Gender$ and for $A = Race$. Thus, even with a "Simpson's reversal" differential (un)fairness will not increase after summing out a protected attribute.

E Extensions to the Differential Fairness Definition

E.1 Dealing with Confounder Variables

As we have seen, differential fairness can be used to measure the inequity between the outcome probabilities for the protected groups and their intersections at different levels of measurement granularity, although it does not determine whether the inequities were due to systemic factors and/or discrimination. In the case study above, a confounding variable which could explain the Simpson's reversal is the decision of the prospective student on the department to apply to. The ϵ-DF criterion is appropriate when the differences are believed to be due to systems of oppression, as posited by intersectionality theory, and confounder variables are not present. With confounders, parity in outcomes between intersectional protected groups, which ϵ-DF rewards, may no longer be desirable (see Figure 6). For example, a confounder variable which could partly explain gender or racial disparities in a university's admissions is the choice of department to apply to, some of which are more selective. We propose an alternative fairness definition for when known confounders are present.

Definition E.1. *Let $\theta \in \Theta$ be distributions over $(\mathbf{x}, c)$, where $c \in C$ are confounder variables. A mechanism $M(\mathbf{x})$ is* ϵ-differentially fair with confounders (DFC) *with respect to (A, Θ, C), if for all $c \in C$, $M(\mathbf{x})$ is ϵ-DF with respect to $(A, \Theta_{|c})$, where $\Theta_{|c} = \{P(\mathbf{x}|\theta, c)|\theta \in \Theta\}$.*

In the university admissions case, Definition E.1 penalizes disparity in admissions at the department level, and the most unfair department determines the overall unfairness ϵ-DFC.

Theorem E.1. ***(Confounders Property)*** *Let M be an ϵ-DFC mechanism in (A, Θ, C), Then M is* ϵ-differentially fair *in (A, Θ).*

From Theorem E.1, if we protect differential fairness per department, we obtain differential fairness and its corresponding theoretical economic and privacy guarantees in the University's overall admissions, bounded by the ϵ of the most unfair department, *even in the case of a Simpson's reversal*.

E.2 DF Bias Amplification Measure

Similarly to differential privacy, differences $\epsilon_2 - \epsilon_1$ between two mechanisms $M_2(\mathbf{x})$ and $M_1(\mathbf{x})$ are meaningful (for fixed A and Θ, and for tightly computed minimum values of ϵ), and measure the additional "fairness cost" of using one mechanism instead of the other. When ϵ_1 is the differential fairness of a labeled dataset and ϵ_2 is the differential fairness of a classifier measured on the same dataset, $\epsilon_2 - \epsilon_1$ is a measure of the extent to which the classifier increases the unfairness over the original data, a phenomenon that [29] refer to as *bias amplification*.

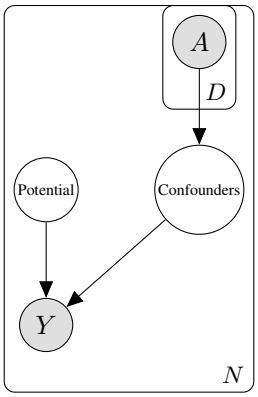


Figure 6: Ideal-world intersectional fairness but with counfounder variables present. Disparity in overall outcomes between protected groups may occur.

Definition E.2. *A mechanism $M(\mathbf{x})$ satisfies $(\epsilon_2 - \epsilon_1)$-*DF bias amplification *with respect to $(A, \Theta, D, \mathcal{M})$ if it is ϵ_2-DF and D is a labeled dataset which is ϵ_1-DF with respect to model $\mathcal{M}$.*

Politically speaking, ϵ-DF is a relatively progressive notion of fairness which we have motivated based on intersectionality (*disparities in societal outcomes are largely due to systems of oppression*), and which is reminiscent of demographic parity [11]. On the other hand, $(\epsilon_2 - \epsilon_1)$-$DF$ **bias amplification** is a more **politically conservative fairness metric** which does not seek to correct unfairness in the original dataset, in line with the principle of **infra-marginality** (*a system is biased only if disparities in its behavior are worse than those in society*) [26]. Informally, ϵ_2-DF and $(\epsilon_2 - \epsilon_1)$-$DF$ bias amplification represent "upper and lower bounds" on the unfairness of the system in the case where the relative effect of structural oppression on outcomes is unknown.

