

Creative Commons Attribution 4.0 International (CC BY 4.0)

<https://creativecommons.org/licenses/by/4.0/>

Access to this work was provided by the University of Maryland, Baltimore County (UMBC) ScholarWorks@UMBC digital repository on the Maryland Shared Open Access (MD-SOAR) platform.

Please provide feedback

Please support the ScholarWorks@UMBC repository by emailing scholarworks-group@umbc.edu and telling us what having access to this work means to you and why it's important to you. Thank you.

Article

Differential Fairness: An Intersectional Framework for Fair AI [†]

Rashidul Islam ^{1,‡,§}, Kamrun Naher Keya ¹, Shimei Pan ¹, Anand D. Sarwate ²  and James R. Foulds ^{1,*} 

¹ Department of Information Systems, University of Maryland, Baltimore County, Baltimore, MD 21250, USA; islam.rashidul@umbc.edu (R.I.); kkeya1@umbc.edu (K.N.K.); shimei@umbc.edu (S.P.)

² Department of Electrical and Computer Engineering, Rutgers, The State University of New Jersey, New Brunswick, NJ 08854, USA; anand.sarwate@rutgers.edu

* Correspondence: jfoulds@umbc.edu

[†] This paper is an extended version of our paper published in ICDE 2020, Dallas, TX, USA, 20–24 April 2020.

[‡] Current address: Visa Research, Visa Inc., Atlanta, GA 30309, USA. The work was done while at UMBC.

[§] These authors contributed equally to this work.

Abstract: We propose definitions of fairness in machine learning and artificial intelligence systems that are informed by the framework of intersectionality, a critical lens from the legal, social science, and humanities literature which analyzes how interlocking systems of power and oppression affect individuals along overlapping dimensions including gender, race, sexual orientation, class, and disability. We show that our criteria behave sensibly for any subset of the set of protected attributes, and we prove economic, privacy, and generalization guarantees. Our theoretical results show that our criteria meaningfully operationalize AI fairness in terms of real-world harms, making the measurements interpretable in a manner analogous to differential privacy. We provide a simple learning algorithm using deterministic gradient methods, which respects our intersectional fairness criteria. The measurement of fairness becomes statistically challenging in the minibatch setting due to data sparsity, which increases rapidly in the number of protected attributes and in the values per protected attribute. To address this, we further develop a practical learning algorithm using stochastic gradient methods which incorporates stochastic estimation of the intersectional fairness criteria on minibatches to scale up to big data. Case studies on census data, the COMPAS criminal recidivism dataset, the HHP hospitalization data, and a loan application dataset from HMDA demonstrate the utility of our methods.

Keywords: fairness in AI; AI and society; intersectionality; 80% rule; privacy



Citation: Islam, R.; Keya, K.N.; Pan, S.; Sarwate, A.D.; Foulds, J.R. Differential Fairness: An Intersectional Framework for Fair AI. *Entropy* **2023**, *25*, 660. <https://doi.org/10.3390/e25040660>

Academic Editors: Prasanna Sattigeri and Yuheng Bu

Received: 20 January 2023

Revised: 25 March 2023

Accepted: 1 April 2023

Published: 14 April 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

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 [1–3]. A number of studies have subsequently demonstrated that bias and fairness issues in AI are both harmful and pervasive [4–6]. The AI community has responded by developing a broad array of mathematical formulations of fairness and learning algorithms which aim to satisfy them [7–12]. Fairness, however, is not a purely technical construct, having social, political, philosophical, and legal facets [13]. At this juncture, the necessity has become clear for interdisciplinary analyses of fairness in AI and its relationship to society, to civil rights, and to the social goals that mathematical fairness definitions are intended to support, which have not always been made explicit [14].

In particular, it is important to connect fairness and bias in algorithms to the broader context of fairness and bias in society, which has long been the concern of civil rights and feminist scholars and activists [3,15]. 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 [16]. More concretely, intersectionality analyzes how

systems of oppression in our society interact and interlock to create patterns of advantage and disadvantage that affect people differently at the intersection of multiple protected dimensions. Thus, in an AI fairness context, intersectionality implies not only that fairness criteria should consider multiple protected attributes simultaneously, but also that they should account for the impact of these unfair systems of society. While multi-attribute fairness definitions have previously been proposed [17,18], they fall short when it comes to the latter aspect of intersectionality. This research aims to fill this gap.

In this work, we propose **intersectional AI fairness criteria** and perform a *comprehensive, interdisciplinary analysis* of their relation to the concerns of diverse fields including the **humanities, law, privacy, economics, and statistical machine learning**. In order to create our quantitative measure of fairness, we must first understand how the principles that underlie intersectionality can be manifested in terms of decision making and in particular how harms are manifested via causal assumptions and how ideal outcomes differ from prior models of fairness. For instance, as we will discuss in Section 3, fairness measures that consider intersections of sensitive attributes or protected classes may not explicitly address how complex and interdependent systems of power and privilege impact those intersecting groups. In Section 2 we will formulate desirable properties that *any* intersectional fairness metric should have, motivating our approach.

We propose a quantitative measure of fairness, called *differential fairness*, and show how it can be used to capture some aspects of intersectionality. In particular, we show how it can be used to quantify harms from fairness violations. To obtain this interpretable guarantee, our approach builds upon ideas from differential privacy, a mathematical quantification of the privacy-preserving properties of an algorithmic mechanism [19,20]. Differential privacy is an information-theoretic privacy notion, in that it is a special case of Rényi differential privacy, a more general notion that is defined in terms of the Rényi divergence [21]. The definition of differential privacy, due to [19,20], follows:

Definition 1. $M(\mathbf{x})$ is ϵ -differentially private if

$$\frac{P(M(\mathbf{x}) \in \mathcal{S})}{P(M(\mathbf{x}') \in \mathcal{S})} \leq e^\epsilon \quad (1)$$

for all outcomes \mathcal{S} and pairs of databases \mathbf{x}, \mathbf{x}' differing in a single element.

Here, $M(\mathbf{x})$ is a randomized algorithmic mechanism, and ϵ is the degree of differential privacy achieved by $M(\mathbf{x})$ (lower is better). Essentially, differential privacy is a promise; if an individual contributes their data to a dataset, their resulting utility, due to algorithms applied to that dataset, will not be substantially affected if ϵ is small. A particular value of ϵ corresponds to an economic guarantee: a bound on the change in the expected utility for any individual who contributes their data. As such, differential privacy *operationalizes algorithmic privacy* in terms of provable guarantees on harms due to the algorithm. Differential privacy thus has *operational significance*, i.e., its measurement represents the construct of “privacy” in a meaningful way as it has a direct relationship to the real-world privacy-related harms that may occur. The potential data contributors and other stakeholders can hence interpret what the value of the differential privacy parameter ϵ means to them.

More generally, mathematical definitions that aim to operationalize abstract constructs ideally should similarly have operational significance. For example, the source coding theorem shows that Shannon entropy operationalizes the notion of uncertainty by providing an interpretation with real-world implications: a bound on the number of bits per symbol that data from a given distribution can be compressed down to. It is natural to ask whether these ideas from algorithmic privacy can be translated to the related problem of algorithmic fairness: *can we operationalize fairness with a definition that provides an interpretable guarantee on the fairness harms due to the algorithm?*

In this work, we answer this question in the affirmative. We propose a fairness definition, which we call *differential fairness*, that adapts differential privacy to the fairness

setting. Crucially, it provably inherits privacy and economic guarantees that are analogous to those of differential privacy, allowing suitably informed stakeholders to similarly interpret its meaning. Hence, similar to differential privacy, it has *operational significance*, in the sense that it meaningfully represents the construct of “fairness” in terms of bounds on possible harms due to the algorithm. (Here, “possible harms” refers to direct harms that the algorithm produces in an idealized setting. This does not account for additional harms due to its interaction with a larger sociotechnical system [22] or harms that are not captured by the fairness metric itself [23]. See Section 1.3 for further discussion). The meaningful interpretation of the metric is particularly important for AI fairness, as it is crucial that affected individuals and stakeholders are able to understand their level of risk in order to make informed decisions, both at the personal level and at the policy level. The legal obligation to provide information relating to these risks is trending upwards globally; see, e.g., Europe’s General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA). (The interpretability properties of the fairness metric do not guarantee that stakeholders with non-technical backgrounds will be able to interpret it correctly. Addressing this is an ongoing research challenge [24–26]. See Section 1.3).

1.1. Differential Fairness: A Definition

We adapt the notation of [27] 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, \dots, S_p are discrete-valued protected attributes, $A = S_1 \times S_2 \times \dots \times S_p$, 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). We suppose that the user of the assigned outcomes (the *vendor*), who may not be the *data owner*, may be *untrusted* and should not access the input data [7]. The setting is illustrated in Figure 1.

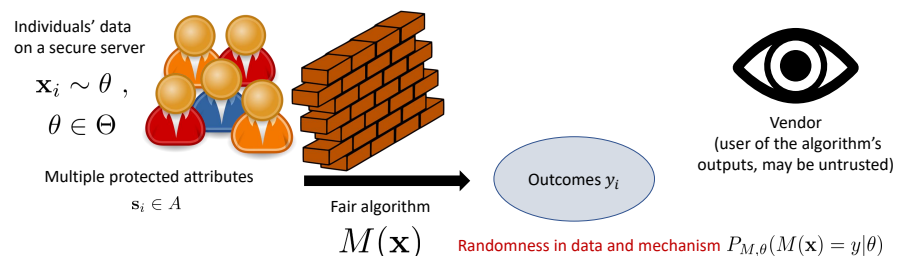


Figure 1. Diagram of the setting for the proposed differential fairness criterion.

Similar to differential privacy, our proposed *differential fairness* definition bounds ratios of probabilities of outcomes resulting from a mechanism. (Differential fairness measures the fairness of the mechanism and not the fairness of a larger sociotechnical system in which it is embedded (see Section 1.3). For example, if the mechanism is an AI that is used to make recommendations to the vendor, it does not account for the vendor injecting additional bias into a subsequent decision. If appropriate for a particular application, human decision making can also be factored into the “mechanism,” thereby measuring the fairness of the human–AI super-system instead of just that of the AI).

Definition 2. 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 $y \in \text{Range}(M)$,

$$e^{-\epsilon} \leq \frac{P_{M,\theta}(M(\mathbf{x}) = y | \mathbf{s}_i, \theta)}{P_{M,\theta}(M(\mathbf{x}) = y | \mathbf{s}_j, \theta)} \leq e^{\epsilon}, \quad (2)$$

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 (2), $\mathbf{s}_i, \mathbf{s}_j \in A$ 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} . (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 [27]. Continuous protected attributes are also possible, in which case sums are replaced by integrals in our proofs). 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. The probabilities $P_{M,\theta}$ are over the randomness in the data as well as in the mechanism, so it is possible to achieve the definition when the mechanism is deterministic, unlike for differential privacy.

1.2. Our Contribution

While the similarity of the mathematical formulation to differential privacy is clear from a comparison of Equations (1) and (2), in what follows we will further motivate this definition from interdisciplinary perspectives, i.e., based on principles from the law and the humanities. In particular, we will provide arguments that motivate differential fairness as upholding the values associated with *intersectionality* [16], a framework for examining fairness, which we will discuss in Section 2.

Our contributions include:

1. A critical analysis of the consequences of intersectionality in the particular context of fairness for AI (Section 2);
2. Three novel fairness metrics: *differential fairness* (DF), which aims to uphold intersectional fairness for AI and machine learning systems (Section 4), *DF bias amplification*, a slightly more conservative fairness definition than DF (Section 6), and *differential fairness with confounders* (DFC) (Section 7);
3. Illustrative worked examples to aid with understanding (Section 5);
4. Proofs of the desirable intersectionality, privacy, economic, and generalization properties of our metrics (Section 8);
5. A simple learning algorithm which enforces our criteria using batch or deterministic gradient methods (Section 9.1);
6. A practical and scalable learning algorithm that enforces our fairness criteria using stochastic gradient methods (Section 9.2);
7. Extensive experiments on census, criminal recidivism, hospitalizations, and loan application data to demonstrate our methods' practicality and benefits (Section 10).

As AI fairness is a problem that extends well beyond the boundaries of traditional computer science, our work aims to cross disciplinary boundaries in order to tackle it. We view the significance of this work as more than simply the introduction of mathematical definitions and associated algorithms and proofs. Our definitions and proofs build upon existing work on differential privacy, which leads to results that are elegant in their simplicity and yet are relatively straightforward. We do not claim that the technical novelty of our results or the difficulty of proofs are the most significant aspects of the research. Rather, we view the primary significance of our proposed differential fairness definition as threefold:

- It operationalizes AI fairness in a manner that is more meaningful than previous definitions, in the sense that it can be interpreted in terms of economic and privacy guarantees, leveraging connections to differential privacy.
- It is the first AI fairness definition to implement the principles and values behind intersectionality, an important notion of fairness from the humanities and legal literature. As such, part of the significance of this work is philosophical, not only mathematical.
- It is a stepping stone toward the holistic, trans-disciplinary research endeavor that is necessary to adequately solve the real-world sociotechnical problem of AI fairness.

1.3. Limitations and Caveats

In order to properly contextualize our contributions, we must consider the limitations of the use of AI fairness metrics, including but not limited to our proposed metrics,

in pursuit of the real-world social goal of *fairness*, or more specifically, *intersectional fairness*. In general terms, because AI algorithms are embedded in complicated sociotechnical systems, enforcing a fairness definition—even one with provable guarantees—is not generally sufficient to ensure that the overall decision-making process, and its higher-order consequences, are necessarily fair. We discuss several aspects of these limitations and challenges below.

Accounting for the Sociotechnical Context: Algorithmic systems do not generally exist in a vacuum. They are deployed within the context of larger real-world sociotechnical systems. Our provable guarantees on the harms caused by an algorithmic system pertain to the direct impact of the system’s output, but not to the complicated interaction between that algorithm and its broader context, which can be difficult to account for [22].

One important aspect of the sociotechnical context is whether the AI algorithm’s output is the final decision that is made on an individual or whether it is used to support such a decision that is ultimately made by a human or another algorithm. For example, in AI-generated criminal justice risk assessments such as those made by the COMPAS system [4], the risk score for an incarcerated individual is typically given to a judge who takes it into account, along with additional information on the individual and their case, in order to make bail and sentencing decisions.

Fairness issues can arise at the nexus between a human decision maker and the AI. For example, Green and Chen [28] found that the presence of an algorithmic risk assessment leads humans to increase the prioritization of risk at the expense of other factors. Humans are frequently poor at overseeing AI systems due to issues such as *automation bias* (deferring to an algorithm even when it is not appropriate), and human oversight can legitimize flawed human–AI decision-making processes [29]. Like other AI fairness metrics, differential fairness does not account for such issues.

Indeed, our theoretical analysis assumes that *the decisions made on the individuals correspond to the output of the mechanism, and not decisions made by a human decision maker who is merely informed by the mechanism*. More specifically, we assume a particular protocol in which the individuals’ data are kept on a secure server and the output of the mechanism is the only information available to the user of these outputs, referred to as the *vendor* (see Figure 1). If this protocol is not respected, e.g., if the vendor is allowed to look at the individuals’ data and could choose to completely ignore the output of the mechanism in making decisions, or deviate from the algorithm’s outputs in systematically biased ways, all bets are clearly off regarding the fairness of the overall sociotechnical system [22]. In the scenario where a judge uses knowledge of the defendant in addition to the AI-generated risk score to make a bail or sentencing decision, our assumed protocol has been violated (since the vendor, i.e., the judge, looked at the individual’s data, and they made a decision that differed from the algorithm’s output), and so the overall sociotechnical decision-making process will not enjoy the provable guarantees afforded by the differential fairness definition, as measured on the AI risk scoring system.

One way around this is to view the human decision maker as part of the mechanism. We can straightforwardly estimate the differential fairness of this human–AI super-system empirically (see Section 4.2). However, our learning algorithm would not be applicable since it is not currently possible to train a human via backpropagation!

Mismatch Between Fairness Metrics and Real-World Fairness Harms: The measurement of fairness depends heavily on assumptions made when modeling the problem at hand, particularly when defining and measuring class labels that encode abstract constructs and are intended to act as proxies for decisions. For example, the COMPAS system defines its prediction target, *recidivism*, as a new misdemeanor or felony arrest within two years [4]. This assumption conflates arrests with crimes, thereby encoding systemic bias in policing into the class label itself [23]. Due to these types of issues, there is generally a gap in alignment between mathematical fairness metrics and the desired fairness construct regarding the real-world harms that they are intended to measure and prevent [23]. Whenever we

interpret our mathematical guarantees on fairness-related harms due to an algorithm, it is important to recognize that the mathematical definition may not perfectly align with real-world harms.

Countering Structural Oppression: Recently, Kong [30] critiqued fairness definitions based on multiple tuples of attributes [17,18,31] regarding the extent to which they implement intersectionality. It was argued that parity-based fairness does not place the proper emphasis on structural oppression, which is a core tenet of intersectionality, and may not go far enough to rectify such oppression. We broadly agree with this statement. However, as we argue here and in prior work [32], parity-based fairness is an important starting point toward addressing structural oppression which is appropriate for many applications. In particular, we try to explicitly incorporate the causal structure of interlocking structural oppression. We believe our definition can be extended to correct societal unfairness beyond the point of parity, as advocated by [30]. However, we leave this for future work.

Stakeholder Interpretation: We have motivated our work via the interpretation that our fairness definition’s provable guarantees provide. We note, however, that it is not always straightforward to ensure that stakeholders correctly interpret the fairness properties of AI systems, particularly when they do not have technical backgrounds, and there is a growing body of research on this challenge and how best to address it [24–26]. Studying how stakeholders interpret and perceive differential fairness is beyond the scope of this paper, but is an important question for future research.

Contextual Considerations: The impact of systems of oppression and privilege may vary contextually. For example, gender discrimination varies widely by area of study and this is a major factor in the choices women make in selecting college majors [33]. Similarly, in some contexts individuals from a particular race, gender, or sexual orientation may be able to “pass” as another, thus potentially avoiding discrimination, and in other contexts where their “passing” status is revealed, those individuals may suffer additional discrimination [34]. Our fairness framework does not presently consider these contextual factors.

Knowledge of Protected Attributes: In order to compute (or enforce) our differential fairness criterion, we must observe the individuals’ protected attributes. This is not always possible in certain applications. For instance, social media users may not declare their gender, race, or other protected demographic information. Future work could address the extension of our methods to handle this scenario, e.g., via an imputation strategy.

2. Intersectionality and Fairness in AI

Our definition(s) differ *quantitatively* from prior fairness definitions (which we address in Section 3) but more importantly represent a *qualitatively* different approach to defining fairness. In order to understand what an intersectional analysis would ask from a fairness criterion, we must describe the challenges posed by legal, political, and humanistic analyses of fairness and discrimination. We now turn to describing these challenges and our desiderata for any fairness definition which seeks to address them.

Intersectionality is a lens for examining societal unfairness which originally arose from the observation that sexism and racism have intertwined effects, in that the harm done to Black women by these two phenomena is more than the sum of the parts [16,35]. The notion of intersectionality was later extended to include overlapping injustices along more general axes [36]. In its general form, 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 [16,35–39]. (Please note that bell hooks prefers that her name be written in lower case, cf. <https://www.nytimes.com/2006/09/28/books/28chic.html> accessed on 4 April 2023.) These systems are interlocking in their effects on individuals at *each intersection of the affected dimensions*.

The term *intersectionality* was introduced by Kimberlé Crenshaw in the 1980s [16] and popularized in the 1990s, e.g., by Patricia Hill Collins [36], although the ideas are much older [35,37]. In the context of machine learning and fairness, intersectionality was considered by

Buolamwini and Gebru [5], who studied the impact of the intersection of gender and skin color on computer vision performance, by Kearns et al. [17] and Hebert-Johnson et al. [18], who aimed to protect certain subgroups in order to prevent “fairness gerrymandering,” and by Yang et al. [40], who considered algorithmic fairness across multiple overlapping groups simultaneously. From a humanities perspective, Noble [3] critiqued the behavior of Google search with an intersectional lens by examining the search results for terms relating to women, people of color, and their intersections, e.g., “Black girls.” Very recently La Cava et al. [41] proposed intersectional fairness definitions combining our differential fairness work and multicalibration [18], and Lett and La Cava [42] advanced a perspective on how intersectionality informs AI fairness in healthcare applications.