F Proofs

We start with a useful lemma.

Lemma F.1. *The ϵ-DF criterion can be rewritten as: for any $\theta \in \Theta$, $y \in$ Range(M),*

$$\begin{aligned} &\log \max_{\mathbf{s} \in A : P(\mathbf{s}|\theta) > 0} P_{M,\theta}(M(\mathbf{x}) = y | \mathbf{s}, \theta) \\ &- \log \min_{s \in A : P(\mathbf{s}|\theta) > 0} P_{M,\theta}(M(\mathbf{x}) = y | \mathbf{s}, \theta) \leq \epsilon \,. \end{aligned} \tag{9}$$

F.1 Proof of Lemma F.1

Proof. The definition of ϵ-differential fairness is, for any $\theta \in \Theta$, $y \in \text{Range}(M)$, $(\mathbf{s}_i, \mathbf{s}_j) \in A \times A$ where $P(\mathbf{s}_i|\theta) > 0$, $P(\mathbf{s}_j|\theta) > 0$,

$$e^{-\epsilon} \leq \frac{P_{M,\theta}(M(\mathbf{x}) = y | \mathbf{s}_i, \theta)}{P_{M,\theta}(M(\mathbf{x}) = y | \mathbf{s}_j, \theta)} \leq e^{\epsilon} \,. \tag{10}$$

Taking the log, we can rewrite this as:

$$\begin{aligned} -\epsilon \leq &\log P_{M,\theta}(M(\mathbf{x}) = y | \mathbf{s}_i, \theta) \\ &- \log P_{M,\theta}(M(\mathbf{x}) = y | \mathbf{s}_j, \theta) \leq \epsilon \,. \end{aligned} \tag{11}$$

The two inequalities can be simplified to:

$$|\log P_{M,\theta}(M(\mathbf{x}) = y | \mathbf{s}_i, \theta) - \log P_{M,\theta}(M(\mathbf{x}) = y | \mathbf{s}_j, \theta)| \leq \epsilon \,. \tag{12}$$

For any fixed θ and y, we can bound the left hand side by plugging in the worst case over $(\mathbf{s}_i, \mathbf{s}_j)$,

$$\begin{aligned} &|\log P_{M,\theta}(M(\mathbf{x}) = y | \mathbf{s}_i, \theta) - \log P_{M,\theta}(M(\mathbf{x}) = y | \mathbf{s}_j, \theta)| \\ &\leq \log \max_{\mathbf{s} : P(\mathbf{s}|\theta) > 0} P_{M,\theta}(M(\mathbf{x}) = y | \mathbf{s}, \theta) \\ &\quad - \log \min_{\mathbf{s} : P(\mathbf{s}|\theta) > 0} P_{M,\theta}(M(\mathbf{x}) = y | \mathbf{s}, \theta) \,. \end{aligned} \tag{13}$$

Plugging in this bound, which is achievable and hence is tight, the criterion is then equivalent to:

$$\log \max_{\mathbf{s}:P(\mathbf{s}|\theta)>0} P_{M,\theta}(M(\mathbf{x}) = y|\mathbf{s},\theta) \\ - \log \min_{\mathbf{s}:P(\mathbf{s}|\theta)>0} P_{M,\theta}(M(\mathbf{x}) = y|\mathbf{s},\theta) \leq \epsilon\,. \tag{14}$$

□

F.2 Proof of Theorem 3.1 (Intersectionality Property)

Proof. Define $E = S_1 \times \ldots \times S_{a-1} \times S_{a+1} \ldots \times S_{k-1} \times S_{k+1} \times \ldots \times S_p$, the Cartesian product of the protected attributes included in A but not in D. Then for any $\theta \in \Theta$, $y \in \text{Range}(M)$,

$$\begin{aligned}
&\log \max_{\mathbf{s}\in D:P(\mathbf{s}|\theta)>0} P_{M,\theta}(M(\mathbf{x}) = y|D = s,\theta) \\
=&\log \max_{\mathbf{s}\in D:P(\mathbf{s}|\theta)>0} \sum_{e\in E} P_{M,\theta}(M(\mathbf{x}) = y|E = e,\mathbf{s},\theta) P_\theta(E = e|\mathbf{s},\theta) \\
\leq&\log \max_{\mathbf{s}\in D:P(\mathbf{s}|\theta)>0} \sum_{e\in E} \max_{e'\in E:P_\theta(E=e'|\mathbf{s},\theta)>0} \\
&\qquad \big(P_{M,\theta}(M(\mathbf{x}) = y|E = e',\mathbf{s},\theta)\big) \times P_\theta(E = e|\mathbf{s},\theta) \\
=&\log \max_{\mathbf{s}\in D:P(\mathbf{s}|\theta)>0} \max_{e'\in E:P_\theta(E=e'|\mathbf{s},\theta)>0} P_{M,\theta}(M(\mathbf{x}) = y|E = e',\mathbf{s},\theta) \\
=&\log \max_{\mathbf{s}'\in A:P(\mathbf{s}'|\theta)>0} P_{M,\theta}(M(\mathbf{x}) = y|\mathbf{s}',\theta)
\end{aligned}$$

By a similar argument, $\log \min_{\mathbf{s}\in D:P(\mathbf{s}|\theta)>0} P_{M,\theta}(M(\mathbf{x}) = y|D = \mathbf{s},\theta) \geq \log \min_{\mathbf{s}'\in A:P(\mathbf{s}'|\theta)>0} P_{M,\theta}(M(\mathbf{x}) = y|\mathbf{s}',\theta)$. Applying Lemma F.1, we hence bound ϵ in (D,Θ) as

$$\begin{aligned}
&\log \max_{\mathbf{s}\in D:P(\mathbf{s}|\theta)>0} P_{M,\theta}(M(\mathbf{x}) = y|D = \mathbf{s},\theta) \\
&- \log \min_{\mathbf{s}\in D:P(\mathbf{s}|\theta)>0} P_{M,\theta}(M(\mathbf{x}) = y|D = \mathbf{s},\theta) \\
\leq&\log \max_{\mathbf{s}'\in A:P(\mathbf{s}'|\theta)>0} P_{M,\theta}(M(\mathbf{x}) = y|\mathbf{s}',\theta) \\
&- \log \min_{\mathbf{s}'\in A:P(\mathbf{s}'|\theta)>0} P_{M,\theta}(M(\mathbf{x}) = y|\mathbf{s}',\theta) \leq\ \epsilon\,.
\end{aligned} \tag{15}$$

□

Discussion of Intersectionality Property: This property is philosophically concordant with intersectionality, which emphasizes empathy with all overlapping marginalized groups. However, its benefits are mainly practical: in principle, one could protect all higher-level groups in SF by specifying $\sum_{j=1}^{p} \binom{p}{j} K^j$ binary indicator protected groups, where K is the number of values per protected attribute. This quickly becomes computationally and statistically infeasible. For example, Figure 7 counts the number of protected groups that must be explicitly considered under the two intersectional fairness definitions. The intersectionality property (Theorem 3.1) implies that when the the bottom-level intersectional groups are protected (blue curve), differential fairness will automatically protect all higher-level groups and subgroups (red curve). Since subgroup fairness does not have this property, all of the groups and subgroups (red curve) must be protected explicitly with their own group indicators $g(\mathbf{s})$. Although the number of bottom-level groups grows exponentially in the number of protected attributes, the total number of groups grows much faster, at the combinatorial rate of $\sum_{j=1}^{p} \binom{p}{j} K^j$.