Intersectionality has implications for AI fairness beyond the use of multiple *protected attributes*. Many 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 police stop) should be taken into account when determining whether bias has occurred [43]. (More specifically, Simoiu et al. [43] argue that risk scores are meritocratic, that fairness corresponds to thresholding risk scores at the same point for different demographics, and that parity-based notions of fairness are hence problematic because they may be incompatible with this version of fairness. They refer to this concern as the *infra-marginality problem*. While “infra-marginality” refers to a technical point, we use the term to also refer to the philosophical viewpoint underlying this argument which emphasizes meritocracy and takes an anti-parity stance. For more discussion, please see [32].) A closely related argument is that parity of outcomes between groups is at odds with accuracy [7,8]. Intersectionality theory provides a counterpoint: these differences in risk/merit, while acknowledged, are frequently due to systemic structural disadvantages such as racism, sexism, inter-generational poverty, the school-to-prison pipeline, mass incarceration, and the prison-industrial complex [16,37,38,44,45]. Systems of oppression can lead individuals to perform below their potential, for instance, by reducing available cognitive bandwidth [46] or by increasing the probability of incarceration [44,47]. 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** (see Figure 2).

As an example of a scenario affected by unfair processes, consider the task of predicting prospective students’ academic performance for use in college admissions decisions. As discussed in detail by [46] and references therein, individuals belonging to marginalized and non-majority groups are disproportionately impacted by challenges of poverty and racism (in its structural, overt, and covert forms), including chronic stress, access to healthcare, under-treatment of mental illness, micro-aggressions, stereotype threat, disidentification with academics, and belongingness uncertainty. Similarly, LGBT and especially transgender, non-binary, and gender non-conforming students disproportionately suffer bullying, discrimination, self-harm, and the burden of concealing their identities. These challenges are often further magnified at the intersection of affected groups. A survey of 6,450 transgender and gender non-conforming individuals found that the most serious discrimination was experienced by people of color, especially Black respondents [48]. Verschelden explains the impact of these challenges as a tax on the “cognitive bandwidth” of minoritized students, which in turn affects their academic performance. She states that the evidence is clear

“...that racism (and classism, homophobia, etc.) has made people physically, mentally, and spiritually ill and dampened their chance at a fair shot at higher education (and at life and living).”

A classifier trained to predict students’ academic performance from historical data hence aims to emulate outcomes that were substantially affected by unfair factors [1]. An *accurate predictor* for a student’s GPA may therefore not correspond to a *fair decision-making*

procedure [49]. We can resolve this apparent conflict if we are careful to distinguish between the *statistical problem* of classification and the *economic problem* of the assignment of outcomes (e.g., admission decisions) to individuals based on classification. (Recall that we assume the outputs of the mechanism correspond to the final decisions made on the individuals.) Viewing the classifier's task as a policy question, it becomes clear that high accuracy need not be the primary goal of the system, especially when we consider that "accuracy" is measured on unfair data. (Amazon abandoned a job candidate classifier which was found to be gender-biased [50]. We speculate that this was likely due to similar issues.)

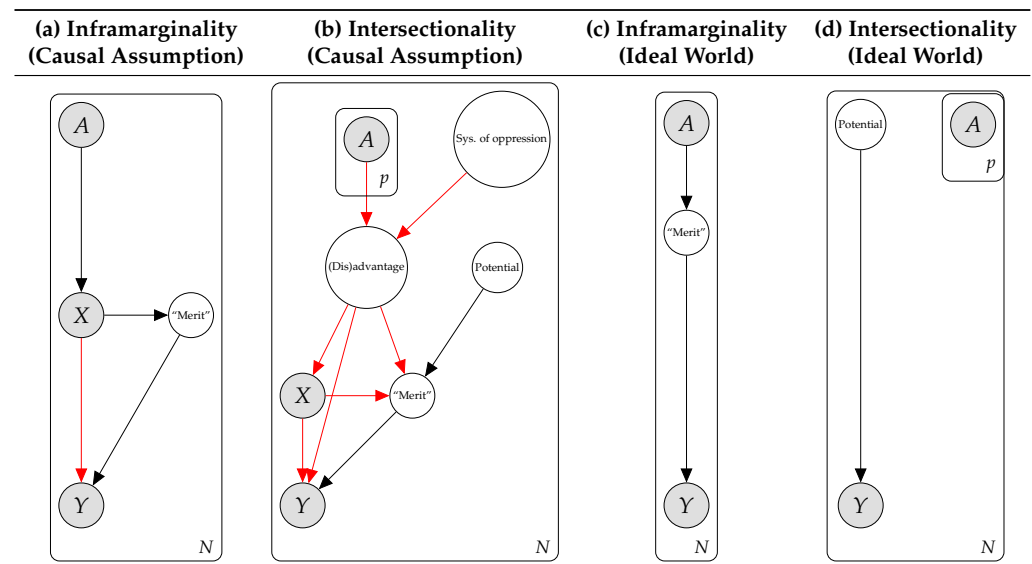


Figure 2. Implicit causal assumptions (a,b) and values-driven ideal-world scenarios (c,d) for infra-marginality and intersectionality notions of fairness. Here, A denotes protected attributes, X observed attributes, Y outcomes, N individuals, p number of protected attributes. Red arrows denote potentially 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.

In Figure 2 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 [43]. 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 [16]. In the ideal world that an algorithmic (or other) intervention aims to achieve, infra-marginality-based fairness desires that individual "merit" be the sole determiner of outcomes (c) [8,43], which can lead to disparity between groups [7]. In ideal intersectional fairness (d), since ability to succeed is affected by unfair processes, it is desired that this unfairness be corrected and individuals achieve their true potential [46]. Assuming potential does not substantially differ across protected groups, this implies that parity between groups is typically desirable. (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 scenario in Section 7.)

In light of the above, we argue that an *intersectional* definition of fairness in AI should satisfy the following criteria:

- A Multiple protected attributes should be considered.
- B All of the intersecting values of the protected attributes, e.g., *Black women*, should be protected by the definition.
- C The definition should still also ensure that protection is provided on individual protected attribute values, e.g., *women*.
- D The definition should ensure protection for minority groups, who are particularly affected by discrimination in society [36].
- E The definition should ensure that systematic differences between the protected groups, assumed to be due to structural oppression, are rectified, rather than codified.

These desiderata do not uniquely specify a fairness definition, but they provide a set of guidelines to which legal, political, and contextual considerations can then be applied to determine an appropriate fairness measure for a particular task.

3. Related Work on Fairness

Before providing our new fairness definition, we discuss some of the existing fairness definitions and their relation to the aforementioned criteria. An overview of AI fairness research can be found in [49].

3.1. Models for Fairness

The 80% rule: Our criterion is related to the *80% rule*, also known as the *four-fifths rule*, a guideline for identifying unintentional discrimination in a legal setting which identifies disparate impact in cases where $P(y = 1|s_1)/P(y = 1|s_2) \leq 0.8$, for a favorable outcome $y = 1$, disadvantaged group s_1 , and best performing group s_2 [51]. This corresponds to testing that $\epsilon \geq -\log 0.8 = 0.2231$, in a version of Equation (2) where only the outcome $y = 1$ is considered.

Demographic Parity: Dwork et al. [7] defined (and criticized) the fairness notion of *demographic parity*, also known as *statistical parity*, which requires that $P(y|s_i) = P(y|s_j)$ for any outcome y and pairs of protected attribute values s_i, s_j (here assumed to be a single attribute). This can be relaxed, e.g., by requiring the total variation distance between the distributions to be less than ϵ . Differential fairness is closely related, as it also aims to match probabilities of outcomes but measures differences using ratios and allows for multiple protected attributes. The criticisms of [7] are mainly related to ways in which subgroups of the protected groups can be treated differently while maintaining demographic parity, which they call “subset targeting,” and which [17] term “fairness gerrymandering.” Differential fairness explicitly protects the intersection of multiple protected attributes, which can be used to mitigate some of these abuses.

Equalized Odds: To address some of the limitations with demographic parity, Hardt et al. [8] propose to instead ensure that a classifier has equal error rates for each protected group. This fairness definition, called *equalized odds*, can loosely be understood as a notion of “demographic parity for error rates instead of outcomes.” Unlike demographic parity, equalized odds rewards accurate classification, and penalizes systems only performing well on the majority group. However, theoretical work has shown that equalized odds is typically incompatible with correctly calibrated probability estimates [52]. It is also a relatively weak notion of fairness from a civil rights perspective compared to demographic parity, as it does not ensure that outcomes are distributed equitably. Hardt et al. also propose a variant definition called *equality of opportunity*, which relaxes equalized odds to only apply to a “deserving” outcome. It is straightforward to extend differential fairness to a definition analogous to equalized odds, although we leave the exploration of this for future work. A more recent algorithm for enforcing equalized odds and equality of opportunity for kernel methods was proposed by [53].

Individual Fairness (“Fairness Through Awareness”): The *individual fairness* definition, due to [7], mathematically enforces the principle that *similar individuals should achieve similar outcomes* under a classification algorithm. An advantage of this approach is that it preserves the privacy of the individuals, which can be important when the user of the

classifications (the *vendor*), e.g., a banking corporation, cannot be trusted to act in a fair manner. However, this is difficult to implement in practice as one must define “similar” in a fair way. The individual fairness property also does not necessarily generalize beyond the training set. In this work, we take inspiration from Dwork et al.’s *untrusted vendor* scenario and the use of a privacy-preserving fairness definition to address it.

Counterfactual Fairness: Kusner et al. [54] propose a causal definition of fairness. Under their *counterfactual fairness* definition, changing protected attributes A , while holding things which are not causally dependent on A constant, will not change the predicted distribution of outcomes. While theoretically appealing, there are difficulties in implementing this in practice. First, it requires an accurate causal model at the fine-grained individual level, while even obtaining a correct population-level causal model is generally very difficult. To implement it, we must solve a challenging causal inference problem over unobserved variables, which generally requires approximate inference algorithms. Finally, to achieve counterfactual fairness, the predictions (usually) cannot make direct use of any descendant of A in the causal model. This generally precludes using *any of the observed features* as inputs.

Threshold Tests: Simoiu et al. [43] address *infra-marginality* by modeling risk probabilities for different subsets (i.e., attribute values) within each protected category and requiring algorithms to threshold these probabilities at the same points when determining outcomes. In contrast, based on *intersectionality* theory, our proposed differential fairness criterion specifies protected categories whose intersecting subsets should be treated equally, regardless of differences in risk across the subsets. Our definition is appropriate when the differences in risk are due to structural systems of oppression, i.e., the risk probabilities themselves are impacted by an unfair process. We also provide a *bias amplification* version of our metric, following [10], which is more in line with the *infra-marginality* perspective, namely that differences in probabilities of outcomes for different groups may be considered legitimate.

3.2. Subgroup Fairness and Multicalibration

Relevant fairness definitions aim to detect and prevent discriminatory bias with respect to a *set of protected attributes*, such as *gender*, *race*, and *disability status*. Given criterion A, we focus on multi-attribute definitions. The two dominant multi-attribute approaches in the literature are *subgroup fairness* [17] and *multicalibration* [18].

Definition 3 (Statistical Parity Subgroup Fairness [17]). Let \mathcal{G} be a collection of protected group indicators $g : A \rightarrow \{0, 1\}$, 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 $g \in \mathcal{G}$,

$$|P_{M,\theta}(M(\mathbf{x}) = 1) - P_{M,\theta}(M(\mathbf{x}) = 1|g(\mathbf{s}) = 1)| \times P_\theta(g(\mathbf{s}) = 1) \leq \gamma. \quad (3)$$

Note that for $\gamma \in [0, 1]$, smaller is better. The first term penalizes a difference between the probability of the *positive* class label for group g and the population average of this probability. The term $P_\theta(g(\mathbf{s}) = 1)$ weights the penalty by the size of group g as a proportion of the population. *Statistical parity subgroup fairness* (SF) is a multi-attribute definition satisfying criterion A. To satisfy B and C criteria, \mathcal{G} can be *all* intersectional subgroups (e.g., *Black women*) and top-level groups (e.g., *men*). The first term in Equation (3), which encourages similar outcomes between groups, enforces criterion E.

From an intersectional perspective, one concern with SF is that it does not satisfy criterion D, the protection of minority groups. The term $P_\theta(g(\mathbf{s}) = 1)$ weights the “per-group (un)fairness” for each group g , i.e., Equation (3) applied to g alone, by its proportion of the population, thereby *specifically downweighting the consideration of minorities*. In Figure 3, we show an example where varying the size of a minority group $P_\theta(g(\text{minority}) = 1)$ 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 (introduced in Section 4), is constant in $P_\theta(g(\text{minority}) = 1)$.

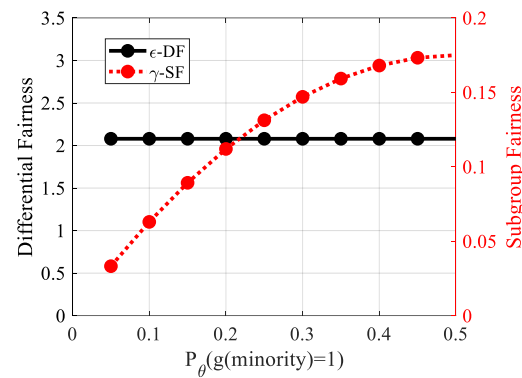


Figure 3. Toy example: probability of the “positive” class is 0.8 for a majority group, 0.1 for a minority group, varying $P_{\theta}(g(\text{minority}) = 1)$.

Figure 4 reports “per-group” γ ’s on the UCI Adult census dataset, i.e., Equation (3) applied separately to each group is empirically seen to have an increasing relationship with $P(\text{group})$ (simplified notation of $P_{\theta}(g(\text{minority}) = 1)$). The final γ -SF is determined by the worst case of the per-group γ ’s. A small minority group thereby will most likely not directly affect γ -SF, since the downweighting makes it unlikely to be the “most unfair” group.

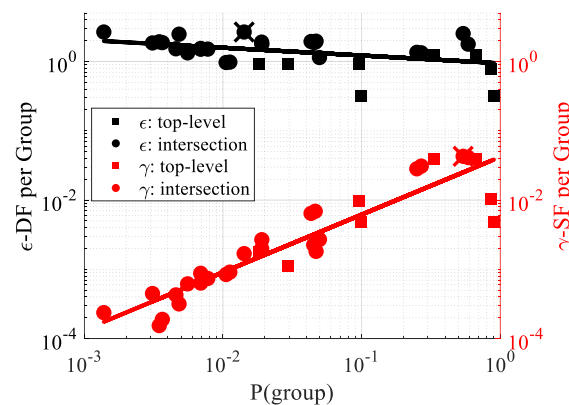


Figure 4. “Per-group” γ -SF and our proposed ϵ -DF vs. probability (i.e., size) of groups, UCI Adult dataset. Circles: intersectional subgroups (e.g., Black women of USA). Squares: top-level groups (e.g., men). Fitted least squares line demonstrates the overall trend between fairness measures and group size. Largest per-group ϵ -DF (black “X”) and γ -SF (red “X”) determines the overall ϵ -DF and γ -SF, respectively.

Kearns et al. [17] justify the use of the $P_{\theta}(g(s) = 1)$ term via statistical considerations, as it is useful to prove generalization guarantees to extrapolate from empirical estimates of γ (see Section 8.4). From a different ethical perspective, total utilitarianism, increasing the utility (i.e., reducing unfairness) for a large group of individuals at the expense of smaller groups (i.e., downweighting small minority groups) could also be justified by the increase in the total utility of the population. The problem with total utilitarianism, of course, is that it admits a scenario where many people possess low utility (i.e., higher unfairness). We do not intend to dismiss SF as a valid notion of fairness. Our claim here, rather, is simply that due to its treatment of minority groups, SF does not fully encapsulate the principles of fairness advocated by intersectional feminist scholars and activists [16,35,36,38,39].

Other candidate multi-attribute fairness definitions include *false positive subgroup fairness* [17] and *multicalibration* [18]. These definitions are similar to SF, but they concern false positive rates and calibration of prediction probabilities, respectively. Since they focus on reliability of estimation rather than allocation of outcomes, they do not directly address the issues (criterion E) raised by the civil rights/feminist perspective. This does not preclude their use for intersectional fairness scenarios in which harms are caused by

incorrect predictions (e.g., higher error rates for unprivileged groups), rather than unfair outcome assignments (e.g., higher false positive rates for privileged groups); indeed, this is the type of approach [5] take for studying intersectional fairness in computer vision applications. Nevertheless, we will not consider them further here.

4. Differential Fairness (DF)

We now introduce our proposed fairness measures which satisfy our intersectionality criteria. There are multiple conceivable fairness definitions which satisfy these criteria. For example, SF could be adapted to address criterion D by simply dropping the $P_\theta(g(s) = 1)$ term, at the loss of its associated generalization guarantees. We instead select an alternative formulation, which is similar to this approach in spirit, but which has additional beneficial properties from a societal perspective regarding the *law*, *privacy*, and *economics*, as we shall discuss below. The privacy and economic guarantees enjoyed by our definition demonstrate that it has operational significance, in that it provably prevents harms, analogously to differential privacy. (Notwithstanding the gap between the behavior of an idealized algorithmic system and the complicated impacts of deploying it within a broader real-world sociotechnical system [22] and the gap between fairness metrics and the real-world harms they are intended to measure [23]. See Section 1.3). Our formalism has a particularly elegant *intersectionality property*, in that Criterion C (protecting higher-level groups) follows automatically from Criterion B (protecting intersectional subgroups).

We motivate our criteria from a legal perspective. Consider the 80% rule, established in the Code of Federal Regulations [51] 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:

$$P(M(x) = 1 | \text{group A}) / P(M(x) = 1 | \text{group B}) < 0.8. \quad (4)$$

Our first 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 [19]. Differential fairness measures the **fairness cost** of mechanism $M(x)$ with a parameter ϵ . Combining these elements, we obtain our differential fairness criterion, which we stated above as Definition 2.

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 x . If the probabilities are always equal, then $\epsilon = 0$; otherwise, $\epsilon > 0$.

4.1. Relationship to Privacy Definitions

Differential Privacy: Like differential privacy [19,20], the differential fairness definition bounds ratios of probabilities of outcomes resulting from a mechanism. There are several important differences, however. When bounding these ratios, differential fairness considers different values of a set of protected attributes, rather than databases that differ in a single element. It posits a specified set of possible distributions which may generate the data, while differential privacy implicitly assumes that the data are independent [27]. Finally, since differential fairness considers randomness in data as well as in the mechanism, it can be satisfied with a deterministic mechanism, while differential privacy can only be satisfied with a randomized mechanism.

The study in Jagielski et al. [55] developed methods which aim to satisfy differential privacy and equalized-odds fairness simultaneously. This is a different goal to our work, which aims to ensure intersectional fairness using a fairness definition that was designed leveraging differential privacy.

Pufferfish: We have arrived at our criterion based on the 80% rule, but it can also be derived as an application of *pufferfish* [27], a generalization of differential privacy [20] which uses a variation of Equation (2) to hide the values of an arbitrary set of secrets. Different choices of the secrets and the data that the Pufferfish privacy-preserving mechanism operates on lead to both differential privacy and differential fairness.

Definition 4. A mechanism $M(\mathbf{x})$ is ϵ -pufferfish private [27] 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 $y \in \text{Range}(M)$,

$$e^{-\epsilon} \leq \frac{P_{M,\theta}(M(\mathbf{x}) = y | \mathbf{s}_i, \theta)}{P_{M,\theta}(M(\mathbf{x}) = y | \mathbf{s}_j, \theta)} \leq e^{\epsilon}, \quad (5)$$

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

While the formulation for differential fairness is mathematically very similar to Pufferfish, it is used with the goal of obtaining fairness, which leads to its more specific instantiation of the Pufferfish secrets (protected attributes), the input data (each individual's feature vectors), and the mechanism (the behavior of the classifier itself under the distribution of the data).

4.2. Empirical DF Estimation

A major challenge for measuring fairness in an intersectional context, either via ϵ -DF (differential fairness), γ -SF (subgroup fairness), or related notions, is to estimate $M(x)$'s marginal behavior $P_{M,\theta}(y | s, \theta)$ for each (y, s) pair, with potentially little data for each of these [56]. If $P_{M,\theta}$ is unknown, it can be estimated using the empirical distribution, or via a probabilistic model of the 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

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

whenever $N_{\mathbf{s}_i} > 0$ and $N_{\mathbf{s}_j} > 0$. However, in the intersectional setting, the counts $N_{y,\mathbf{s}}$ at the intersection of the values of the protected attributes become rapidly smaller as the dimensionality and cardinality of protected attributes increase. The $N_{y,\mathbf{s}}$ counts may even be 0, which can make the estimate of ϵ in Equation (6) infinite/undefined.

Alternatively, to address the zero count issue, we can estimate ϵ -DF via the posterior predictive distribution of a Dirichlet-multinomial as

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

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:

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

5. 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 they 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 ($y = \text{no}$). Presuming that the two groups are roughly equally capable of performing the job overall, this is clearly unsatisfactory in terms of fairness.