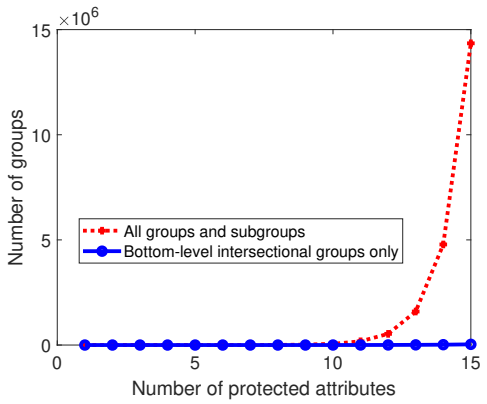


Figure 7: The number of groups and intersectional subgroups to protect when varying the number of protected attributes, with 2 values per protected attribute.

F.3 Proof of Utility Guarantee

Let $u(y) : \text{Range}(M(\mathbf{x})) \to \mathbb{R}_{\geq 0}$ be a utility function. Then:

$$E_{P_{M,\theta}}\big[u(y)|\mathbf{s}_i\big] = \int P_{M,\theta}(y|\mathbf{s}_i)u(y)dy \tag{16}$$
$$\leq \int e^{\epsilon} P_{M,\theta}(y|\mathbf{s}_j)u(y)dy = e^{\epsilon} E_{P_{M,\theta}}\big[u(y)|\mathbf{s}_j\big] \ .$$

F.4 Proof of Theorem 3.2 (Generalization Property)

Kearns et al. [18] proved that empirical estimates of the quantities per group which determine subgroup fairness, $P_{M,\theta}(y=1|g(\mathbf{s})=1)P_\theta(g(\mathbf{s})=1)$, will be similar to their true values, with enough data relative to the VC dimension of the classification model's concept class $\mathcal{H}$. We state their result below.

Theorem F.1. ***[18]'s Theorem 2.11 (SP Uniform Convergence).*** *Fix a class of functions $\mathcal{H}$ and a class of group indicators $\mathcal{G}$. For any distribution P, let $S \sim P^m$ be a dataset consisting of m examples $(\mathbf{x}_i, y_i)$ sampled i.i.d. from P. Then for any $0 < \delta < 1$, with probability $1-\delta$, for every $h \in \mathcal{H}$ and $g \in \mathcal{G}$, we have:*

$$\begin{aligned}&|P(y=1|g(\mathbf{s})=1,h)P(g(\mathbf{s})=1)\\ &- P_S(y=1|g(\mathbf{s})=1,h)P_S(g(\mathbf{s})=1)|\\ &\leq \tilde{O}\Big(\sqrt{\frac{(\mathit{VCDIM}(\mathcal{H}) + \mathit{VCDIM}(\mathcal{G}))\log m + \log(1/\delta)}{m}}\Big)\ .\end{aligned} \tag{17}$$

Here, $\tilde{O}$ hides logarithmic factors, and P_S is the empirical distribution from the S samples. It is natural to ask whether a similar result holds for differential fairness. As [18] note, the SF definition was chosen for statistical reasons, revealed in the above equation: the $P_\theta(g(\mathbf{s})=1)$ term in SF arises naturally in their generalization bound. For DF, we specifically avoid this term due to its impact on minority groups, and must instead bound $P_{M,\theta}(y|\mathbf{s})$ per group $\mathbf{s}$.

We now make use of [18]'s theorem above to prove our Theorem 3.2.

Proof. **(Theorem 3.2, Generalization Property)** Let $g(\mathbf{s}') = 1$ when $\mathbf{s}' = \mathbf{s}$ and 0 otherwise, and let $\mathcal{G} = \{g(\mathbf{s}')\}$. We see that $\mathcal{G}$ has a VC-dimension of 0. The result follows directly by applying Theorem F.1 ([18]'s Theorem 2.11) to $\mathcal{H}$ and $\mathcal{G}$, and considering the bound for the distributions P over $(\mathbf{x}, y)$ where $P(g(\mathbf{s}')=1) = 1$. □

Discussion of Generalization Property: While SF has generalization bounds which depend on the overall number of data points, DF's generalization guarantee requires that we obtain a reasonable

Models		DF-Classifier			SF-Classifier		Typical Classifier
		$\epsilon_1 = 0.0$	$\epsilon_1 = 0.2231$	$\epsilon_1 = \epsilon_{data}$	$\gamma_1 = 0.0$	$\gamma_1 = \gamma_{data}$	
Performance Measures	Accuracy	0.811	0.823	0.839	0.835	0.839	0.839
	F1 Score	0.470	0.520	0.600	0.550	0.590	0.602
	ROC AUC	0.849	0.862	0.885	0.882	0.886	0.892
Fairness Measures (using soft counts)	ϵ-DF	0.428	**0.379**	1.629	1.334	1.590	1.646
	γ-SF	**0.006**	0.012	0.039	0.026	0.034	0.041
	Bias Amp-DF	-0.952	**-1.001**	0.249	-0.046	0.210	0.266
	Bias Amp-SF	**-0.027**	-0.021	0.006	-0.007	0.001	0.008
Fairness Measures (using hard counts)	ϵ-DF	**1.602**	1.676	2.034	1.843	1.843	2.115
	γ-SF	**0.003**	0.010	0.034	0.017	0.026	0.040
	Bias Amp-DF	**-0.303**	-0.229	0.129	-0.062	-0.062	0.210
	Bias Amp-SF	**-0.037**	-0.030	-0.006	-0.023	-0.014	0.000

Table 2: Comparison of intersectionally fair classifiers with the typical classifier on the Adult dataset ($\epsilon_1 = 0.2231$ is the 80% rule).

number of data points for each intersectional group in order to accurately estimate ϵ-DF. This difference, the price of removing the minority-biasing term, should be interpreted in the context of the differing goals of our work and [18], who aimed to **prevent fairness gerrymandering** by protecting every conceivable subgroup that could be targeted by an adversary.

In contrast, our goal is to **uphold intersectionality**, which simply aims to enact a more nuanced understanding of unfairness than with a single protected dimension such as gender or race. In practice, consideration of 2 or 3 intersecting protected dimensions already improves the nuance of assessment. Sufficient data per intersectional group can often be readily obtained in such cases, e.g. [5] studied the intersection of gender and skin color on fairness. Similarly, [18] focus on the challenge of *auditing* subgroup fairness when the subgroups cannot easily be enumerated, which is important in the fairness gerrymandering setting. In contrast, in our intended applications of preserving intersectional fairness the number of intersectional groups is often only around 2^2 – 2^5.

F.5 Proof of Theorem E.1 (Confounders Property)

Proof. Let $\theta \in \Theta$, $y \in \text{Range}(M)$, $c \in C$, and $(\mathbf{s}_i, \mathbf{s}_j) \in A \times A$ where $P(\mathbf{s}_i|\theta) > 0$ and $P(\mathbf{s}_j|\theta) > 0$. We have:

$$\begin{aligned}
&\frac{P_{M,\theta}(M(x) = y|\mathbf{s}_i, \theta)}{P_{M,\theta}(M(x) = y|\mathbf{s}_j, \theta)} \\
&= \frac{\sum_{c \in C} P_{M,\theta}(M(x) = y|\mathbf{s}_i, c, \theta) P_{M,\theta}(c|\mathbf{s}_i, \theta)}{\sum_{c \in C} P_{M,\theta}(M(x) = y|\mathbf{s}_j, c, \theta) P_{M,\theta}(c|\mathbf{s}_j, \theta)} \\
&= \frac{\sum_{c \in C} \frac{P_{M,\theta}(M(x)=y|\mathbf{s}_i,c,\theta)}{P_{M,\theta}(M(x)=y|\mathbf{s}_j,c,\theta)} P_\theta(c|\mathbf{s}_i, \theta)}{\sum_{c \in C} \frac{P_{M,\theta}(M(x)=y|\mathbf{s}_j,c,\theta)}{P_{M,\theta}(M(x)=y|\mathbf{s}_j,c,\theta)} P_\theta(c|\mathbf{s}_j, \theta)} \\
&= \sum_{c \in C} \frac{P_{M,\theta}(M(x) = y|\mathbf{s}_i, c, \theta)}{P_{M,\theta}(M(x) = y|\mathbf{s}_j, c, \theta)} P_\theta(c|\mathbf{s}_i, \theta) \\
&\leq \sum_{c \in C} e^\epsilon P_\theta(c|\mathbf{s}_i, \theta) = e^\epsilon \,. \qquad (18)
\end{aligned}$$