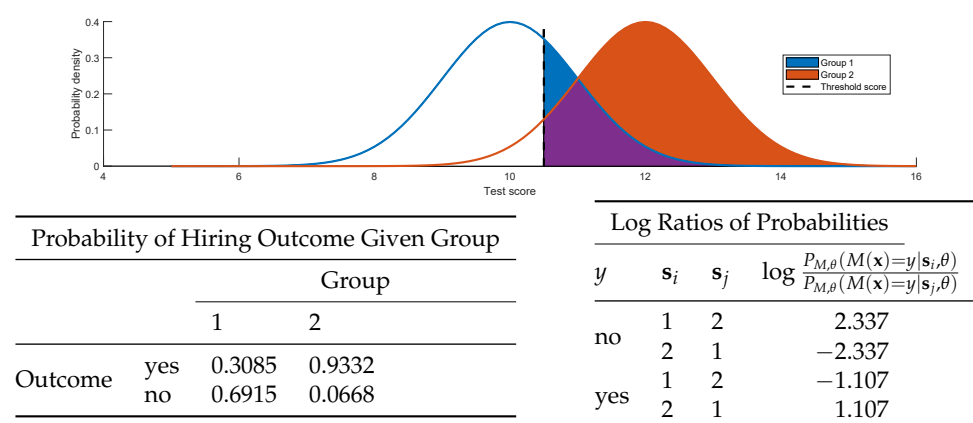


Figure 5. Worked example of differential fairness from Section 5. The calculations above show that $\epsilon = 2.337$.

Differential fairness specifically addresses the *intersectional* setting by considering outcome probabilities at each intersection of a set of protected variables. We illustrate this with an example on admissions of prospective students to a particular University X (Table 1). The protected attributes are *gender* and *race*, and the mechanism is the admissions process, with a binary outcome. Our data are adapted from a real-world scenario involving treatments for kidney stones, often used to demonstrate Simpson's paradox [57,58]. 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.

Table 1. Intersectional example: Simpson’s paradox.

Probability of Being Admitted to University X				
		Gender		Overall
		A	B	
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)	

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 $A = \text{gender} \times \text{race}$. By calculating ϵ using the admission probabilities in the *overall* row (*gender*) and the *overall* column (*race*), we find that $\epsilon = 0.2329$ for $A = \text{gender}$, and $\epsilon = 0.8667$ for $A = \text{race}$. We will prove in Theorem 2 that ϵ with $A = \text{gender} \times \text{race}$ is an upper bound on ϵ -DF for $A = \text{gender}$ and for $A = \text{race}$. Thus, even with a “Simpson’s reversal,” differential (un)fairness will not increase after summing out a protected attribute.

6. DF Bias Amplification Measure

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

Similar 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 [10] refer to as *bias amplification*.

Definition 6. 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 labeled dataset D is ϵ_1 -DF in A 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 [7]. On the other hand, $(\epsilon_2 - \epsilon_1)$ -DF **bias amplification** is a more **conservative fairness metric** than DF which does not seek to correct unfairness in the original dataset (i.e., it relaxes criterion E), in line with the principle of **infra-marginality** (*a system is biased only if disparities in its behavior are worse than those in society, which are presumed to be legitimate*) [43]. 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.

7. Dealing with Confounder Variables

As we have seen, ϵ -DF measures inequity between protected groups and their intersections at different levels of granularity, although it does not determine whether the

inequities were due to systemic factors and/or discrimination. *Confounding variables* can potentially be a legitimate source of disparity. For example, a study on U.C. Berkeley admissions in the 1970s [59] found that a disparity in overall admission rates between men and women was due to a confounding variable, the department that a prospective student chooses to apply to, with women being more likely to apply to more selective departments. This scenario was similar to the worked example in Table 1. With confounders, parity in outcomes between intersectional protected groups, which ϵ -DF rewards, may no longer be desirable (see Figure 6). We propose an alternative fairness definition for when known confounders are present.

Definition 7. 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, Θ_c) , where $\Theta_c = \{P(\mathbf{x}|\theta, c) | \theta \in \Theta\}$.

In the university admissions case, Definition 7 penalizes disparity in admissions at the department level, and the most unfair department determines the overall unfairness ϵ -DFC.

Theorem 1. Let M be an ϵ -DFC mechanism in (A, Θ, C) , then M is ϵ -differentially fair in (A, Θ) .

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:

$$\begin{aligned}
 & \frac{P_{M,\theta}(M(x) = y|\mathbf{s}_i, \theta)}{P_{M,\theta}(M(x) = y|\mathbf{s}_j, \theta)} \\
 &= \frac{\sum_{c \in C} P_{M,\theta}(M(x) = y|\mathbf{s}_i, c, \theta) P_{M,\theta}(c|\mathbf{s}_i, \theta)}{\sum_{c \in C} P_{M,\theta}(M(x) = y|\mathbf{s}_j, c, \theta) P_{M,\theta}(c|\mathbf{s}_j, \theta)} \\
 &= \frac{\sum_{c \in C} \frac{P_{M,\theta}(M(x)=y|\mathbf{s}_i, c, \theta)}{P_{M,\theta}(M(x)=y|\mathbf{s}_j, c, \theta)} P_{\theta}(c|\mathbf{s}_i, \theta)}{\sum_{c \in C} \frac{P_{M,\theta}(M(x)=y|\mathbf{s}_j, c, \theta)}{P_{M,\theta}(M(x)=y|\mathbf{s}_j, c, \theta)} P_{\theta}(c|\mathbf{s}_j, \theta)} \\
 &= \sum_{c \in C} \frac{P_{M,\theta}(M(x) = y|\mathbf{s}_i, c, \theta)}{P_{M,\theta}(M(x) = y|\mathbf{s}_j, c, \theta)} P_{\theta}(c|\mathbf{s}_i, \theta) \\
 &\leq \sum_{c \in C} e^{\epsilon} P_{\theta}(c|\mathbf{s}_i, \theta) = e^{\epsilon}.
 \end{aligned} \tag{9}$$

Reversing \mathbf{s}_i and \mathbf{s}_j shows the other inequality. \square

From Theorem 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*. If confounder variables are latent, we can attempt to infer them probabilistically in order to apply DFC. Alternatively, $(\epsilon_2 - \epsilon_1)$ -DF bias amplification can still be used to study the impact of an algorithm on fairness.

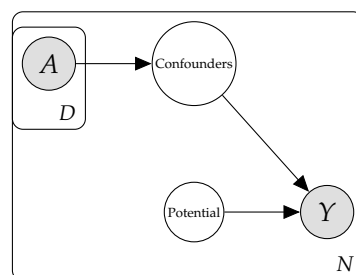


Figure 6. Ideal-world intersectional fairness with confounder variables. Disparity in overall outcomes between protected groups may be considered legitimate.

8. Properties of Differential Fairness

We now discuss the theoretical properties of our definitions.

8.1. Differential Fairness and Intersectionality

Differential fairness explicitly encodes protection of intersectional groups (Criterion B). For DF, we prove that this automatically implies fairness for *each of the protected attributes individually* (Criterion C), and indeed, *any subset* of the protected attributes. For example, if a loan approval mechanism $M(\mathbf{x})$ is ϵ -DF in $A = \text{gender} \times \text{race} \times \text{nationality}$, it is also ϵ -DF in, e.g., $A = \text{gender}$ by itself, or $A = \text{gender} \times \text{nationality}$. In other words, by ensuring fairness at the intersection of *gender*, *race*, and *nationality*, we also ensure the same degree of fairness between *genders* overall, between *gender/nationality* pairs overall, and so on. Here, ϵ is a worst case, and DF may also hold for lower values of ϵ .

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

$$\begin{aligned} & \log \max_{\mathbf{s} \in A: P(\mathbf{s}|\theta) > 0} P_{M,\theta}(M(\mathbf{x}) = y|\mathbf{s}, \theta) \\ & - \log \min_{\mathbf{s} \in A: P(\mathbf{s}|\theta) > 0} P_{M,\theta}(M(\mathbf{x}) = y|\mathbf{s}, \theta) \leq \epsilon. \end{aligned} \quad (10)$$

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 $P(\mathbf{s}_i|\theta) > 0$, $P(\mathbf{s}_j|\theta) > 0$,

$$e^{-\epsilon} \leq \frac{P_{M,\theta}(M(\mathbf{x}) = y|\mathbf{s}_i, \theta)}{P_{M,\theta}(M(\mathbf{x}) = y|\mathbf{s}_j, \theta)} \leq e^{\epsilon}. \quad (11)$$

Taking the log, we can rewrite this as:

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

The two inequalities can be simplified to:

$$|\log P_{M,\theta}(M(\mathbf{x}) = y|\mathbf{s}_i, \theta) - \log P_{M,\theta}(M(\mathbf{x}) = y|\mathbf{s}_j, \theta)| \leq \epsilon. \quad (13)$$

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

$$\begin{aligned} & |\log P_{M,\theta}(M(\mathbf{x}) = y|\mathbf{s}_i, \theta) - \log P_{M,\theta}(M(\mathbf{x}) = y|\mathbf{s}_j, \theta)| \\ & \leq \log \max_{\mathbf{s}: P(\mathbf{s}|\theta) > 0} P_{M,\theta}(M(\mathbf{x}) = y|\mathbf{s}, \theta) \\ & - \log \min_{\mathbf{s}: P(\mathbf{s}|\theta) > 0} P_{M,\theta}(M(\mathbf{x}) = y|\mathbf{s}, \theta). \end{aligned} \quad (14)$$

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

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

□

Theorem 2. (Intersectionality Property) *Let M be an ϵ -differentially fair mechanism in (A, Θ) , $A = S_1 \times S_2 \times \dots \times S_p$, and let $D = S_a \times \dots \times S_k$ 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) .*

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

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

By a similar argument,

$$\log \min_{\mathbf{s} \in D: P(\mathbf{s}|\theta) > 0} P_{M,\theta}(M(\mathbf{x}) = y | D = \mathbf{s}, \theta) \geq \log \min_{\mathbf{s}' \in A: P(\mathbf{s}'|\theta) > 0} P_{M,\theta}(M(\mathbf{x}) = y | \mathbf{s}', \theta). \quad (16)$$

Applying Lemma 1, we hence bound ϵ in (D, Θ) as

$$\begin{aligned}
 & \log \max_{\mathbf{s} \in D: P(\mathbf{s}|\theta) > 0} P_{M,\theta}(M(\mathbf{x}) = y | D = \mathbf{s}, \theta) \\
 & - \log \min_{\mathbf{s} \in D: P(\mathbf{s}|\theta) > 0} P_{M,\theta}(M(\mathbf{x}) = y | D = \mathbf{s}, \theta) \\
 & \leq \log \max_{\mathbf{s}' \in A: P(\mathbf{s}'|\theta) > 0} P_{M,\theta}(M(\mathbf{x}) = y | \mathbf{s}', \theta) \\
 & - \log \min_{\mathbf{s}' \in A: P(\mathbf{s}'|\theta) > 0} P_{M,\theta}(M(\mathbf{x}) = y | \mathbf{s}', \theta) \leq \epsilon. \quad (17)
 \end{aligned}$$

□

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 γ -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, in order to respect the intersectional fairness Criteria B and C. The intersectionality property (Theorem 2) implies that when the bottom-level intersectional groups are protected (blue curve), differential fairness will automatically protect all higher-level groups (red curve). Since statistical parity 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 $\sum_{j=1}^p \binom{p}{j} K^j$.

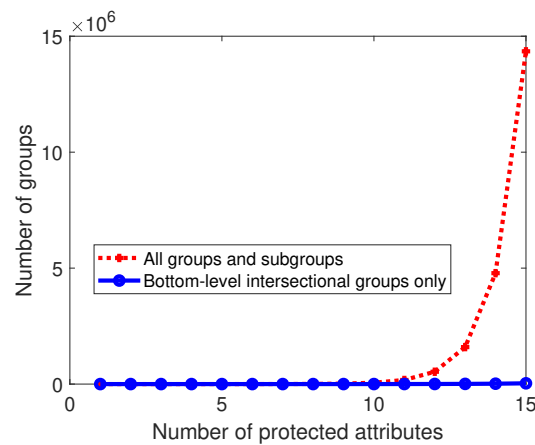


Figure 7. The number of groups and intersectional subgroups to protect when varying the number of protected attributes, with 2 values per protected attribute.

8.2. Privacy Interpretation

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 :

$$e^{-\epsilon} \frac{P(\mathbf{s}_i|\theta)}{P(\mathbf{s}_j|\theta)} \leq \frac{P(\mathbf{s}_i|M(\mathbf{x}) = y, \theta)}{P(\mathbf{s}_j|M(\mathbf{x}) = y, \theta)} \leq e^{\epsilon} \frac{P(\mathbf{s}_i|\theta)}{P(\mathbf{s}_j|\theta)}. \quad (18)$$

For example, if a loan is given to an individual, an adversary's Bayesian posterior beliefs about their *race* and *gender* will not be substantially changed. Thus, the adversary will be unable to infer that “this individual was given a loan, so they are probably white and male.” Our definition thereby provides fairness guarantees when the user of $M(\mathbf{x})$ is untrusted, cf. [7], by preventing subsequent discrimination, e.g., in retaliation to a fairness correction. Although DF is a *population-level* definition, it provides a privacy guarantee for *individuals*. 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.” The privacy guarantee is inherited from *pufferfish*, a general privacy framework which DF instantiates [27].

8.3. Economic Guarantees

We also show that an economic guarantee for DF holds. 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 that could be chosen. For example, consider a loan approval process, where the utility of being given a loan is 1, and being denied is 0. Suppose the approval process is $\ln(3)$ -differentially fair. The process could then be three times as likely to award a loan to white men as to white women, and thus award white men three times the expected utility as white women. The proof follows the case of differential privacy [20]. Let $u(y) : \text{Range}(M(\mathbf{x})) \rightarrow \mathbb{R}_{\geq 0}$ be a utility function. Then:

$$\begin{aligned} E_{P_{M,\theta}}[u(y)|\mathbf{s}_i] &= \int P_{M,\theta}(y|\mathbf{s}_i)u(y)dy \\ &\leq \int e^{\epsilon} P_{M,\theta}(y|\mathbf{s}_j)u(y)dy = e^{\epsilon} E_{P_{M,\theta}}[u(y)|\mathbf{s}_j]. \end{aligned} \quad (19)$$

Similarly, for $(\epsilon_2 - \epsilon_1)$ -DF bias amplification, $M(\mathbf{x})$ admits at most an $\exp(\epsilon_2 - \epsilon_1) \approx 1 + \epsilon_2 - \epsilon_1$ (for small values of $\epsilon_2 - \epsilon_1$) multiplicative increase in the disparity of expected

utility between pairs of protected intersections of groups with $\mathbf{s}_i \in A$, $\mathbf{s}_j \in A$, relative to the data-generating process \mathcal{M} , for any utility function.

Note that the privacy and economic guarantees afforded by differential fairness (Equations (18) and (19), respectively) show that the definition is operationally significant, in the sense that the definition provides a guarantee on the harms due to the algorithm (or at least, its direct harms, neglecting its broader impacts within the complicated sociotechnical system in which it is embedded [22]). Impacted individuals and stakeholders can interpret the mechanism's value of ϵ in terms of how its privacy and utility properties might impact them. This interpretability is a key advantage over competing definitions.

8.4. Generalization Guarantees

In order to ensure that an algorithm is truly fair, it is important that the fairness properties obtained on a dataset will extend to the underlying population. Kearns et al. [17] 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 3. [17]'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:

$$\begin{aligned} & |P(y = 1|g(\mathbf{s}) = 1, h)P(g(\mathbf{s}) = 1) \\ & - P_S(y = 1|g(\mathbf{s}) = 1, h)P_S(g(\mathbf{s}) = 1)| \\ & \leq \tilde{O}\left(\sqrt{\frac{(\text{VCDIM}(\mathcal{H}) + \text{VCDIM}(\mathcal{G})) \log m + \log(1/\delta)}{m}}\right). \end{aligned} \quad (20)$$

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 Kearns et al. [17] 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} . For this case, we prove the following generalization guarantee.

Theorem 4. Fix a class of functions \mathcal{H} , which w.l.o.g. aim to discriminate the outcome $y = 1$ from any other value, denoted 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:

$$\begin{aligned} & |P(y = 1|\mathbf{s}, h) - P_S(y = 1|\mathbf{s}, h)| \\ & \leq \tilde{O}\left(\sqrt{\frac{\text{VCDIM}(\mathcal{H}) \log m + \log(1/\delta)}{m}}\right). \end{aligned} \quad (21)$$

Proof. 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 3 (Kearns et al. [17]'s Theorem 2.11) to \mathcal{H} and \mathcal{G} and considering the bound for the distributions P over (\mathbf{x}, y) where $P(g(\mathbf{s}') = 1) = 1$. \square

While SF has generalization bounds which depend on the overall number of data points, DF's generalization guarantee requires that we obtain a reasonable 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 Kearns et al. [17], 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, the consideration of two or three intersecting protected dimensions already improves the nuance of assessment. Sufficient data per intersectional group can often be readily obtained in such cases, e.g., Buolamwini and Gebru [5] studied the intersection of *gender* and *skin color* on fairness. Similarly, Kearns et al. [17] 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 between around 2^2 (e.g., binary gender and race, as originally considered by [16]) to around 2^5 (e.g., the Civil Rights Act of 1964 prohibits discrimination along lines of five attributes: race, color, religion, sex, and national origin).

9. Learning Algorithm

In this section we introduce simple and practical learning algorithms for differentially fair classifiers (*DF-Classifiers*). Our algorithms use 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 such as cross-entropy loss plus a penalty on unfairness which is weighted by a tuning parameter $\lambda > 0$. In practice, we found that a warm start optimizing loss function L only for several “burn-in” iterations improves convergence.

The fairness penalty term $R_{\mathbf{X}}(\epsilon)$ for training data \mathbf{X} is designed as

$$R_{\mathbf{W}}(\epsilon) = \max(0, \epsilon_{M_{\mathbf{W}}(\mathbf{x})} - \epsilon_1), \quad (22)$$

where $\epsilon_{M_{\mathbf{W}}(\mathbf{x})}$ is the ϵ -DF measures for $M_{\mathbf{W}}(\mathbf{x})$. 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. We develop two approaches of learning algorithm that ensure the reliable and data-efficient estimation of $\epsilon_{M_{\mathbf{W}}(\mathbf{x})}$ in the training phase.

9.1. Batch Method

We process all the training examples $\mathcal{D} = (\mathbf{x}, y)$ simultaneously for batch DF (BDF) model $M_{\mathbf{W}^B}(\mathbf{x})$. The learning objective for BDF becomes

$$\min_{\mathbf{W}^B} \left[\frac{1}{n_{\mathcal{D}}} \sum_{i=1}^{n_{\mathcal{D}}} L(f(\mathbf{x}^{(i)}; \mathbf{W}^B), \mathbf{y}^{(i)}) + \lambda R_{\mathbf{W}^B}^{\mathcal{D}}(\epsilon) \right], \quad (23)$$

where $R_{\mathbf{W}^B}^{\mathcal{D}}(\epsilon) = \max(0, \epsilon_{M_{\mathbf{W}^B}(\mathbf{x})} - \epsilon_1)$ represents the fairness penalty term in the batch method, and $f(\mathbf{x}; \mathbf{W}^B)$ is the predicted output (analogous to $P(y|\mathbf{x})$ in Equation (8)) for $n_{\mathcal{D}}$ training examples. $\epsilon_{M_{\mathbf{W}^B}(\mathbf{x})}$ is measured for the entire training data \mathcal{D} using soft counts (Equation (8)) to make the objective differentiable. We train BDF using batch or deterministic gradient descent (GD) with learning step size schedule $\rho_t^{\mathbf{W}^B}$ on the objective (Equation (23)). Pseudo-code to train the BDF model is given in Algorithm 1. In practice, we use *Adam* optimization [60] via backpropagation (BP) [61] and automatic differentiation (autodif) [62] with a fixed step size $\rho^{\mathbf{W}^B}$. The main limitation of the BDF algorithm is that it involves calculating the relevant probabilities for all intersecting protected groups in every iteration, which is exponential in the number of top-level groups. To address this, we present a stochastic method in the following section.

Algorithm 1: Training Algorithm for Batch Differential Fair (BDF) Model**Require:** Train data $\mathcal{D} = (x, y)$ **Require:** Tuning parameter $\lambda > 0$ **Require:** Learning step size schedule $\rho_1^{\mathbf{W}^B}, \rho_2^{\mathbf{W}^B}, \dots$ **Require:** Randomly initialize model parameters \mathbf{W}^B **Output:** BDF model $M_{\mathbf{W}^B}(\mathbf{x})$

- For each burn-in iteration t_b :
 - Apply update using batch GD with $\rho_{t_b}^{\mathbf{W}^B}$:

$$\mathbf{W}^B := \mathbf{W}^B - \rho_{t_b}^{\mathbf{W}^B} \frac{1}{n_{\mathcal{D}}} \nabla_{\mathbf{W}^B} \sum_{i=1}^{n_{\mathcal{D}}} L(f(\mathbf{x}^{(i)}; \mathbf{W}^B), \mathbf{y}^{(i)})$$
- For each iteration t :
 - Compute $\epsilon_{M_{\mathbf{W}^B}(\mathbf{x})} \leq \frac{\sum_{\mathbf{x} \in \mathcal{D}: A=s_j} f(\mathbf{x}; \mathbf{W}^B) + \alpha}{N_{s_j} + |\mathcal{Y}| \alpha} \frac{N_{s_j} + |\mathcal{Y}| \alpha}{\sum_{\mathbf{x} \in \mathcal{D}: A=s_j} f(\mathbf{x}; \mathbf{W}^B) + \alpha} \leq \epsilon_{M_{\mathbf{W}^B}(\mathbf{x})}$
 - Apply update using batch GD with $\rho_t^{\mathbf{W}^B}$:

$$\mathbf{W}^B := \mathbf{W}^B - \rho_t^{\mathbf{W}^B} \frac{1}{n_{\mathcal{D}}} \nabla_{\mathbf{W}^B} [\sum_{i=1}^{n_{\mathcal{D}}} L(f(\mathbf{x}^{(i)}; \mathbf{W}^B), \mathbf{y}^{(i)}) + \lambda R_{\mathbf{W}^B}^{\mathcal{D}}(\epsilon)]$$

// in practice, Adam optimization via BP and autodif with a fixed $\rho^{\mathbf{W}^B}$

9.2. Stochastic Method

The stochastic DF (SDF) method $M_{\mathbf{W}^S}(\mathbf{x})$ uses a minibatch or a small subset of all the training examples $\mathcal{D}_m \in \mathcal{D}$ for each iteration of learning. Pseudo-code to train SDF is given in Algorithm 2. The objective for SDF with minibatch size $n_{\mathcal{D}_m}$ is equivalent to that of the batch method:

$$\min_{\mathbf{W}^S} \left[\frac{1}{n_{\mathcal{D}_m}} \sum_{i=1}^{n_{\mathcal{D}_m}} L(f(\mathbf{x}^{(i)}; \mathbf{W}^S), \mathbf{y}^{(i)}) + \lambda R_{\mathbf{W}^S}^{\mathcal{D}_m}(\epsilon) \right], \quad (24)$$

where $R_{\mathbf{W}^S}^{\mathcal{D}_m}(\epsilon) = \max(0, \epsilon_{M_{\mathbf{W}^S}(\mathbf{x})} - \epsilon_1)$ represents the fairness penalty term for a minibatch. However, the reliable estimation of $\epsilon_{M_{\mathbf{W}^S}(\mathbf{x})}$ for a minibatch becomes statistically challenging due to data sparsity of intersectional groups [56]. For example, one or more missing intersectional groups for a minibatch is a typical scenario in the stochastic models that can lead to inaccurate estimation of intersectional fairness, i.e., ϵ -DF or γ -SF, which affects the training of the models.

We develop a stochastic approximation-based approach to address data sparsity in ϵ -DF estimation for training the SDF model. Our approach is mainly inspired by the online EM algorithm of [63]. It is a general-purpose method for learning latent variable models in an online setting that alternates between a standard M-step that maximizes the EM lower bound with respect to model parameters β and a stochastic expectation step that updates exponential family sufficient statistics Φ with an online average.

$$\Phi := (1 - \rho_t) \Phi + \rho_t \hat{\Phi}(\mathbf{x}^{(n+1)}; \beta), \quad (25)$$

where $\mathbf{x}^{(n+1)}$ is a new data point, ρ_t is the step size at time step t , and $\hat{\Phi}(\mathbf{x}^{(n+1)}; \beta)$ is the estimation of the sufficient statistics based on the latest model parameters β . Inspired by the online EM algorithm, stochastic variational inference was proposed [64,65] to scale models such as topic models up to an enormous number of documents [66–68].

Algorithm 2: Training Algorithm for Stochastic Differential Fair (SDF) Model**Require:** Train data $\mathcal{D} = (x, y)$ **Require:** Tuning parameter $\lambda > 0$ **Require:** Constant learning step size schedule $\rho_1^{N^G}, \rho_2^{N^G}, \dots$ and $\rho_1^{W^S}, \rho_2^{W^S}, \dots$ **Require:** Randomly initialize model parameters \mathbf{W}^S and count parameters $N_{y,s}^G$ and N_s^G **Output:** SDF model $M_{\mathbf{W}^S}(\mathbf{x})$

- For each burn-in epoch:
 - For each iteration t_b :
 - * Draw a minibatch $\mathcal{D}_m \in \mathcal{D}$
 - * Apply update using SGD with $\rho_{t_b}^{W^S}$:

$$\mathbf{W}^S := \mathbf{W}^S - \rho_{t_b}^{W^S} \frac{1}{n_{\mathcal{D}_m}} \nabla_{\mathbf{W}^S} \sum_{i=1}^{n_{\mathcal{D}_m}} L(f(\mathbf{x}^{(i)}; \mathbf{W}^S), \mathbf{y}^{(i)})$$
- For each epoch:
 - For each iteration t_o in the outer loop:
 - * For each iteration t_i in the inner loop:
 - Draw a minibatch $\mathcal{D}_m \in \mathcal{D}$
 - Empirically estimate $\hat{N}_{y,s}^L = \sum_{\mathbf{x} \in \mathcal{D}_m: A=s} f(\mathbf{x}; \mathbf{W}^S)$ and \hat{N}_s^L
 - Apply update: $N_{y,s}^G := (1 - \rho_{t_i}^{N^G}) N_{y,s}^G + \rho_{t_i}^{N^G} \frac{n_{\mathcal{D}}}{n_{\mathcal{D}_m}} \hat{N}_{y,s}^L$
 - Apply update: $N_s^G := (1 - \rho_{t_i}^{N^G}) N_s^G + \rho_{t_i}^{N^G} \frac{n_{\mathcal{D}}}{n_{\mathcal{D}_m}} \hat{N}_s^L$
 - * Compute $\epsilon_{M_{\mathbf{W}^S}(\mathbf{x})} \leq \frac{N_{y,s_i}^G + \alpha}{N_{s_i}^G + |\mathcal{Y}| \alpha} \frac{N_{s_j}^G + |\mathcal{Y}| \alpha}{N_{y,s_j}^G + \alpha} \leq \epsilon_{M_{\mathbf{W}^S}(\mathbf{x})}$
 - * Draw a minibatch $\mathcal{D}_m \in \mathcal{D}$
 - * Apply update using SGD with $\rho_{t_o}^{W^S}$:

$$\mathbf{W}^S := \mathbf{W}^S - \rho_{t_o}^{W^S} \frac{1}{n_{\mathcal{D}_m}} \nabla_{\mathbf{W}^S} [\sum_{i=1}^{n_{\mathcal{D}_m}} L(f(\mathbf{x}^{(i)}; \mathbf{W}^S), \mathbf{y}^{(i)}) + \lambda R_{\mathbf{W}^S}^{\mathcal{D}_m}(\epsilon)]$$

To compute $\epsilon_{M_{\mathbf{W}^S}(\mathbf{x})}$, we iterate over an inner loop t_i for each outer loop t_o , with multiple minibatches drawn from training examples $\mathcal{D}_m \in \mathcal{D}$. First, we empirically estimate noisy expected counts per group $\hat{N}_{y,s}^L$ and \hat{N}_s^L for each minibatch in the inner loop and update global expected counts per group $N_{y,s}^G$ and N_s^G (shared within inner and outer loop iterations) with learning step size schedule $\rho_{t_i}^{N^G}$, which is typically annealed towards zero, as follows:

$$N_{y,s}^G := (1 - \rho_{t_i}^{N^G}) N_{y,s}^G + \rho_{t_i}^{N^G} \frac{n_{\mathcal{D}}}{n_{\mathcal{D}_m}} \hat{N}_{y,s}^L, \quad (26)$$

$$N_s^G := (1 - \rho_{t_i}^{N^G}) N_s^G + \rho_{t_i}^{N^G} \frac{n_{\mathcal{D}}}{n_{\mathcal{D}_m}} \hat{N}_s^L, \quad (27)$$

where $\hat{N}_{y,s}^L = \sum_{\mathbf{x} \in \mathcal{D}_m: A=s} f(\mathbf{x}; \mathbf{W}^S)$ are empirically estimated for each minibatch \mathcal{D}_m with respect to the latest \mathbf{W}^S (from the last outer loop iteration), while \hat{N}_s^L is the total population per group for the corresponding minibatch. Note that it is too time-consuming to run the above inner loop for several minibatch examples. In practice, we found that one inner loop pass with a fixed learning step size ρ^{N^G} is enough to train SDF model successfully, since $N_{y,s}^G$ and N_s^G are continuously updated and shared in the outer loop iterations. Then, $\epsilon_{M_{\mathbf{W}^S}(\mathbf{x})}$ is estimated in the outer loop via the posterior predictive distribution of a Dirichlet-multinomial (similar to Equation (7)) as

$$\epsilon_{M_{\mathbf{W}^S}(\mathbf{x})} \leq \frac{N_{y,s_i}^G + \alpha}{N_{s_i}^G + |\mathcal{Y}| \alpha} \frac{N_{s_j}^G + |\mathcal{Y}| \alpha}{N_{y,s_j}^G + \alpha} \leq \epsilon_{M_{\mathbf{W}^S}(\mathbf{x})}. \quad (28)$$

Finally, the SDF model is trained using stochastic gradient descent (SGD) with step size schedule $\rho_{t_o}^{W^S}$ in the outer loop. In practice, we use *Adam* optimization algorithm on the

objective in Equation (24) via BP and autodif with a fixed step size $\rho^{\mathbf{W}^S}$. We also provide the pseudo-code to our more practical approach for the SDF model in Algorithm 3.

Algorithm 3: Practical Approach for Stochastic Differential Fair (SDF) Model

Require: Train data $\mathcal{D} = (x, y)$

Require: Tuning parameter $\lambda > 0$

Require: Constant learning step size ρ^{N^G} and $\rho^{\mathbf{W}^S}$

Require: Randomly initialize model parameters \mathbf{W}^S and count parameters $N_{y,s}^G$ and N_s^G

Output: SDF model $M_{\mathbf{W}^S}(x)$

- For each burn-in epoch:
 - For each iteration t_b :
 - * Draw a minibatch $\mathcal{D}_m \in \mathcal{D}$
 - * Apply Adam optimizer with $\rho^{\mathbf{W}^S}$ via BP and autodif on the objective:

$$\min_{\mathbf{W}^S} \left[\frac{1}{n_{\mathcal{D}_m}} \sum_{i=1}^{n_{\mathcal{D}_m}} L(f(\mathbf{x}^{(i)}; \mathbf{W}^S), \mathbf{y}^{(i)}) \right]$$
 - For each epoch:
 - For each iteration t :
 - * Draw a minibatch $\mathcal{D}_m \in \mathcal{D}$
 - * Empirically estimate $\hat{N}_{y,s}^L = \sum_{\mathbf{x} \in \mathcal{D}_m: A=s} f(\mathbf{x}; \mathbf{W}^S)$ and \hat{N}_s^L
 - * Apply update: $N_{y,s}^G := (1 - \rho^{N^G})N_{y,s}^G + \rho^{N^G} \frac{n_{\mathcal{D}_m}}{n_{\mathcal{D}}} \hat{N}_{y,s}^L$
 - * Apply update: $N_s^G := (1 - \rho^{N^G})N_s^G + \rho^{N^G} \frac{n_{\mathcal{D}_m}}{n_{\mathcal{D}}} \hat{N}_s^L$
 - * Compute $\epsilon_{M_{\mathbf{W}^S}}(x) \leq \frac{N_{y,s_i}^G + \alpha}{N_{s_i}^G + |\mathcal{Y}| \alpha} \frac{N_{s_j}^G + |\mathcal{Y}| \alpha}{N_{y,s_j}^G + \alpha} \leq \epsilon_{M_{\mathbf{W}^S}}(x)$
 - * Apply Adam optimizer with $\rho^{\mathbf{W}^S}$ via BP and autodif on the objective:

$$\min_{\mathbf{W}^S} \left[\frac{1}{n_{\mathcal{D}_m}} \sum_{i=1}^{n_{\mathcal{D}_m}} L(f(\mathbf{x}^{(i)}; \mathbf{W}^S), \mathbf{y}^{(i)}) + \lambda R_{\mathbf{W}^S}^{\mathcal{D}_m}(\epsilon) \right]$$
-

9.3. Convergence Analysis for Stochastic Method

Our convergence analysis is conducted using the Robbins–Monro stochastic approximation (SA) [69] interpretation of the stochastic ϵ -DF estimation in the SDF algorithm. We have the following theorem:

Theorem 5. *If $0 < \rho_t \leq 1 \forall t$, $\sum_{t=1}^{\infty} \rho_t = \infty$, and $\lim_{t \rightarrow \infty} \rho_t = 0$, then in the limit as the number of steps t approaches infinity stochastic ϵ -DF estimation converges to a stationary point of the objective function for the SDF model.*

Proof. Consider the inner loop for stochastic ϵ -DF estimation in the SDF algorithm, with an update schedule for t steps. Noisy expected counts $\hat{N}_{y,s}^L$ and \hat{N}_s^L are empirically estimated with respect to the SDF parameters \mathbf{W}^S and drawn minibatch $\mathcal{D}_m \in \mathcal{D}$ at the iteration step t of the inner loop. Given \mathbf{W}^S , we do not need to maintain $\hat{N}_{y,s}^L$ and \hat{N}_s^L between inner loop iterations. Thus, stochastic ϵ -DF estimation is operated only on the global expected count parameters $N_{y,s}^G$ and N_s^G . For each count parameter $c \in \{N_{y,s}^G, N_s^G\}$, let $f_c(\mathcal{D}, \mathbf{W}^S, \phi) : \Phi_c \rightarrow \Phi_c$ be a mapping from a current value to the updated value after an iteration t , where Φ_c is the space of possible assignments for c .

Let $\bar{\phi} = (N_{y,s}^G, N_s^G)$ be an assignment of the expected count parameters, with $\bar{\phi}_c$ referring to a parameter c , and $\bar{\phi}^{(t)}$ be the c parameters at step t . Furthermore, let $\hat{\phi}_c(\mathcal{D}_m^{(t+1)}, \mathbf{W}^S)$ be the noisy estimate of $f_c(\mathcal{D}, \mathbf{W}^S, \phi)$ based on the minibatch $\mathcal{D}_m^{(t+1)}$ examined at step $t + 1$. Finally, let $\xi^{(t+1)} = \hat{\phi}_c(\mathcal{D}_m^{(t+1)}, \mathbf{W}^S) - f_c(\mathcal{D}, \mathbf{W}^S, \phi^{(t)})$ be the stochastic error made at step $t + 1$, and observe that $E[\xi^{(t+1)}] = 0$. We can rewrite the update for each c parameter as

$$\begin{aligned}
\bar{\phi}_c^{(t+1)} &= (1 - \rho_{t+1}^c) \bar{\phi}_c^{(t)} + \rho_{t+1}^c \hat{\phi}_c(\mathcal{D}_m^{(t+1)}, \mathbf{W}^S) \\
&= \bar{\phi}_c^{(t)} + \rho_{t+1}^c (-\bar{\phi}_c^{(t)} + \hat{\phi}_c(\mathcal{D}_m^{(t+1)}, \mathbf{W}^S)) \\
&= \bar{\phi}_c^{(t)} + \rho_{t+1}^c (f_c(\mathcal{D}, \mathbf{W}^S, \bar{\phi}_c^{(t)}) - \bar{\phi}_c^{(t)} + \hat{\phi}_c(\mathcal{D}_m^{(t+1)}, \mathbf{W}^S) - f_c(\mathcal{D}, \mathbf{W}^S, \bar{\phi}_c^{(t)})) \\
&= \bar{\phi}_c^{(t)} + \rho_{t+1}^c (f_c(\mathcal{D}, \mathbf{W}^S, \bar{\phi}_c^{(t)}) - \bar{\phi}_c^{(t)} + \zeta^{(t+1)}). \tag{29}
\end{aligned}$$

In this form, we can see that iterating each of the c parameter updates over the inner loop to ϵ -DF estimation corresponds to the Robbins–Monro SA algorithm [69] for finding the zeros of $f_c(\mathcal{D}, \mathbf{W}^S, \phi^{(t)}) - \phi_c^{(t)}$, i.e., the fixed points for ϕ_c that lead to the stationary point for $\epsilon_{M_{\mathbf{W}^S}(\mathbf{x})}$ estimation in the SDF algorithm. Theorem 2.3 of [70] states that under the existence of a Lyapunov function along with a boundedness condition, this implies that such an SA algorithm will converge with step size schedules such as mentioned above. Note that a Lyapunov function can be viewed as an objective function in the absence of stochastic noise and the SA algorithm would improve the function monotonically with small enough steps in the direction of the updates. \square

10. Experiments and Results

In this section, we performed an extensive experimental analysis of the proposed fairness metric in terms of fair learning algorithms and intersectionality. Our implementation's source code is provided in the GitHub (https://github.com/rashid-islam/Differential_Fairness accessed on 10 April 2023).

10.1. Datasets

We performed all experiments with the following four small and large-scale datasets. *We recognize that using risk assessment datasets such as COMPAS or HMDA for fairness research is contested due to the biases inherent in criminal justice and financial systems [71]. Our results should be taken as an illustration of how our definition is operationally different rather than as an endorsement of using automated decision making for these applications.*

- **COMPAS:** The COMPAS dataset regarding a system that is used to predict criminal recidivism, and which has been criticized as potentially biased [4]. We used *race* and *gender* as protected attributes. Gender was coded as binary. Race originally had six values, but we merged “Asian” and “Native American” with “other,” as all three contained very few instances. We predict “actual recidivism,” which is binary, within a 2-year period for 7.22 K individuals.
- **Adult:** The Adult 1994 U.S. census income data from the UCI ML-repository [72] consists of 14 attributes regarding work, relationships, and demographics for individuals, who are labeled according to whether their income exceeds USD 50,000/year (Predicted income used for consequential decisions such as housing approval may result in *digital redlining* [1].) pre-split into a training set of 32.56 K and a test set of 16.28 K instances. We considered *race*, *gender*, and *nationality* as the protected attributes. Gender was coded as binary. As most instances have U.S. nationality, we treat nationality as binary also between U.S. and “other.” The race attribute originally had five values, but we merged the “Native American” with “other,” as both contained very few instances.
- **HHP:** This is a medium-sized dataset which is derived from the Heritage Health Prize (HHP) milestone 1 challenge (<https://www.kaggle.com/c/hhp> accessed on 10 April 2023). The dataset contains information for 171.07 K patients over a 3-year period. Our goal is to predict whether the *Charlson Index*, an estimation of patient mortality, is greater than zero. Following Song et al. [73], we consider *age* and *gender* as the protected attributes, where there are nine possible *age* values and two possible *gender* values.

- HMDA:** This is the largest dataset in our study, which is derived from the loan application register form for the Home Mortgage Disclosure Act (HMDA) [74]. The HMDA is a federal act approved in 1975 which requires mortgage lenders to keep records of information regarding their lending practices to create greater transparency and borrower protections in the residential mortgage market (<https://www.ffiec.gov/hmda/> accessed on 10 April 2023). The downstream task is to predict whether the loan application is approved. We used *ethnicity*, *race*, and *gender* as protected attributes, where ethnicity and gender were coded as binary. The race attribute originally had five values such as “American-Indian/Alaska-Native,” “Asian,” “Black/African-American,” “Native-Hawaiian or Other Pacific-Islander,” and “White.” we merged “American-Indian/Alaska-Native,” “Asian,” and “Native-Hawaiian or Other Pacific-Islander” to the same category as “other” since all three contained very few instances. Note that we filtered out individuals who did not declare information, i.e., incomplete or missing, related to the protected and/or other attributes. After pre-processing, the HMDA data consist of 30 attributes regarding loan type, property type, loan purpose, etc., for 2.33 M individuals.