Reversing $\mathbf{s}_i$ and $\mathbf{s}_j$ and taking the reciprocal shows the other inequality. □

G Learning Algorithm

In this section we introduce a simple, practical learning algorithm for differentially fair classifiers (*DF-Classifiers*). Our algorithm uses the fairness cost as a regularizer to balance the trade-off between fairness and accuracy. We minimize, with respect to the classifier $M_{\mathbf{W}}(\mathbf{x})$'s parameters $\mathbf{W}$, a loss function $L_{\mathbf{X}}(\mathbf{W})$ plus a penalty on unfairness which is weighted by a tuning parameter $\lambda > 0$. We train fair neural networks using gradient descent (GD) on our objective via backpropagation and automatic differentiation. The learning objective for training data $\mathbf{X}$ becomes:

$$\min_{\mathbf{W}} [L_{\mathbf{X}}(\mathbf{W}) + \lambda R_{\mathbf{X}}(\epsilon)] \qquad (19)$$

Models		DF-Classifier			SF-Classifier		Typical Classifier
		$\epsilon_1 = 0.0$	$\epsilon_1 = 0.2231$	$\epsilon_1 = \epsilon_{data}$	$\gamma_1 = 0.0$	$\gamma_1 = \gamma_{data}$	
Performance Measures	Accuracy	0.686	0.684	0.692	0.690	0.697	0.700
	F1 Score	0.633	0.642	0.643	0.622	0.647	0.641
	ROC AUC	0.730	0.723	0.734	0.719	0.739	0.734
Fairness Measures (using soft counts)	ϵ-DF	**0.180**	0.281	0.410	0.404	0.468	0.773
	γ-SF	**0.006**	0.021	0.033	0.007	0.028	0.035
	Bias Amp-DF	**-0.360**	-0.259	-0.130	-0.136	-0.072	0.233
	Bias Amp-SF	**-0.015**	0.000	0.012	-0.014	0.007	0.014
Fairness Measures (using hard counts)	ϵ-DF	**0.207**	0.671	0.884	0.825	0.860	0.897
	γ-SF	**0.015**	0.045	0.060	0.017	0.048	0.062
	Bias Amp-DF	**-0.339**	0.125	0.338	0.279	0.314	0.351
	Bias Amp-SF	**-0.025**	0.005	0.020	-0.023	0.008	0.022

Table 3: Comparison of intersectionally fair classifiers with the typical classifier on the COMPAS dataset ($\epsilon_1 = 0.2231$ is the 80% rule).

where $R_{\mathbf{X}}(\epsilon) = max(0, \epsilon_{M_{\mathbf{W}}(\mathbf{x})} - \epsilon_1)$ represents the fairness penalty term, and $\epsilon_{M_{\mathbf{W}}(\mathbf{x})}$ is the ϵ for $M_{\mathbf{W}}(\mathbf{x})$. To make the objective differentiable, $\epsilon_{M_{\mathbf{W}}(\mathbf{x})}$ is measured using soft counts (Equation 8). If ϵ_1 is 0, this penalizes ϵ-DF, and if ϵ_1 is the data's ϵ, this penalizes bias amplification. Optimizing for bias amplification will also improve ϵ-DF, up to the ϵ_1 threshold. In practice, we found that a warm start optimizing $L_{\mathbf{X}}(\mathbf{W})$ only for several "burn-in" iterations improves convergence. For large datasets, stochastic gradient descent (SGD) can be used instead of batch GD. In this case, we recommend that $\epsilon_{M_{\mathbf{W}}(\mathbf{x})}$ be estimated on a development set $\mathcal{D}$, as minibatch estimates may be unstable in the intersectional data regime.