10.2. Fair Learning Algorithms

The goals of our experiments were to demonstrate the practicality of our *batch differential fair (BDF)* and *stochastic differential fair (SDF)* classifiers in learning intersectionally fair classifiers and to compare their behavior to baseline methods, especially with regards to minorities. Instead of Kearns et al. [17]’s algorithm, we trained subgroup fair classifiers as batch (BSF) and stochastic (SSF) methods using the same approach in Algorithms 1 and 3, respectively, replacing ϵ with γ in Equation (23), i.e., $R_{\mathbf{W}^B}^D(\gamma) = \max(0, \gamma_{M_{\mathbf{W}^B}(\mathbf{x})} - \gamma_1)$, and Equation (24), i.e., $R_{\mathbf{W}^S}^D(\gamma) = \max(0, \gamma_{M_{\mathbf{W}^S}(\mathbf{x})} - \gamma_1)$, respectively. This simplifies and speeds up learning to handle deep neural networks used in our experiments. We also compare the models with a demographic parity-based *p%-Rule* [75] classifier baseline, which generalizes the 80% rule to measure disparate impact toward one protected group. Since demographic parity, by definition, assumes binary protected attributes, we select the most marginalized bottom-level group as the protected group (*black women non-USA* for Adult, *black women* for COMPAS, *women with age ≥ 85* for HHP, and *black hispanic women* for HMDA), compared to its complement. A logistic regression or SVM-based algorithm subject to *p%-Rule* constraint was proposed by Zafar et al. [75], but the performance of their algorithm was very poor compared to our deep learning-based approach. Therefore, we implemented *batch p%-Rule classifier (BPR)* and *stochastic p%-Rule classifier (SPR)* using our DNN-based Algorithms 1 and 3, respectively. Since higher is better for *p%-Rule*, we modified Equation (23) as $R_{\mathbf{W}^B}^D(p\%) = \max(0, p\%_1 - p\%_{M_{\mathbf{W}^B}(\mathbf{x})})$, and Equation (24) as $R_{\mathbf{W}^S}^D(p\%) = \max(0, p\%_1 - p\%_{M_{\mathbf{W}^S}(\mathbf{x})})$ for BPR and SPR, respectively. We also considered deep learning-based typical classifiers in batch (BT) and stochastic (ST) methods that do not incorporate any fairness interventions with their objective functions. The summary of learning algorithms is given in Table 2.

Table 2. Summary of learning algorithms.

Abbreviation	Definition	Description
BT	Batch Typical (vanilla) model	Train deep learning classifier with batch method without any fairness interventions
BPR	Batch <i>p%-Rule</i> model	Train deep learning classifier with batch method that enforces <i>p%-Rule</i> criteria
BSF	Batch Subgroup Fair model	Train deep learning classifier with batch method that enforces γ -SF criteria
BDF	Batch Differential Fair model	Train deep learning classifier with batch method that enforces proposed ϵ -DF criteria
ST	Stochastic Typical (vanilla) model	Train deep learning classifier with stochastic method without any fairness interventions
SPR	Stochastic <i>p%-Rule</i> model	Train deep learning classifier with stochastic method that enforces <i>p%-Rule</i> criteria
SSF	Stochastic Subgroup Fair model	Train deep learning classifier with stochastic method that enforces γ -SF criteria
SDF	Stochastic Differential Fair model	Train deep learning classifier with stochastic method that enforces proposed ϵ -DF criteria

10.2.1. Experimental Settings

All classifiers were trained on a common neural network architecture via adaptive gradient descent optimization (Adam) with learning rate = 0.001 using PyTorch [76]. The configuration of the network was 3 hidden layers, 16 neurons in each layer with drop out probability 0.5, “relu” and “sigmoid” activations for the hidden and output layers, respectively. For all batch fair methods, we trained for 500 iterations, disabling the fairness penalties for an additional 50 “burn-in” iterations. For all stochastic fair methods, we trained for 5 epochs with an additional “burn-in” epoch for the large-scale HMDA dataset, while 100 epochs + 5 “burn-in” epochs were used for the other datasets. Note that we did not use any “burn-in” period for the typical models (both BT and ST), since a warm start is redundant here due to the absence of fairness interventions.

We split the COMPAS dataset into 60% train, 20% development, and 20% test set. For Adult, we used the pre-specified test set and held out 30% from the training data as the development set. We held out 10% and 5% from the HHP and HMDA datasets, respectively, as the test set, using the remainder for training. We further held out 10% and 5% from the HHP and HMDA training sets, respectively, as the development set for each dataset. Since HMDA is a very large dataset, it was not feasible in terms of runtime and memory to train the models in batch setting for the entire HMDA training set. Therefore, we trained the batch models on HMDA by randomly taking a subset with 10% of the HMDA training set.

We learned fair classifiers in several settings; we set the target thresholds (1) to perfect fairness in terms of the corresponding fairness metric, $\epsilon_1 = 0.0$, $\gamma_1 = 0.0$, $p\%_1 = 100.0\%$ for DF-Classifiers (BDF and SDF), SF-Classifiers (BSF and SSF), and $p\%$ -Rule classifiers (BPR and SPR), respectively, and (2) to penalize bias amplification by the algorithm, by setting the thresholds to $\epsilon_1 = \epsilon_{data}$ and $\gamma_1 = \gamma_{data}$ for DF-Classifiers and SF-Classifiers, respectively. Finally, to protect the 80% rule, we set $\epsilon_1 = -\log 0.8 = 0.2231$ for DF-Classifiers. Since there is no straightforward way to enforce the 80% rule for SF-Classifiers, it was not considered in this analysis. To evaluate the predictive performance of the classifiers, we compute Accuracy, F1 score, and ROC AUC for the held-out data. Finally, we compute ϵ -DF, γ -SF, $p\%$ -Rule, bias amplification on held-out data (using Equation (8)) to evaluate the classifiers in terms of fairness.

10.2.2. Fairness and Accuracy Trade-Off

Fair learning algorithms divert the learning objective from accuracy only to both accuracy and fairness, which may hurt the predictive performance of the models. The tuning parameter λ allows the stakeholders to balance between fairness and accuracy. We chose λ for all fair learning algorithms via rigorous grid search on the development set based on a pre-defined rule: *select λ that provides the fairest (under the corresponding fairness metric, e.g., ϵ for DF-Classifiers, γ for SF-Classifiers) model on the development set, allowing up to 5% degradation in accuracy from the typical classifier*. Since fairness is a multi-stakeholder issue, the amount of slack tolerance on the accuracy can be amended based on the stakeholders' preferences when deploying these methods in practice.

In Figure 8, we show the impact of the tuning parameter λ on the accuracy and corresponding fairness metric for the SDF and SSF classifiers on the development set of Adult and HHP datasets. *Black circles* in the plots represent the SDF or SSF model corresponding to different λ values (larger to smaller from left to right), while *red square* is the stochastic typical classifier (ST) (no fairness interventions in the objective). Larger λ incorporated with the corresponding fairness penalty for the SDF or SSF model allows us to achieve more fairness, but with greater loss in predictive performance, while smaller λ has the opposite impact on the models' output. So, λ needs to be chosen as a trade-off between accuracy and fairness. *Parallel dotted lines* represent the valid region of 5% degraded accuracy from the ST model, and *green asterisk* is the fairest SDF or SSF model, in terms of their corresponding fairness metric, within the valid region for a particular λ value. According to our pre-defined rule, these λ values (*green arrow*) are then selected for

SDF and SSF models. The trade-off between fairness and accuracy was similar for other datasets and is provided in the Appendix A (see Figure A1).

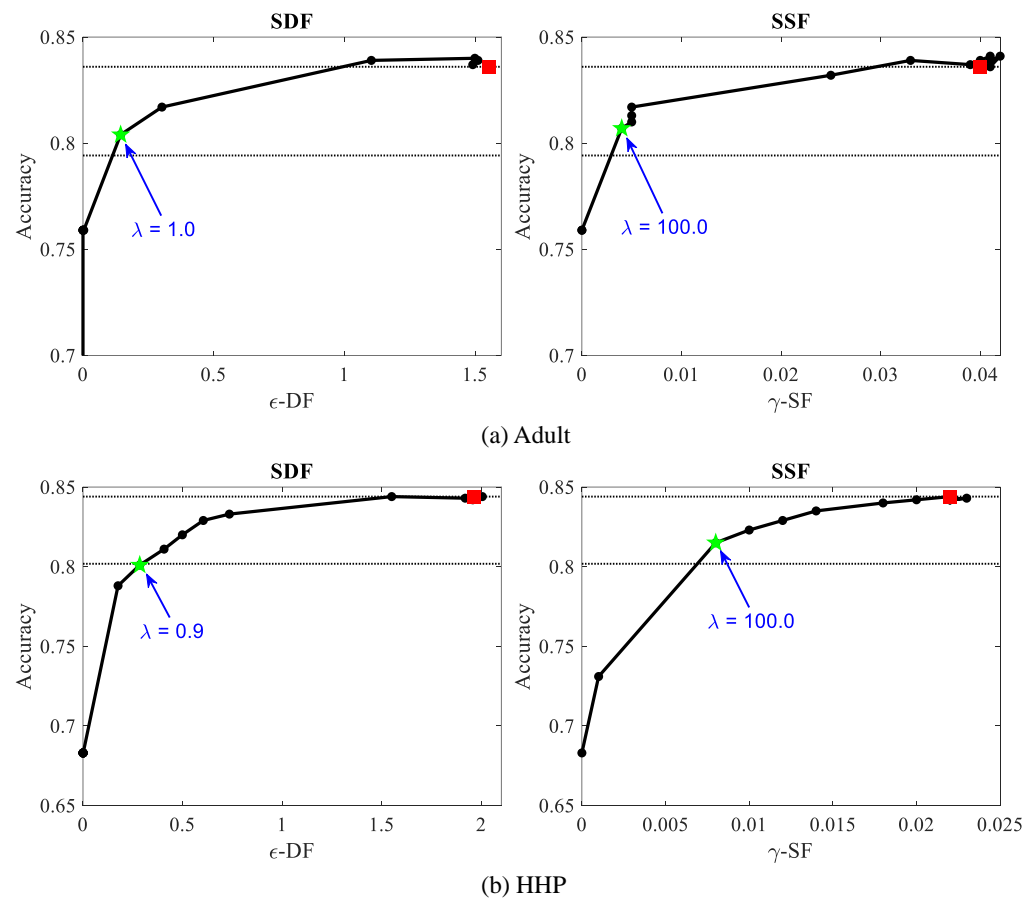


Figure 8. Fairness and accuracy trade-off plots for stochastic DF (SDF) and SF (SSF) models on the development set of (a) Adult and (b) HHP datasets. *Red square*: stochastic typical (ST) classifier that does not incorporate fairness penalty in the objective; *dotted lines*: indicate the valid region of 5% degraded accuracy from ST; *black circles*: fair models (SDF or SSF) correspond to different λ values (larger to smaller from left to right); *green asterisk*: fairest model within the valid region for a particular λ value.

10.2.3. Comparative Analysis for Batch and Stochastic Methods

In this experiment, we compare the accuracy and fairness for the batch and stochastic fair models (DF-Classifiers and SF-Classifiers) in terms of the training algorithm's runtime. Stochastic fair models are scalable to very large data due to their minibatch setting. For example, we trained the stochastic models on the entire HMDA training data (largest dataset in our study) without any additional complexity, while the batch models were trained on a small subset of the HMDA training data due to extremely high memory requirements and runtime. The purpose of this analysis is to show that the noisy updates used to train stochastic fair models do not introduce any additional harm to the accuracy or fairness of the models as we estimate the fairness penalty using noisy updates of the empirical counts (see Equations (26) and (27)).

Figure 9 shows ϵ -DF and accuracy in terms of runtime for the BDF and SDF models on the development set of Adult and HHP datasets. The *red X* indicates the end of burn-in iterations and burn-in epochs for the BDF and SDF models, respectively. Recall that, in the case of these two datasets, we disabled the fairness penalty for 50 burn-in iterations and 5 burn-in epochs in the training process of BDF and SDF models, respectively. In the SDF model, the burn-in epochs were enough to obtain a model that shows similar performance to the typical model (high accuracy and high ϵ -DF), while the model started declining in

accuracy and improving ϵ -DF (lower is better) after burn-in due to the fairness interventions in the objective. Due to the fewer burn-in iterations in the training of the BDF model, we found that the model performed with relatively lower accuracy and less extreme ϵ -DF, since the predicted probability of the outcomes across protected groups were approximately similar. After burn-in, the BDF model improved the predictive performance and also maintained a low ϵ -DF from imposing the fairness penalty in the objective. Note that λ was pre-selected based on our pre-defined rule, as discussed in Section 10.2.2. Since both the BDF and SDF models were constrained to comply with the pre-defined rule for λ selection, after a certain period of runtime, both models converged to approximately similar solutions for both datasets (Adult in Figure 9a and HHP in Figure 9b) in terms of ϵ -DF and accuracy. We show a similar trend in Figure 10 for BSF and SSF models by comparing the γ -SF and accuracy with respect to the runtime. We conclude that stochastic models do not cause any additional harm to the model's fairness or predictive performance comparing to the batch methods. See the appendix for similar results with the other datasets (Figures A2 and A3).

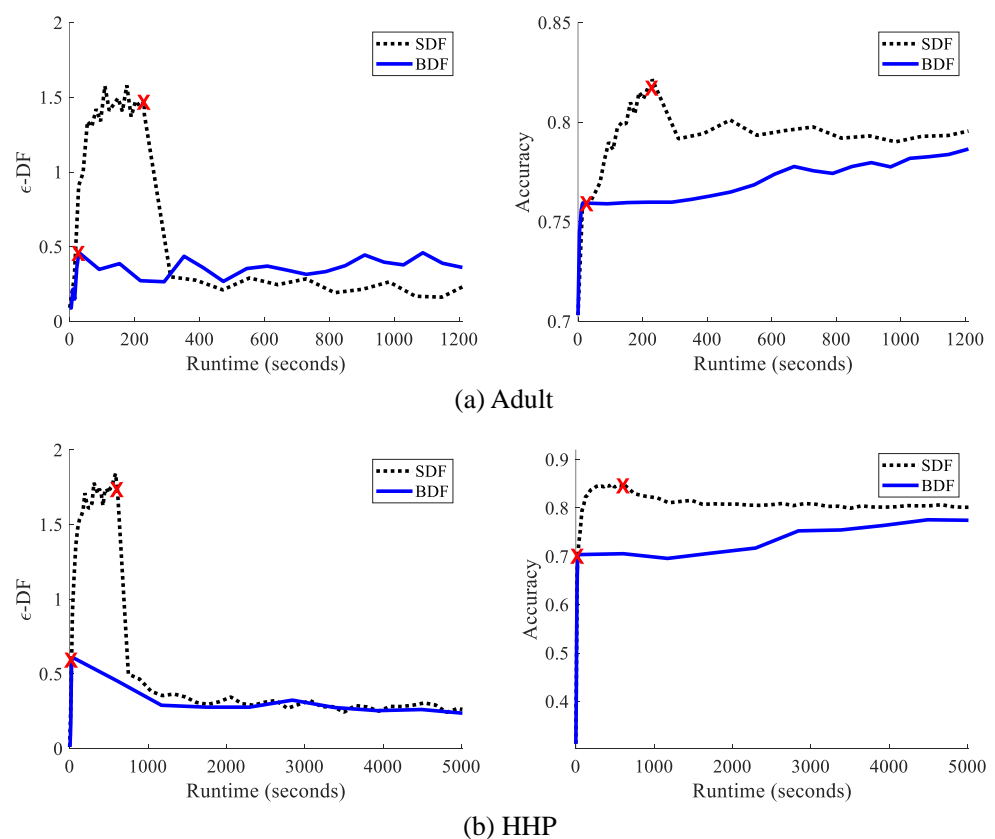


Figure 9. Comparison of ϵ -DF and accuracy in terms of runtime for the batch differential fair (BDF) and stochastic differential fair (SDF) models on the development set of (a) Adult and (b) HHP datasets. Red X: end of burn-in (disabling the fairness interventions) iterations and epochs for the BDF and SDF models, respectively.

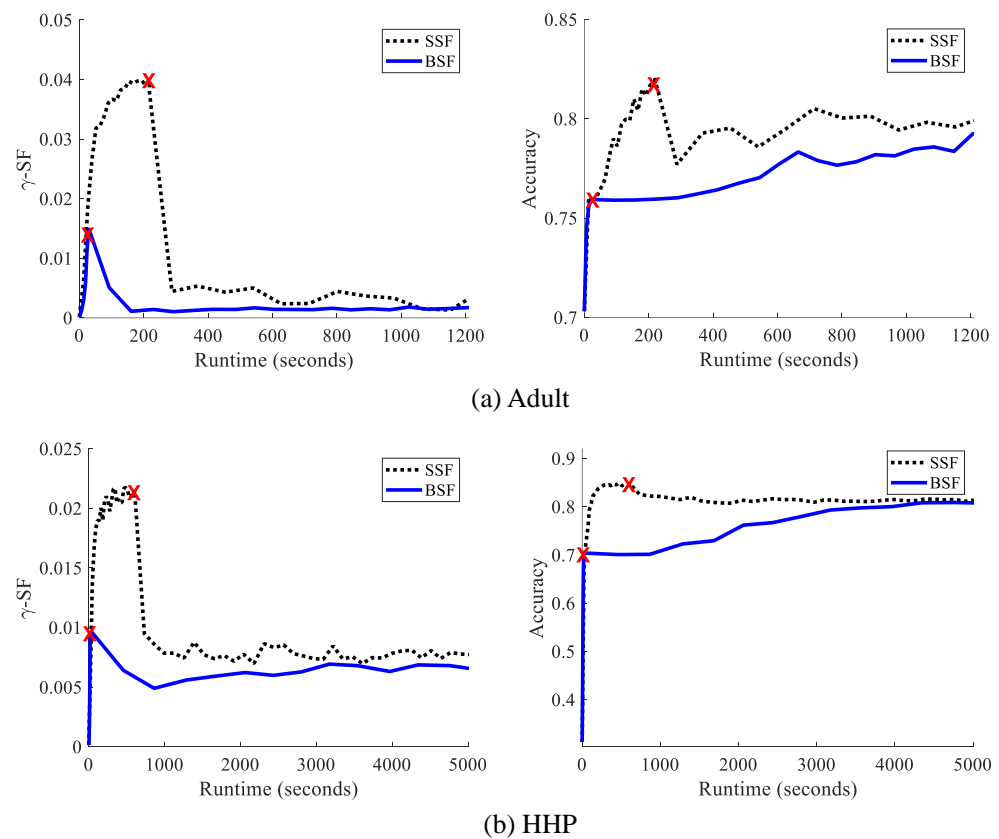


Figure 10. Comparison of γ -SF and accuracy in terms of runtime for the batch subgroup fair (BSF) and stochastic subgroup fair (SSF) models on the development set of (a) Adult and (b) HHP datasets. Red X: end of burn-in (disabling the fairness interventions) iterations and epochs for the BSF and SSF models, respectively.

10.2.4. Performance for Fair Classifiers

In this experiment, we evaluate the accuracy-based performance and fairness metrics for batch and stochastic classifiers on the unseen test set of all datasets. Table 3 compares the batch methods on the Adult and COMPAS datasets. Both the BDF and BSF classifiers were able to **substantially improve their fairness metrics over the BPR and BT classifiers, 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 . The BDF model with $\epsilon_1 = 0$ improved from $\epsilon = 0.743$ to $\epsilon = 0.205$ on COMPAS data without any loss in accuracy, while also improving all other fairness metrics substantially. Surprisingly, BDF (and also the BPR classifier) slightly improved accuracy on this data compared to BT. This counter-intuitive result is presumably due to the regularization behavior of the fairness penalty on the objective, which can sometimes lead fair models to reduce overfitting to some degree compared to the typical model, a phenomenon which we observed in our previous work on the equitable allocation of healthcare resources [12]. On Adult, it improved from $\epsilon = 1.591$ to $\epsilon = 0.305$, corresponding to a worst-case difference in utility between groups of a factor of $e^\epsilon \approx 1.4$ (Equation (19)), with a loss of 3.6 percentage points in accuracy. In terms of γ and $p\%$ -Rule, the BSF model with $\gamma_1 = 0$ showed the highest improvement on the Adult data.

Table 3. Comparison of batch differential fair (BDF), batch subgroup fair (BSF), batch $p\%$ -Rule (BPR), and batch typical (BT) classifiers on the test set of COMPAS and Adult datasets ($\epsilon_1 = 0.2231$ is the 80% rule). Higher is better for measures with \uparrow ; lower is better for measures with \downarrow .

COMPAS Dataset								
Models		$\epsilon_1 = 0.0$	BDF-Classifier $\epsilon_1 = 0.2231$	$\epsilon_1 = \epsilon_{data}$	BSF-Classifier $\gamma_1 = 0.0$	$\gamma_1 = \gamma_{data}$	BPR-Classifier	BT-Classifier
Tuning Parameter	λ	0.800	5.000	50.000	25.000	100.000	0.010	X
Performance Measures	Accuracy \uparrow	0.672	0.674	0.665	0.660	0.667	0.674	0.671
	F1 Score \uparrow	0.560	0.580	0.580	0.560	0.580	0.620	0.630
	ROC AUC \uparrow	0.700	0.698	0.687	0.701	0.701	0.700	0.686
Fairness Measures	ϵ -DF \downarrow	0.205	0.252	0.308	0.249	0.303	0.630	0.743
	γ -SF \downarrow	0.003	0.016	0.019	0.003	0.012	0.035	0.034
	$p\%$ -Rule \uparrow	66.027	65.278	64.276	63.853	62.296	63.096	60.062
	Bias Amp-DF \downarrow	-0.333	-0.286	-0.230	-0.289	-0.235	0.092	0.205
	Bias Amp-SF \downarrow	-0.017	-0.004	-0.001	-0.017	-0.008	0.015	0.014
Adult Dataset								
Models		$\epsilon_1 = 0.0$	BDF-Classifier $\epsilon_1 = 0.2231$	$\epsilon_1 = \epsilon_{data}$	BSF-Classifier $\gamma_1 = 0.0$	$\gamma_1 = \gamma_{data}$	BPR-Classifier	BT-Classifier
Tuning Parameter	λ	0.100	0.100	10.000	10.000	25.000	0.001	X
Performance Measures	Accuracy \uparrow	0.795	0.802	0.802	0.798	0.818	0.827	0.831
	F1 Score \uparrow	0.380	0.440	0.420	0.420	0.500	0.580	0.600
	ROC AUC \uparrow	0.836	0.847	0.860	0.835	0.876	0.876	0.883
Fairness Measures	ϵ -DF \downarrow	0.305	0.442	1.194	0.733	1.355	1.362	1.591
	γ -SF \downarrow	0.007	0.011	0.031	0.001	0.030	0.033	0.042
	$p\%$ -Rule \uparrow	55.590	54.855	46.988	58.112	47.128	54.071	45.019
	Bias Amp-DF \downarrow	-1.075	-0.938	-0.186	-0.647	-0.025	-0.018	0.211
	Bias Amp-SF \downarrow	-0.026	-0.022	-0.002	-0.032	-0.003	0.000	0.009

We found a similar trend in performance for stochastic methods. Table 4 compares the accuracy and fairness measures for the stochastic methods on the HHP and HMDA datasets. Similar to batch methods, SDF and SSF substantially improve fairness over the SPR and ST classifiers with modest costs in accuracy-based measures. On the HHP dataset, the SDF method with $\epsilon_1 = 0$ improved from $\epsilon = 1.791$ to $\epsilon = 0.21$, corresponding to a worst-case difference in utility between groups of a factor of $e^\epsilon \approx 1.2$, with an accuracy loss of 4.2 percentage points. SDF also showed the best improvement on this dataset in terms of γ (improved from $\gamma = 0.022$ to $\gamma = 0.004$). In the case of HMDA, it improved from $\epsilon = 0.673$ to $\epsilon = 0.24$ with a loss of 4.4 percentage points in accuracy. When BDF and SDF were trained to protect the 80% rule (i.e., $\epsilon_1 = 0.2231$) or to prevent bias amplification (i.e., $\epsilon_1 = \epsilon_{data}$), the fairness metrics were improved with relatively little reduction in accuracy.

Our overall conclusions from Tables 3 and 4 are as follows: the DF-Classifiers (BDF and SDF) typically had comparable accuracy under the same settings compared to the SF-Classifiers (BSF and SSF) and $p\%$ -Rule Classifiers (BPR and SPR), while the **DF-Classifiers often greatly improved γ -SF and $p\%$ -Rule, but the SF- and $p\%$ -Rule Classifiers had modest improvements, or even increases, in ϵ -DF.**

An important goal of this work was to address fairness for minority groups. In Figures 11 and 12, we report the “per-group fairness”, defined as Equation (3) (γ -SF) and Equation (2) (ϵ -DF) with the group held fixed, versus the group’s probability (i.e., size) for batch and stochastic models, respectively. Both the DF-Classifiers and SF-Classifiers generally improved their corresponding unfairness per group over the Typical-Classifiers (BT or ST). On the other hand, similar to Figure 4, the γ -SF metric assigns the highest per-group γ -SF values, which determines overall fairness, to large groups, so **minority groups were not able to influence the overall γ -SF**. For example, overall γ -SF is determined by a top-level group (red “X”) for BSF on both Adult and COMPAS datasets and SSF on both HHP and HMDA datasets. This was not the case for the **ϵ -DF metric, where groups of various sizes had similarly high per-group ϵ -DF values**. For example, the overall ϵ -DF is

determined by an intersectional subgroup (blue “X”) regardless of the subgroup’s size for both the BDF and SDF models on all the example datasets. Furthermore, the **DF-Classifiers improved the per-group fairness under both metrics for groups of all sizes, while the SF-classifiers made per-group γ -SF and ϵ -DF worse for very small groups**. Our overall conclusion is that the *DF-Classifiers are able to achieve intersectionally fair classification with minor loss in performance, while providing greater protection to minority groups than when enforcing SF*. Similar experimental results on the other datasets are provided in Appendix A (see Tables A1 and A2 and Figures A4 and A5).

Table 4. Comparison of stochastic differential fair (SDF), stochastic subgroup fair (SSF), stochastic p %-Rule (SPR), and stochastic typical (ST) classifiers on the test set of HHP and HMDA datasets ($\epsilon_1 = 0.2231$ is the 80% rule). Higher is better for measures with \uparrow ; lower is better for measures with \downarrow .

HHP Dataset								
Models		SDF-Classifer			SSF-Classifer		SPR-Classifer	ST-Classifer
		$\epsilon_1 = 0.0$	$\epsilon_1 = 0.2231$	$\epsilon_1 = \epsilon_{data}$	$\gamma_1 = 0.0$	$\gamma_1 = \gamma_{data}$		
Tuning Parameter	λ	0.900	0.900	1000.000	100.000	500.000	0.100	X
Performance Measures	Accuracy \uparrow	0.801	0.801	0.832	0.813	0.829	0.818	0.843
	F1 Score \uparrow	0.580	0.580	0.680	0.620	0.660	0.650	0.720
	ROC AUC \uparrow	0.796	0.803	0.859	0.870	0.887	0.873	0.895
Fairness Measures	ϵ -DF \downarrow	0.210	0.229	0.762	1.164	1.444	1.688	1.791
	γ -SF \downarrow	0.004	0.005	0.013	0.007	0.010	0.022	0.022
	p %-Rule \uparrow	61.507	61.292	56.183	53.678	51.169	61.720	49.860
	Bias Amp-DF \downarrow	−1.250	−1.231	−0.698	−0.296	−0.016	0.228	0.331
	Bias Amp-SF \downarrow	−0.009	−0.008	0.000	−0.006	−0.003	0.009	0.009
HMDA Dataset								
Models		SDF-Classifer			SSF-Classifer		SPR-Classifer	ST-Classifer
		$\epsilon_1 = 0.0$	$\epsilon_1 = 0.2231$	$\epsilon_1 = \epsilon_{data}$	$\gamma_1 = 0.0$	$\gamma_1 = \gamma_{data}$		
Tuning Parameter	λ	0.600	0.600	1.000	50.000	1000.000	0.010	X
Performance Measures	Accuracy \uparrow	0.819	0.821	0.848	0.851	0.858	0.858	0.863
	F1 Score \uparrow	0.890	0.890	0.900	0.900	0.910	0.910	0.910
	ROC AUC \uparrow	0.903	0.903	0.917	0.921	0.927	0.926	0.931
Fairness Measures	ϵ -DF \downarrow	0.240	0.296	0.417	0.436	0.549	0.554	0.673
	γ -SF \downarrow	0.004	0.005	0.007	0.002	0.005	0.008	0.011
	p %-Rule \uparrow	71.018	69.938	66.822	70.260	64.643	71.256	62.045
	Bias Amp-DF \downarrow	−0.421	−0.365	−0.244	−0.225	−0.112	−0.107	0.012
	Bias Amp-SF \downarrow	−0.005	−0.004	−0.002	−0.007	−0.004	−0.001	0.002

10.3. Inequity of Fairness Measures

We have seen that the γ -SF metric downweights the consideration of minorities (cf. Figures 4, 11 and 12). In this experiment, we quantify the resulting inequity of fairness consideration using the *Gini coefficient* [77], a commonly used measure of statistical dispersion which is often used to represent the inequity of income. We calculate the *Gini coefficient* G of a fairness metric F as

$$G = \frac{1}{2\mu} \sum_{i=1}^n \sum_{j=1}^n P(\mathbf{s}_i)P(\mathbf{s}_j)|F_{\mathbf{s}_i} - F_{\mathbf{s}_j}|, \quad (30)$$

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.

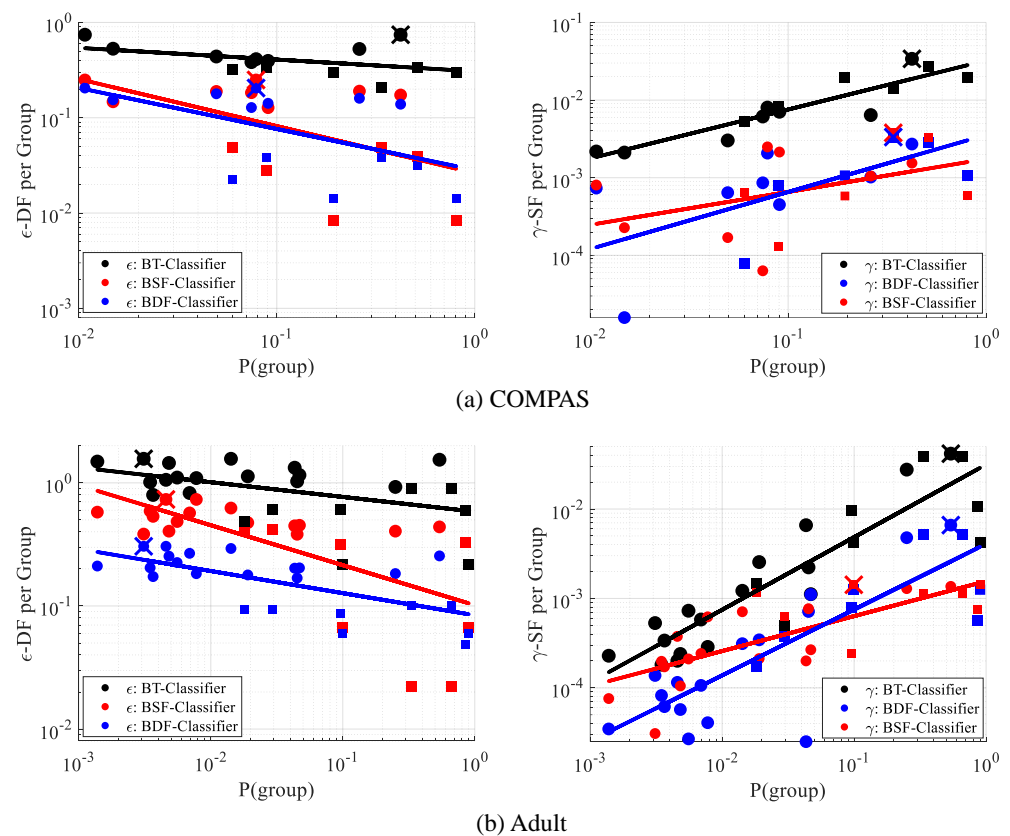


Figure 11. Per-group measurements of ϵ -DF and γ -SF for the batch differential fair (BDF), batch subgroup fair (BSF), batch $p\%$ -Rule (BPR), and batch typical (BT) classifiers vs. group size (probability), (a) COMPAS and (b) Adult datasets, calculated using Equation (3) (γ -SF) and Equation (2) (ϵ -DF) with the group held fixed. Circles: intersectional subgroups. Squares: top-level groups. Fitted least squares line demonstrates the overall trend between fairness measures and group size. Largest per-group ϵ -DF and γ -SF, indicated using “X”, determines the overall ϵ -DF and γ -SF, respectively.

Table 5 shows a comparison of G values for the ϵ -DF and γ -SF metrics on all 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.

Table 5. Comparison of the inequity in the per-group allocation of the ϵ -DF and γ -SF metrics via the Gini coefficient.

Dataset	$G(\epsilon_{Data})$	Gini Coefficient (G)		
		$G(\gamma_{Data})$	$G(\epsilon_{LR})$	$G(\gamma_{LR})$
COMPAS	0.151	0.376	0.135	0.343
Adult	0.099	0.256	0.126	0.257
HHP	0.113	0.311	0.105	0.305
HMDA	0.073	0.423	0.094	0.358

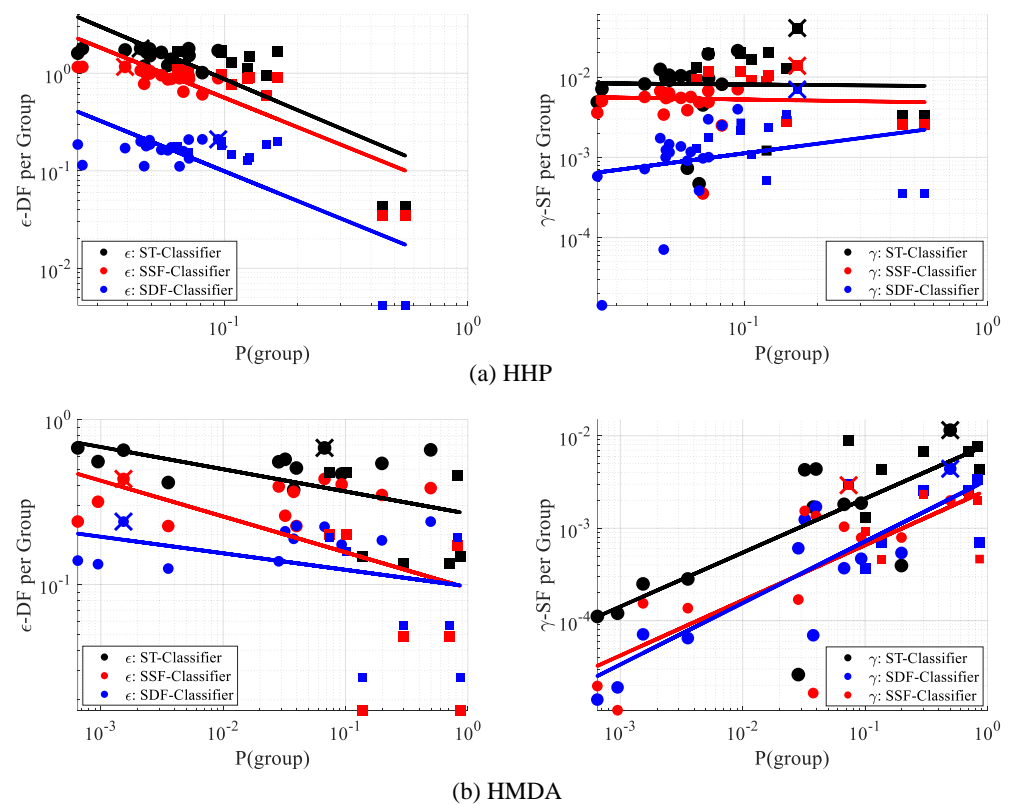


Figure 12. Per-group measurements of ϵ -DF and γ -SF for the stochastic differential fair (SDF), stochastic subgroup fair (SSF), stochastic $p\%$ -Rule (SPR), and stochastic typical (ST) classifiers vs. group size (probability), (a) HHP and (b) HMDA datasets, calculated using Equation (3) (γ -SF) and Equation (2) (ϵ -DF) with the group held fixed. Circles: intersectional subgroups. Squares: top-level groups. Fitted least squares line demonstrates the overall trend between fairness measures and group size. Largest per-group ϵ -DF and γ -SF, indicated using “X”, determines the overall ϵ -DF and γ -SF, respectively.

10.4. Evaluation of Intersectionality Property

In our final experiment (Table 6), we studied the ability of γ -SF to preserve the intersectionality property shown for ϵ -DF in Theorem 2, by measuring fairness with different sets of protected attributes on all datasets.

Table 6. Protection of intersectionality property (Theorem 2). The cases in red are where γ -SF violates the intersectionality property enjoyed by ϵ -DF. Note that ϵ -DF and γ -SF are in different scales, so their values are not comparable. The purpose of this experiment is not to compare the values for ϵ -DF and γ -SF, but rather to empirically verify whether they respect the intersectionality property.

COMPAS Dataset			HHP Dataset		
Protected Attributes	ϵ -DF	γ -SF	Protected Attributes	ϵ -DF	γ -SF
race	0.1003	0.0070	gender	0.0505	0.0039
gender	0.9255	0.0656	age	2.0724	0.0469
race, gender	1.3156	0.0604	gender, age	2.2505	0.0241
Adult Dataset			HMDA Dataset		
Protected Attributes	ϵ -DF	γ -SF	Protected Attributes	ϵ -DF	γ -SF
nationality	0.2177	0.0045	ethnicity	0.1846	0.0056
race	0.9188	0.0128	race	0.6067	0.0126
gender	1.0266	0.0434	gender	0.1702	0.0088
gender, nationality	1.1511	0.0431	gender, ethnicity	0.3221	0.0118
race, nationality	1.1534	0.0163	race, ethnicity	0.6911	0.0167
race, gender	1.7511	0.0451	race, gender	0.6855	0.0137
race, gender, nationality	1.9751	0.0455	race, gender, ethnicity	0.8498	0.0163

The property is violated if removing a protected attribute increases the metric. As expected, ϵ -DF obeyed the intersectionality property for all datasets, but γ -SF violated it as γ for *gender* > γ for *race* \times *gender* (COMPAS); γ for *gender* > γ for *gender* \times *nationality* (Adult); γ for *age* > γ for *gender* \times *age* (HHP); and γ for *race* \times *ethnicity* > γ for *race* \times *gender* \times *ethnicity* (HMDA).

11. Discussion

In this work, we introduced fairness measures for AI/ML algorithms and data with regard to multiple protected attributes. Our proposed *differential fairness* (DF) metrics are informed by the framework of intersectionality, which analyzes how interlocking systems of power and oppression affect individuals along overlapping dimensions including race, gender, sexual orientation, class, and disability. DF has a particularly elegant intersectionality property that protects higher-level groups (e.g., *women*) automatically by protecting intersectional subgroups (*Black women*). There are several other attractive properties of our formalism and we provided proofs for the privacy, economic, and generalization properties. DF is *lightweight*, in that it does not require the specification or estimation of a causal model [54] or a latent model of risk distributions [43]. It is nevertheless able to make use of a probabilistic model of the data, when available, but does not require one. While no fairness definition is suitable in all contexts, due to these properties we recommend the use of differential fairness when an intersectional measure of fairness is appropriate.

It has been recently shown that many fairness notions can be captured for fair classifications as a combination of equality constraints between protected groups using Lagrangian dual approaches [78–80]. However, these methods were developed to enforce group fairness notions and cannot be directly adapted to learn fairness in an intersectional context, especially for learning in stochastic settings, due to the data sparsity of multi-dimensional protected attributes. We developed simple and practical learning algorithms which enforce DF, thereby ensuring fairness in a manner which behaves sensibly for any subset of the set of protected attributes. Our experimental results on multiple benchmark datasets demonstrate that the proposed algorithms substantially improve DF metrics over the baseline models, with modest costs in accuracy. Furthermore, unlike the baseline models, our algorithms often greatly improve other group-level (*p*%-Rule) and subgroup-level (SF) fairness metrics. Finally, we showed that our DF metric-based methods are able to achieve intersectionally fair classification with little loss in predictive performance, while providing greater protection to minority groups (in terms of population size per group) compared to SF metric-based methods.

However, the measurement of fairness becomes statistically challenging in the intersectional setting due to data sparsity, which increases rapidly in the number of dimensions (e.g., *gender*, *race*, *age*, etc.), and in the values per dimension (e.g., *race* attributes may include *black*, *white*, *asian*, etc.). In this paper, we addressed this by developing a learning algorithm for the reliable and stochastic approximation-based estimation of intersectional fairness penalty, incorporated with the objective function of deep neural networks. In our previous work [56], we proposed using hierarchical Bayesian probabilistic modeling to manage uncertainty of the intersectional fairness measures via MCMC or variational inference. It would be interesting to extend our approach to use this fully Bayesian uncertainty estimation strategy during training, although it would be costly to compute and differentiate MCMC samples or a variational posterior per minibatch for deep learning-based methods trained via stochastic gradient descent.

In other future work, differential privacy has been extended and generalized in various ways, including notably its generalization to Rényi differential privacy [21]. It would be interesting to investigate analogous generalizations for differential fairness. Another potential extension of the definition which is important for applications is a version that protects error rates as opposed to outcomes, following [8].

More generally, an important aspect of this work is to integrate research on fairness and bias issues in our society from disciplines such as the humanities, social science, and law,

in our case regarding intersectionality, with technical methods for fair AI/machine learning. AI fairness is not a purely technical issue, and hence the field of computer science cannot solve it alone. There is an urgent need for more research on interdisciplinary approaches and perspectives on AI fairness, and this work is one small step in that direction. A special issue on this topic recently published in the IEEE Data Engineering Bulletin, edited by two of the authors of this paper, also aims to promote this emerging direction of research [81].