H Additional Experimental Results

H.1 Fair Learning Algorithm

The goals of our experiments were to demonstrate the practicality of our *DF-Classifier* method in learning an intersectionally fair classifier, and to compare its behavior to a learned subgroup fair *SF-Classifier* and a typical classifier (without the fairness penalty term of Equation 19), especially with regards to minorities. Instead of [18]'s algorithm, we trained the SF-Classifier using the same GD+backpropagation approach, replacing ϵ with γ in Equation 19, i.e. $R_{\mathbf{X}}(\gamma) = max(0, \gamma_{M_{\mathbf{W}}(\mathbf{x})} - \gamma_1)$. This simplifies and speeds up learning to handle deep neural networks.

All classifiers were trained on a common neural network architecture via adaptive gradient descent optimization (Adam) with learning rate = 0.01 using pyTorch. The configuration of the neural network was 3 hidden layers, 16 neurons in each layer, "relu" and "sigmoid" activations for the hidden and output layers, respectively. We trained for 500 iterations, disabling the fairness penalties for the first 50 "burn-in" iterations. We chose λ as 0.1 and 1.0 for *DF-Classifier* and *SF-Classifier*, respectively, as a best trade-off value via grid search over the randomly held out 20% development sets.

We learned fair classifiers in several settings: 1) we set the target thresholds to perfect fairness, ϵ_1=0.0 and γ_1=0.0 for DF-Classifier and SF-Classifier, respectively, and 2) to penalize bias amplification by the algorithm, by setting the thresholds to ϵ_1=ϵ_{data} and γ_1=γ_{data} for DF-Classifier and SF-Classifier, respectively. Finally, to protect the 80%-rule we set ϵ_1=$-\log 0.8 = 0.2231$ for DF-Classifier only. Since there is no straightforward way to enforce the 80%-rule for SF-Classifier, it was not considered in this analysis.

Tables 2 and 3 compare the classifiers on the Adult and COMPAS datasets, respectively. Both DF-Classifier and SF-Classifier were able to **substantially improve their fairness metrics over the typical classifier, with modest costs in accuracy, F1 score, and ROC AUC**, and the trade-off varied roughly monotonically in the target value ϵ_1 or γ_1. Based on *soft count* estimation (Equation 8), the DF-Classifier with $\epsilon_1 = 0$ improved from $\epsilon = 1.646$ to $\epsilon = 0.428$ on Adult with a loss of 2.8 percentage points of accuracy. On COMPAS, it improved from $\epsilon = 0.773$ to $\epsilon = 0.180$, corresponding to a worst-case difference in utility between groups of a factor of $e^{\epsilon} \approx 1.2$, with a loss of just 1.4 percentage points of accuracy. When trained to prevent **bias amplification**, the fairness metrics were improved with **little** (COMPAS) **to no** (Adult) **reduction in accuracy**. While SF-Classifier typically had slightly higher accuracy under the same settings, **DF-Classifier often greatly improved γ-SF as well, while SF-Classifier enjoyed only modest improvements in ϵ-DF**.

	Gini Coefficient (G)			
Dataset	ϵ_{Data}	γ_{Data}	ϵ_{LR}	γ_{LR}
Adult	**0.099**	0.256	**0.126**	0.257
COMPAS	**0.151**	0.376	**0.135**	0.343

Table 4: Comparison of the inequity in the per-group allocation of the ϵ-DF and γ-SF metrics via the Gini coefficient (lower is better).

COMPAS Dataset		
Protected attributes	ϵ-DF	γ-SF
race	0.1003	0.0070
gender	0.9255	**0.0656**
race, gender	1.3156	0.0604

Adult Dataset		
Protected attributes	ϵ-DF	γ-SF
nationality	0.2177	0.0045
race	0.9188	0.0128
gender	1.0266	**0.0434**
gender, nationality	1.1511	0.0431
race, nationality	1.1534	0.0163
race, gender	1.7511	0.0451
race, gender, nationality	1.9751	0.0455

Table 5: Protection of intersectionality by DF metric on COMPAS and Adult dataset. The cases in red are where γ-SF violates the intersectionality property enjoyed by ϵ-DF (Theorem 3.1).