12. Conclusions

In this research we took an interdisciplinary approach to fairness in AI and machine learning systems by developing technical methods to implement intersectionality, a perspective on fairness and bias arising from the humanities, law, and the social sciences disciplines. We introduced three AI fairness definitions satisfying intersectional fairness desiderata, *differential fairness* and its *bias amplification* and *confounder-aware* counterparts, and proved their attractive properties regarding law, privacy, economics, and statistical learning. Our theoretical results show that differential fairness operationalizes fairness in a meaningful and interpretable way, as it quantifies the potential real-world economic and privacy harms due to the algorithm. We developed learning algorithms to enforce our criteria. We further addressed the data sparsity problem of intersectional fairness estimation which arises in SGD minibatches, using a novel intersectionally fair stochastic learning algorithm. In extensive experiments, we showed that our criteria can be practically attained, and that they behave more equitably with regard to minority groups than subgroup fairness. Going forward, we argue that it is essential to deeply integrate research from the humanities and social science on fairness and bias issues with work from the field of computer science, in order to holistically address both the human and algorithmic sides of the AI fairness equation. This work is a step in that direction, but much more remains to be done.

Author Contributions: Conceptualization, J.R.F.; methodology, J.R.F. and R.I.; software, R.I. and K.N.K.; validation, R.I. and K.N.K.; formal analysis, J.R.F. and R.I.; investigation, R.I. and K.N.K.; resources, J.R.F.; data curation, R.I. and A.D.S.; writing—original draft preparation, J.R.F. and R.I.; writing—review and editing, J.R.F., S.P. and A.D.S.; visualization, R.I.; supervision, J.R.F. and S.P.; project administration, J.R.F.; funding acquisition, J.R.F. All authors have read and agreed to the published version of the manuscript.

Funding: 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.'s IIS1927486; IIS1850023; IIS2046381. 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. The work of ADS was supported in part by the US NSF under award CCF-1453432.

Institutional Review Board Statement: Ethical review and approval were waived due to a Not Human Subjects Research designation by the UMBC Office of Research Protections and Compliance.

Data Availability Statement: Publicly available datasets were analyzed in this study, except for the HDMA dataset. Restrictions apply to the availability of these data. The HDMA dataset was obtained from the Federal Financial Institutions Examination Council's (FFIEC) and are available at <https://www.ffiec.gov/hmda/> (accessed on 10 April 2023) with the permission of the FFIEC. Links and/or citations to the sources of the datasets used were provided in Section 10.1.

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

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A. Additional Experimental Results

Appendix A.1. Fairness and Accuracy Trade-Off

We report in Figure A1 additional results for the rest of the datasets (COMPAS and HMDA), for our experiment in Section 10.2.2, regarding fairness and accuracy trade-off. The results validate the conclusions drawn in Section 10.2.2. For the COMPAS dataset, note that we found a noisy relation between fairness and accuracy, presumably due to the unstable prediction by the deep neural network on this small dataset. Therefore, Pareto frontier is used to show the impact of λ on the fair models (Figure A1a).

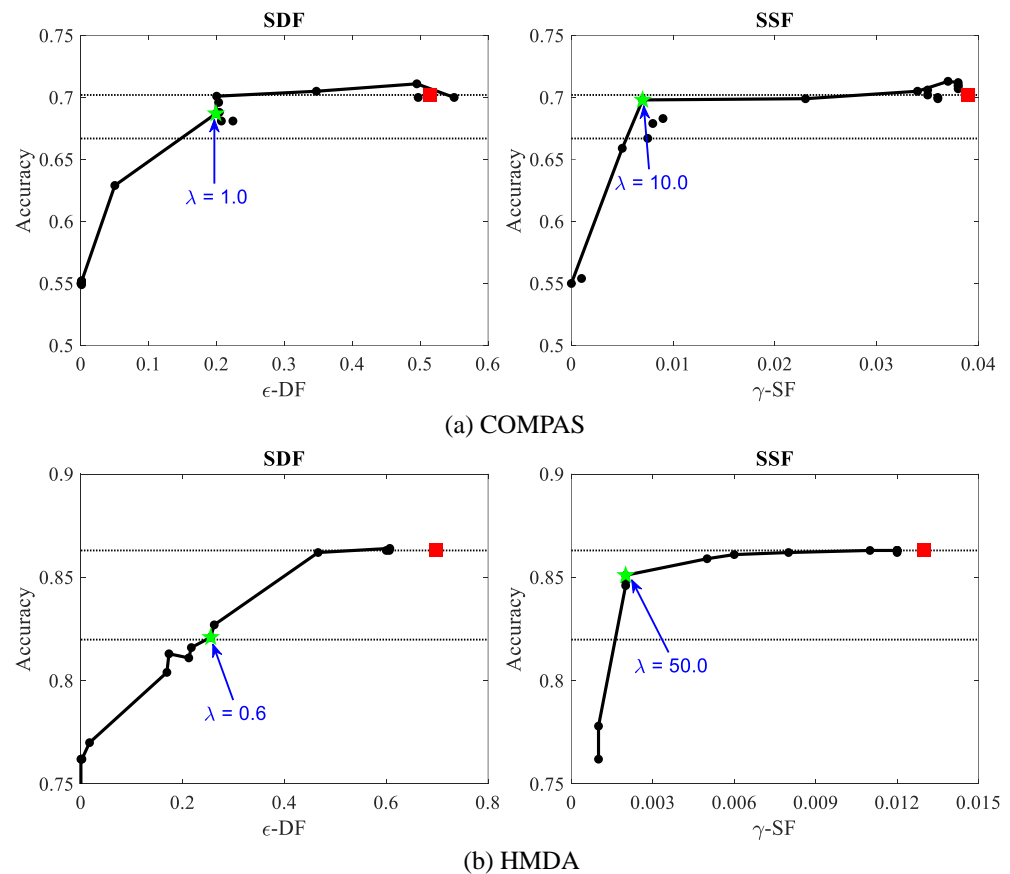


Figure A1. Fairness and accuracy trade-off plots for stochastic DF (SDF) and SF (SSF) models on the development set of (a) COMPAS and (b) HMDA datasets. Red square: stochastic typical (ST) classifier that does not incorporate fairness penalty in the objective; dotted lines: indicate the valid region of 5% degraded accuracy from ST; black circles: fair models (SDF or SSF) correspond to different λ values (larger to smaller from left to right); green asterisk: fairest model within the valid region for a particular λ value.

Appendix A.2. Comparative Analysis for Batch and Stochastic Methods

We report the additional results for rest of the datasets regarding the experiment of comparative analysis in Section 10.2.3 between batch and stochastic fair models. Figure A2 compares the computed ϵ -DF and accuracy for the BDF and SDF models, and Figure A3 compares γ -SF and accuracy for the BSF and SSF models in terms of runtime on the development set of COMPAS and HMDA. Our conclusion for this experiment remains the same: that the stochastic fair models do not cause any additional harm to the model's fairness or predictive performance due to the noisy estimation of fairness metrics for intersectional groups in minibatches during training. Note that it was not feasible to train the models in batch setting for the very large HMDA training data. So, we randomly took a subset with 10% of the whole HMDA training data to train the batch methods.

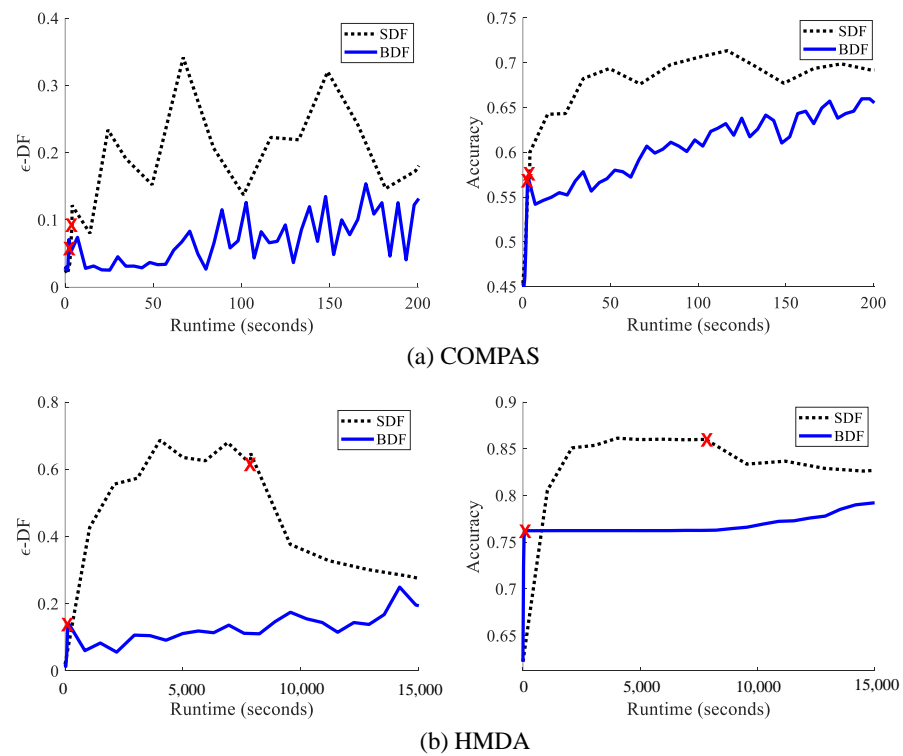


Figure A2. Comparison of ϵ -DF and accuracy in terms of runtime for the batch differential fair (BDF) and stochastic differential fair (SDF) models on the development set of (a) COMPAS and (b) HMDA datasets. Red X: end of burn-in (disabling the fairness interventions) iterations and epochs for the BDF and SDF models, respectively.

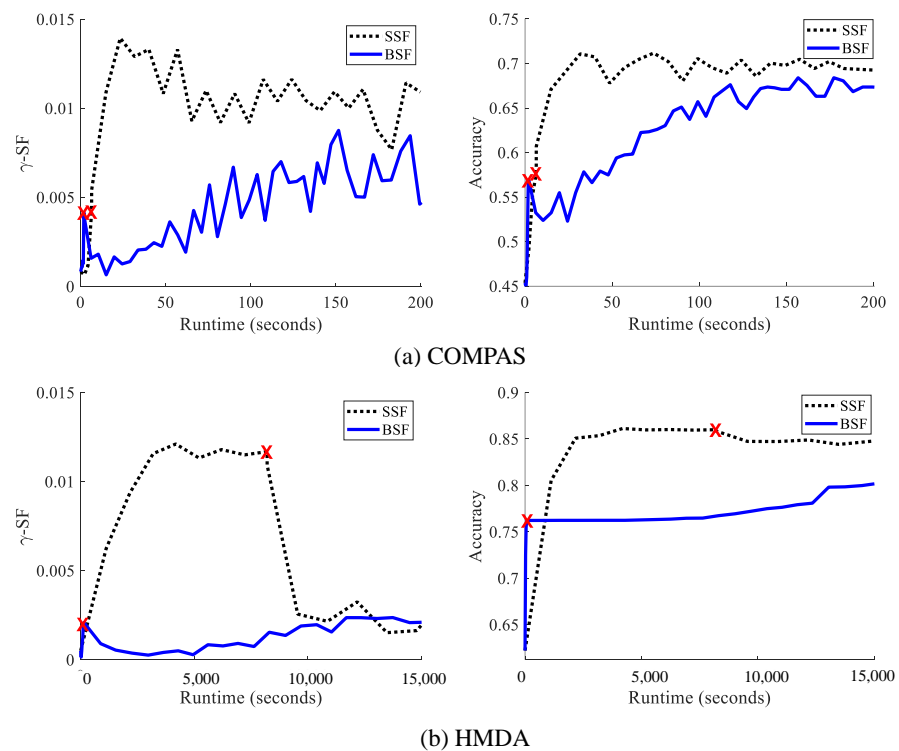


Figure A3. Comparison of γ -SF and accuracy in terms of runtime for the batch subgroup fair (BSF) and stochastic subgroup fair (SSF) models on the development set of (a) COMPAS and (b) HMDA datasets. Red X: end of burn-in (disabling the fairness interventions) iterations and epochs for the BSF and SSF models, respectively.

Appendix A.3. Performance for Fair Classifiers

In this additional experiment, we again compare the batch and stochastic classifiers on the test set for the rest of the datasets to complement the experiment in Section 10.2.4.

In Table A1, we compare the batch methods on the HHP and HMDA test sets, respectively. Both BDF and BSF classifiers were found to substantially improve fairness over the BPR and BT classifiers, however with modest decrements in accuracy, F1 score, and ROC AUC. For the HHP dataset, the BDF with $\epsilon_1 = 0$ improved from $\epsilon = 1.654$ to $\epsilon = 0.215$ with a sacrifice of 4.7 percentage points in accuracy. Similar to HMDA, it improved from $\epsilon = 0.553$ to $\epsilon = 0.242$, while it decreased the accuracy by 4.9. BDF performed as the best fair model in terms of both ϵ -DF and γ -SF on these datasets.

Stochastic fair models also showed similar trends in the performance on the COMPAS and Adult test data (Table A2). On COMPAS, SDF with $\epsilon_1 = 0$ performed again as the best fair model in terms of both ϵ -DF and γ -SF with an accuracy loss of 1.2 percentage points, while it improved from $\epsilon = 1.446$ to $\epsilon = 0.104$, with 2.9 percentage points of decrease in accuracy, on Adult data. When SDF trained to prevent bias amplification (i.e., $\epsilon_1 = \epsilon_{data}$), there were modest improvements in the fairness metrics with relatively little reduction in accuracy.

The performance analysis of fair models (Tables A1 and A2) points to the same conclusion of Section 10.2.4; DF-Classifiers (BDF and SDF) typically had comparable accuracy under the same settings compared to SF-Classifiers (BSF and SSF) and p %-Rule Classifiers (BPR and SPR), while DF-Classifiers often greatly improved γ -SF and p %-Rule, but the SF- and p %-Rule Classifiers had modest improvements, or even increases, in ϵ -DF.

Table A1. Comparison of batch differential fair (BDF), batch subgroup fair (BSF), batch p %-Rule (BPR), and batch typical (BT) classifiers on the test set of HHP and HMDA datasets ($\epsilon_1 = 0.2231$ is the 80% rule). Higher is better for measures with \uparrow ; lower is better for measures with \downarrow .

HHP Dataset								
Models		BDF-Classifier			BSF-Classifier		BPR-Classifier	BT-Classifier
		$\epsilon_1 = 0.0$	$\epsilon_1 = 0.2231$	$\epsilon_1 = \epsilon_{data}$	$\gamma_1 = 0.0$	$\gamma_1 = \gamma_{data}$		
Tuning Parameter	λ	0.100	0.100	50.000	10.000	70.000	0.010	X
Performance Measures	Accuracy \uparrow	0.801	0.810	0.800	0.818	0.835	0.816	0.848
	F1 Score \uparrow	0.590	0.610	0.650	0.640	0.700	0.660	0.730
	ROC AUC \uparrow	0.777	0.792	0.848	0.847	0.868	0.842	0.885
Fairness Measures	ϵ -DF \downarrow	0.215	0.277	1.232	0.996	1.594	1.450	1.654
	γ -SF \downarrow	0.003	0.004	0.017	0.007	0.012	0.021	0.020
	p %-Rule \uparrow	61.610	60.628	53.809	55.060	50.281	62.128	51.122
	Bias Amp-DF \downarrow	-1.245	-1.183	-0.228	-0.464	0.134	-0.010	0.194
	Bias Amp-SF \downarrow	-0.010	-0.009	0.004	-0.006	-0.001	0.008	0.007
HMDA Dataset								
Models		BDF-Classifier			BSF-Classifier		BPR-Classifier	BT-Classifier
		$\epsilon_1 = 0.0$	$\epsilon_1 = 0.2231$	$\epsilon_1 = \epsilon_{data}$	$\gamma_1 = 0.0$	$\gamma_1 = \gamma_{data}$		
Tuning Parameter	λ	0.070	0.100	1.000	5.000	70.000	0.001	X
Performance Measures	Accuracy \uparrow	0.800	0.803	0.856	0.812	0.825	0.839	0.849
	F1 Score \uparrow	0.870	0.870	0.900	0.880	0.880	0.890	0.900
	ROC AUC \uparrow	0.888	0.895	0.924	0.898	0.898	0.910	0.916
Fairness Measures	ϵ -DF \downarrow	0.242	0.375	0.605	0.457	0.390	0.537	0.553
	γ -SF \downarrow	0.001	0.002	0.009	0.001	0.003	0.006	0.008
	p %-Rule \uparrow	70.528	71.846	65.888	69.342	67.231	71.642	66.122
	Bias Amp-DF \downarrow	-0.419	-0.286	-0.056	-0.204	-0.271	-0.124	-0.108
	Bias Amp-SF \downarrow	-0.008	-0.007	0.000	-0.008	-0.006	-0.003	-0.001

To showcase the impact of our work in addressing fairness for minority groups, we report again the “per-group fairness” of the classifiers for the rest of the datasets versus the group’s probability (i.e., size) in Figures A4 and A5. The batch fair models (BDF and BSF) and the stochastic fair models (SDF and SSF) generally improved their corresponding

fairness measure per group over the BT and ST models, respectively. However, the minority groups were not able to influence the overall γ -SF, since the γ -SF metric assigns the highest per-group γ -SF values to large groups (i.e., top-level group). As an example for the HHP and HMDA datasets, the overall γ -SF is determined by a top-level group (Red “X”) for SSF classifiers. On the other hand, the overall ϵ -DF is determined by an intersectional subgroup (Blue “X”) regardless of the subgroup’s size for all the examples in Figures A4 and A5.

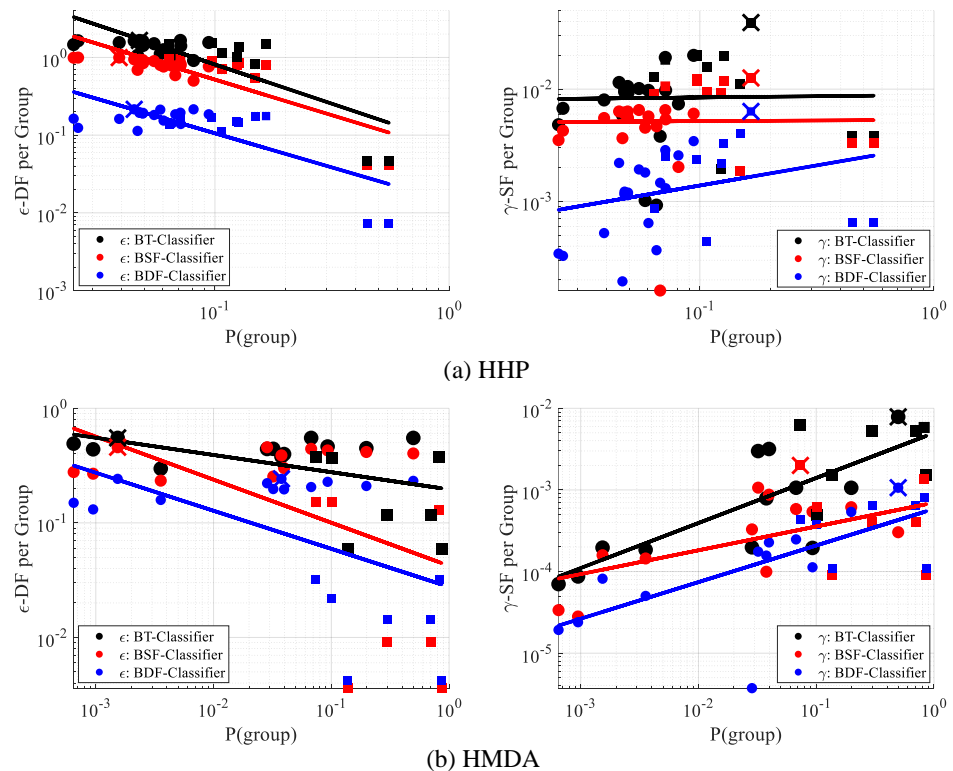


Figure A4. Per-group measurements of ϵ -DF and γ -SF for the batch differential fair (BDF), batch subgroup fair (BSF), batch p %-Rule (BPR), and batch typical (BT) classifiers vs. group size (probability), (a) HHP and (b) HMDA datasets, calculated using Equation (3) (γ -SF) and Equation (2) (ϵ -DF) with the group held fixed. Circles: intersectional subgroups. Squares: top-level groups. Fitted least squares line demonstrates the overall trend between fairness measures and group size. Largest per-group ϵ -DF and γ -SF, indicated using “X”, determines the overall ϵ -DF and γ -SF, respectively.

Table A2. Comparison of stochastic differential fair (SDF), stochastic subgroup fair (SSF), stochastic p %-Rule (SPR), and stochastic typical (ST) on the test set of COMPAS and Adult datasets ($\epsilon_1 = 0.2231$ is the 80% rule). Higher is better for measures with \uparrow ; lower is better for measures with \downarrow .