The conclusions were similar with "hard count" smoothed EDF estimates (Equation 7), but the metrics' estimates were higher.

H.2 Inequity of Fairness Measures

We have seen that the γ-SF metric downweights the consideration of minorities (cf. Figure 2). In this experiment, we quantify the resulting inequity of fairness consideration using the *Gini coefficient* [22], a commonly used measure of statistical dispersion which is often used to represent the inequity of income. The *Gini coefficient (G) of a fairness metric F* is calculated as

$$G = \frac{1}{2\mu} \sum_{i=1}^{n} \sum_{j=1}^{n} P(\mathbf{s}_i) P(\mathbf{s}_j) |F_{\mathbf{s}_i} - F_{\mathbf{s}_j}| \,, \tag{20}$$

where $\mu = \sum_{i=1}^{n} F_{\mathbf{s}_i} P(\mathbf{s}_i)$ and $P(\mathbf{s}_i)$ is the fraction of population belonging to the i^{th} intersectional group, while $F_{\mathbf{s}_i}$ represents the fairness measure (i.e. per-group ϵ or γ) of that group. For a fixed algorithm and data distribution, a fairness metric with a smaller Gini coefficient distributes its (un)fairness consideration more equitably across the population, which is typically desirable in the sense that *the entire population has a voice in the determination of (un)fairness*.

Table 4 shows a comparison of G values for the ϵ-DF and γ-SF metrics on the Adult and COMPAS datasets. Both fairness metrics are measured for the labeled dataset (i.e. ϵ_{Data}) as well as for a logistic regression (LR) classifier (i.e. ϵ_{LR}) trained on the same dataset. In all the experiments, the G value for ϵ-DF is much lower compared to γ-SF's G value. Thus, ϵ-DF was observed to provide a more equitable distribution of its per-group fairness measurements, presumably due to its more inclusive treatment of minority groups.

H.3 Evaluation of Intersectionality Property

In our final experiment (Table 5), we studied the ability of γ-SF to preserve the intersectionality property shown for ϵ-DF in Theorem 3.1, by measuring fairness with different sets of protected attributes. The property is violated if removing a protected attribute increases the metric. As expected, ϵ-DF obeyed the intersectionality property, but γ-SF violated it as γ for *gender* $>$ γ for *race* $\times$ *gender* (COMPAS), and γ for *gender* $>$ γ for *gender* $\times$ *nationality* (Adult).

I Relationship Between DF and Pufferfish Privacy

Pufferfish [19] is a general privacy framework which extends differential privacy to protect arbitrary pairs of secrets. Differential fairness adapts the pufferfish privacy framework to the task of defining algorithmic fairness, by selecting a set of protected attributes as the secrets, and ensuring that the values of these attributes are indistinguishable. Thus, differential fairness provides a closely related privacy guarantee to differential privacy.

Definition I.1. *A mechanism $M(\mathbf{x})$ is* ϵ-pufferfish private *[19] in a framework (S, Q, Θ) if for all $\theta \in \Theta$ with $\mathbf{x} \sim \theta$, for all secret pairs $(\mathbf{s}_i, \mathbf{s}_j) \in Q$ and $y \in$ Range(M),*

$$e^{-\epsilon} \leq \frac{P_{M,\theta}(M(\mathbf{x}) = y|\mathbf{s}_i, \theta)}{P_{M,\theta}(M(\mathbf{x}) = y|\mathbf{s}_j, \theta)} \leq e^{\epsilon}\,, \tag{21}$$

when $\mathbf{s}_i$ and $\mathbf{s}_j$ are such that $P(\mathbf{s}_i|\theta) > 0$, $P(\mathbf{s}_j|\theta) > 0$.