COMPAS Dataset								
Models		SDF-Classifier			SSF-Classifier		SPR-Classifier	ST-Classifier
		$\epsilon_1 = 0.0$	$\epsilon_1 = 0.2231$	$\epsilon_1 = \epsilon_{data}$	$\gamma_1 = 0.0$	$\gamma_1 = \gamma_{data}$		
Tuning Parameter	λ	1.000	1.000	70.000	10.000	70.000	0.010	X
Performance Measures	Accuracy \uparrow	0.671	0.667	0.670	0.660	0.674	0.669	0.683
	F1 Score \uparrow	0.600	0.600	0.590	0.600	0.600	0.610	0.630
	ROC AUC \uparrow	0.704	0.698	0.694	0.694	0.703	0.706	0.713
Fairness Measures	ϵ -DF \downarrow	0.176	0.343	0.374	0.313	0.517	0.676	0.888
	γ -SF \downarrow	0.004	0.014	0.020	0.005	0.012	0.029	0.032
	p %-Rule \uparrow	63.064	62.953	64.306	61.885	62.727	64.684	61.680
	Bias Amp-DF \downarrow	-0.362	-0.195	-0.164	-0.225	-0.021	0.138	0.350
	Bias Amp-SF \downarrow	-0.016	-0.006	0.000	-0.015	-0.008	0.009	0.012

Table A2. Cont.

Adult Dataset								
Models		SDF-Classifier			SSF-Classifier		SPR-Classifier	ST-Classifier
		$\epsilon_1 = 0.0$	$\epsilon_1 = 0.2231$	$\epsilon_1 = \epsilon_{data}$	$\gamma_1 = 0.0$	$\gamma_1 = \gamma_{data}$		
Tuning Parameter	λ	1.000	1.000	25.000	100.000	1000.000	0.100	X
Performance Measures	Accuracy \uparrow	0.803	0.808	0.810	0.802	0.811	0.815	0.832
	F1 Score \uparrow	0.390	0.440	0.420	0.410	0.390	0.480	0.600
	ROC AUC \uparrow	0.821	0.844	0.861	0.843	0.871	0.858	0.885
Fairness Measures	ϵ -DF \downarrow	0.104	0.259	0.648	0.821	1.028	1.268	1.446
	γ -SF \downarrow	0.003	0.012	0.023	0.002	0.017	0.028	0.042
	$p\%$ -Rule \uparrow	59.896	58.054	52.147	59.036	50.220	59.616	46.052
	Bias Amp-DF \downarrow	-1.276	-1.121	-0.732	-0.559	-0.352	-0.112	0.066
	Bias Amp-SF \downarrow	-0.030	-0.021	-0.010	-0.031	-0.016	-0.555	0.009

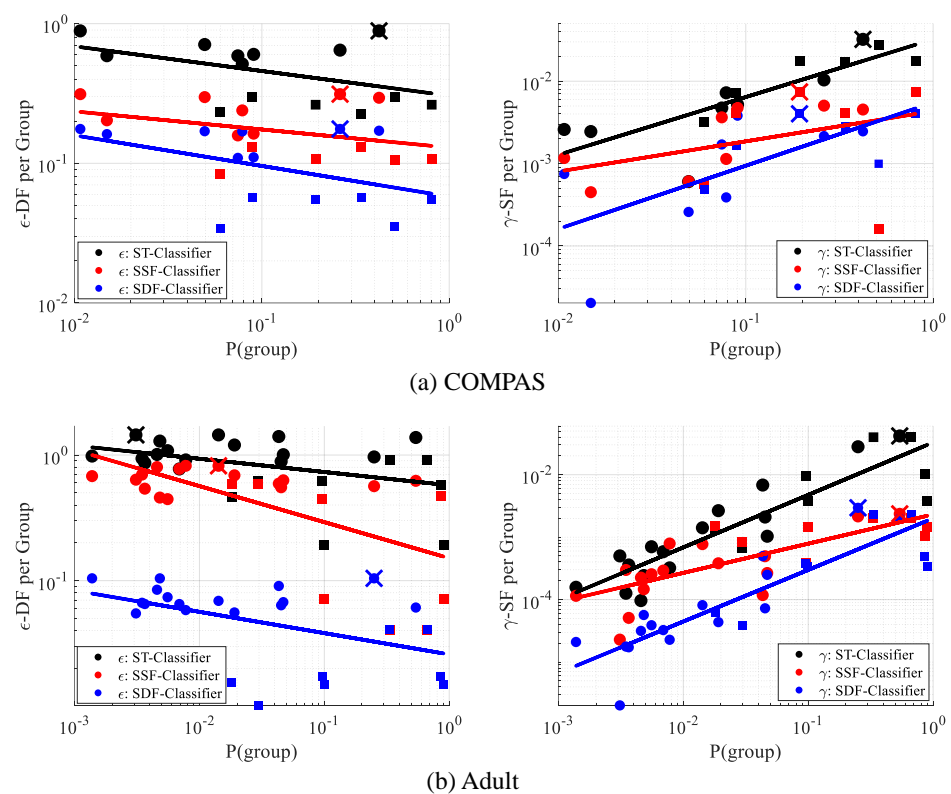


Figure A5. Per-group measurements of ϵ -DF and γ -SF for the stochastic differential fair (SDF), stochastic subgroup fair (SSF), stochastic $p\%$ -Rule (SPR), and stochastic typical (ST) classifiers vs. group size (probability), (a) COMPAS and (b) Adult datasets, calculated using Equation (3) (γ -SF) and Equation (2) (ϵ -DF) with the group held fixed. Circles: intersectional subgroups. Squares: top-level groups. Fitted least squares line demonstrates the overall trend between fairness measures and group size. Largest per-group ϵ -DF and γ -SF, indicated using “X”, determines the overall ϵ -DF and γ -SF, respectively.

Our overall conclusion for this experiment remains the same as in Section 10.2.4: *DF metric-based BDF and SDF models are able to achieve intersectionally fair classification with a minor loss in performance, while providing greater protection to minority groups than when enforcing SF, such as in BSF and SSF models.*

References

1. Barocas, S.; Selbst, A. Big data's disparate impact. *Calif. Law Rev.* **2016**, *104*, 671. [CrossRef]
2. Munoz, C.; Smith, M.; Patil, D. *Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights*; Executive Office of the President: Washington, DC, USA, 2016.

3. Noble, S. *Algorithms of Oppression: How Search Engines Reinforce Racism*; NYU Press: New York, NY, USA, 2018.
4. Angwin, J.; Larson, J.; Mattu, S.; Kirchner, L. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica* **2016**. Available online: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> (accessed on 4 April 2023).
5. Buolamwini, J.; Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Proceedings of the Conference on Fairness, Accountability, and Transparency, New York, NY, USA, 23–24 February 2018.
6. Bolukbasi, T.; Chang, K.W.; Zou, J.; Saligrama, V.; Kalai, A. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In Proceedings of the Advances in NeurIPS, Barcelona, Spain, 5–10 December 2016.
7. Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; Zemel, R. Fairness through awareness. In Proceedings of the Innovations in Theoretical Computer Science (ITCS), Cambridge, MA USA, 8–10 January 2012; pp. 214–226.
8. Hardt, M.; Price, E.; Srebro, N. Equality of opportunity in supervised learning. Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Barcelona, Spain, 5–10 December 2016; pp. 3315–3323.
9. Berk, R.; Heidari, H.; Jabbari, S.; Joseph, M.; Kearns, M.; Morgenstern, J.; Neel, S.; Roth, A. A convex framework for fair regression. In Proceedings of the FAT/ML Workshop, Halifax, NS, Canada, 14 August 2017.
10. Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; Chang, K.W. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Copenhagen, Denmark, 9–11 September 2017.
11. Islam, R.; Keya, K.; Zeng, Z.; Pan, S.; Foulds, J. Debiasing Career Recommendations with Neural Fair Collaborative Filtering. In Proceedings of the Web Conference, Ljubljana, Slovenia, 19–23 April 2021.
12. Keya, K.N.; Islam, R.; Pan, S.; Stockwell, I.; Foulds, J.R. Equitable Allocation of Healthcare Resources with Fair Survival Models. In Proceedings of the 2021 SIAM International Conference on Data Mining (SIAM), Virtual, 29 April–1 May 2021.
13. Campolo, A.; Sanfilippo, M.; Whittaker, M.; Crawford, A.S.K.; Barocas, S. *AI Now 2017 Symposium Report*; AI Now: New York, NY, USA, 2017.
14. Mitchell, S.; Potash, E.; Barocas, S. Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions. *arXiv* **2018**, arXiv:1811.07867.
15. Keyes, O.; Hutson, J.; Durbin, M. A Mulching Proposal: Analysing and Improving an Algorithmic System for Turning the Elderly into High-Nutrient Slurry. In Proceedings of the Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, UK, 4–9 May 2019; ACM: New York, NY, USA, 2019.
16. Crenshaw, K. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *Univ. Chic. Leg. Forum* **1989**, 1989, 139–167.
17. Kearns, M.; Neel, S.; Roth, A.; Wu, Z. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In Proceedings of the International Conference on Machine Learning (ICML), Stockholm, Sweden, 10–15 July 2018; pp. 2569–2577.
18. Hebert-Johnson, U.; Kim, M.; Reingold, O.; Rothblum, G. Multicalibration: Calibration for the (Computationally-Identifiable) Masses. In Proceedings of the International Conference on Machine Learning (ICML), Stockholm, Sweden, 10–15 July 2018; pp. 1944–1953.
19. Dwork, C.; McSherry, F.; Nissim, K.; Smith, A. Calibrating noise to sensitivity in private data analysis. In Proceedings of the Third Theory of Cryptography, New York, NY, USA, 4–7 March 2006; pp. 265–284.
20. Dwork, C.; Roth, A. The algorithmic foundations of differential privacy. *Theor. Comput. Sci.* **2013**, 9, 211–407.
21. Mironov, I. Rényi differential privacy. In Proceedings of the 2017 IEEE 30th Computer Security Foundations symposium (CSF), Santa Barbara, CA, USA, 21–25 August 2017; pp. 263–275.
22. Selbst, A.D.; Boyd, D.; Friedler, S.A.; Venkatasubramanian, S.; Vertesi, J. Fairness and abstraction in sociotechnical systems. In Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA, 29–31 January 2019; pp. 59–68.
23. Jacobs, A.Z.; Wallach, H. Measurement and Fairness. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event, 3–10 March 2021; Association for Computing Machinery: New York, NY, USA, 2021; pp. 375–385. [\[CrossRef\]](#)
24. Cheng, H.F.; Wang, R.; Zhang, Z.; O'Connell, F.; Gray, T.; Harper, F.M.; Zhu, H. Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, UK, 4–9 May 2019; pp. 1–12.
25. Van Berkel, N.; Goncalves, J.; Russo, D.; Hosio, S.; Skov, M.B. Effect of information presentation on fairness perceptions of machine learning predictors. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan, 8–13 May 2021; pp. 1–13.
26. Wang, R.; Harper, F.M.; Zhu, H. Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25–30 April 2020; pp. 1–14.
27. Kifer, D.; Machanavajjhala, A. Pufferfish: A framework for mathematical privacy definitions. *TODS* **2014**, 39, 3. [\[CrossRef\]](#)
28. Green, B.; Chen, Y. Algorithmic risk assessments can alter human decision-making processes in high-stakes government contexts. *Proc. ACM Hum.-Comput. Interact.* **2021**, 5, 1–33. [\[CrossRef\]](#)
29. Green, B. The flaws of policies requiring human oversight of government algorithms. *Comput. Law Secur. Rev.* **2022**, 45, 105681. [\[CrossRef\]](#)

30. Kong, Y. Are “Intersectionally Fair” AI Algorithms Really Fair to Women of Color? A Philosophical Analysis. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, 21–24 June 2022; Association for Computing Machinery: New York, NY, USA, 2022; pp. 485–494. [\[CrossRef\]](#)
31. Foulds, J.R.; Islam, R.; Keya, K.N.; Pan, S. An intersectional definition of fairness. In Proceedings of the 2020 IEEE 36th International Conference on Data Engineering (ICDE), Dallas, TX, USA, 20–24 April 2020; pp. 1918–1921.
32. Foulds, J.; Pan, S. Are Parity-Based Notions of AI Fairness Desirable? *Bull. IEEE Tech. Comm. Data Eng.* **2020**, *43*, 51–73.
33. Ganley, C.M.; George, C.E.; Cimpian, J.R.; Makowski, M.B. Gender equity in college majors: Looking beyond the STEM/Non-STEM dichotomy for answers regarding female participation. *Am. Educ. Res. J.* **2018**, *55*, 453–487. [\[CrossRef\]](#)
34. Piper, A. Passing for white, passing for black. *Transition* **1992**, *58*, 4–32. [\[CrossRef\]](#)
35. Truth, S. *Ain’t I a Woman?* Speech delivered at Women’s Rights Convention: Akron, OH, USA, 1851.
36. Collins, P.H. *Black Feminist thought: Knowledge, Consciousness, and the Politics of Empowerment*; Routledge: London, UK, 2002.
37. Combahee River Collective. A Black Feminist Statement. In *Capitalist Patriarchy and the Case for Socialist Feminism*; Eisenstein, Z., Ed.; Monthly Review Press: New York, NY, USA, 1978.
38. Hooks, B. *Ain’t I a Woman: Black Women and Feminism*; South End Press: Boston, MA, USA, 1981.
39. Lorde, A. Age, race, class, and sex: Women redefining difference. In *Sister Outsider*; Ten Speed Press: Berkeley, CA, USA, 1984; pp. 114–124.
40. Yang, F.; Cisse, M.; Koyejo, O.O. Fairness with Overlapping Groups; a Probabilistic Perspective. In Proceedings of the Advances in Neural Information Processing Systems 33 (NeurIPS 2020), Virtual, 6–12 December 2020.
41. La Cava, W.; Lett, E.; Wan, G. Proportional Multicalibration. *arXiv* **2022**, arXiv:2209.14613.
42. Lett, E.; La Cava, W. Translating Intersectionality to Fair Machine Learning in Health Sciences. *SocArXiv* **2023**. [\[CrossRef\]](#)
43. Simoiu, C.; Corbett-Davies, S.; Goel, S. The problem of infra-marginality in outcome tests for discrimination. *Ann. Appl. Stat.* **2017**, *11*, 1193–1216. [\[CrossRef\]](#)
44. Davis, A. *Are Prisons Obsolete?* Seven Stories Press: New York, NY, USA, 2011.
45. Wald, J.; Losen, D. Defining and redirecting a school-to-prison pipeline. *New Dir. Youth Dev.* **2003**, *2003*, 9–15. [\[CrossRef\]](#)
46. Verschelden, C. *Bandwidth Recovery: Helping Students Reclaim Cognitive Resources Lost to Poverty, Racism, and Social Marginalization*; Stylus: Sterling, VA, USA, 2017.
47. Alexander, M. *The New Jim Crow: Mass Incarceration in the Age of Colorblindness*; T. N. P.: New York, NY, USA, 2012.
48. Grant, J.; Mottet, L.; Tanis, J.; Harrison, J.; Herman, J.; Keisling, M. *Injustice at Every Turn: A Report of the National Transgender Discrimination Survey*; National Center for Transgender Equality: Washington, DC, USA, 2011.
49. Berk, R.; Heidari, H.; Jabbari, S.; Kearns, M.; Roth, A. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociol. Methods Res.* **2018**, *1050*, 28. [\[CrossRef\]](#)
50. Dastin, J. Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women. *Reuters* **2018**. Available online: <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G> (accessed on 4 April 2023).
51. Equal Employment Opportunity Commission. Guidelines on Employee Selection Procedures. In *C.F.R. 29.1607*; Office of the Federal Register: Washington, DC, USA, 1978.
52. Pleiss, G.; Raghavan, M.; Wu, F.; Kleinberg, J.; Weinberger, K. On fairness and calibration. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 5684–5693.
53. Donini, M.; Oneto, L.; Ben-David, S.; Shawe-Taylor, J.S.; Pontil, M. Empirical risk minimization under fairness constraints. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Montreal, QC, Canada, 3–8 December 2018; pp. 2791–2801.
54. Kusner, M.; Loftus, J.; Russell, C.; Silva, R. Counterfactual fairness. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017.
55. Jagielski, M.; Kearns, M.; Mao, J.; Oprea, A.; Roth, A.; Sharifi-Malvajerdi, S.; Ullman, J. Differentially private fair learning. *arXiv* **2018**, arXiv:1812.02696.
56. Foulds, J.; Islam, R.; Keya, K.N.; Pan, S. Bayesian Modeling of Intersectional Fairness: The Variance of Bias. In Proceedings of the 2020 SIAM International Conference on Data Mining, Cincinnati, OH, USA, 5–9 May 2020; pp. 424–432.
57. Charig, C.; Webb, D.; Payne, S.; Wickham, J. Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy. *Br. Med. J.* **1986**, *292*, 879–882. [\[CrossRef\]](#)
58. Julious, S.; Mullee, M. Confounding and Simpson’s paradox. *Br. Med. J.* **1994**, *309*, 1480–1481. [\[CrossRef\]](#) [\[PubMed\]](#)
59. Bickel, P.J.; Hammel, E.A.; O’Connell, J.W. Sex Bias in Graduate Admissions: Data from Berkeley: Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation. *Science* **1975**, *187*, 398–404. [\[CrossRef\]](#) [\[PubMed\]](#)
60. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
61. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [\[CrossRef\]](#) [\[PubMed\]](#)
62. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic differentiation in pytorch. In Proceedings of the Advances in Neural Information Processing Systems (Autodiff Workshop), Long Beach, CA, USA, 4–9 December 2017.
63. Cappé, O.; Moulines, E. On-line expectation–maximization algorithm for latent data models. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2009**, *71*, 593–613. [\[CrossRef\]](#)

64. Hoffman, M.; Bach, F.R.; Blei, D.M. Online learning for latent Dirichlet allocation. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 6–9 December 2010; pp. 856–864.
65. Hoffman, M.D.; Blei, D.M.; Wang, C.; Paisley, J. Stochastic variational inference. *J. Mach. Learn. Res.* **2013**, *14*, 1303–1347.
66. Mimno, D.; Hoffman, M.D.; Blei, D.M. Sparse stochastic inference for latent Dirichlet allocation. In Proceedings of the 29th International Conference on Machine Learning, Edinburgh, Scotland, 26 June–1 July 2012; pp. 1515–1522.
67. Foulds, J.; Boyles, L.; DuBois, C.; Smyth, P.; Welling, M. Stochastic collapsed variational Bayesian inference for latent Dirichlet allocation. In Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, IL, USA, 11–14 August 2013; pp. 446–454.
68. Islam, R.; Foulds, J. Scalable Collapsed Inference for High-Dimensional Topic Models. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; pp. 2836–2845.
69. Robbins, H.; Monro, S. A stochastic approximation method. *Ann. Math. Statist.* **1951**, *22*, 400–407. [[CrossRef](#)]
70. Andrieu, C.; Moulines, É.; Priouret, P. Stability of stochastic approximation under verifiable conditions. *SIAM J. Control Optim.* **2005**, *44*, 283–312. [[CrossRef](#)]
71. Bao, M.; Zhou, A.; Zottola, S.; Brubach, B.; Desmarais, S.; Horowitz, A.; Lum, K.; Venkatasubramanian, S. It's COMPASlicated: The Messy Relationship between RAI Datasets and Algorithmic Fairness Benchmarks. In Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1 (NeurIPS Datasets and Benchmarks 2021), Virtual, 6–14 December 2021.
72. Dua, D.; Graff, C. *UCI Machine Learning Repository*; University of California, School of Information and Computer Science: Irvine, CA, USA, 2017. Available online: <http://archive.ics.uci.edu/ml> (accessed on 10 April 2023).
73. Song, J.; Kalluri, P.; Grover, A.; Zhao, S.; Ermon, S. Learning controllable fair representations. In Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics, Naha, Japan, 16–18 April 2019; pp. 2164–2173.
74. ProPublica/Investigative Reporters and Editors. Home Mortgage Disclosure Act. 2015. Available online: <https://www.consumerfinance.gov/data-research/hmda/> (accessed on 10 April 2023).
75. Zafar, M.; Valera, I.; Rodriguez, M.; Gummadi, K. Fairness constraints: Mechanisms for fair classification. In Proceedings of the AISTATS, Lauderdale, FL, USA, 20–22 April 2017.
76. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Adv. Neural Inf. Process. Syst. (NeurIPS)* **2019**, *32*, 8026–8037.
77. Lorenz, M. Methods of measuring the concentration of wealth. *Publ. Am. Stat. Assoc.* **1905**, *9*, 209–219. [[CrossRef](#)]
78. Agarwal, A.; Beygelzimer, A.; Dudík, M.; Langford, J.; Wallach, H. A reductions approach to fair classification. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 60–69.
79. Fioretto, F.; Van Hentenryck, P.; Mak, T.W.; Tran, C.; Baldo, F.; Lombardi, M. Lagrangian Duality for Constrained Deep Learning. In Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases, ECML PKDD 2020, Ghent, Belgium, 14–18 September 2020; Springer Science and Business Media Deutschland GmbH: Berlin/Heidelberg, Germany, 2021; pp. 118–135.
80. Tran, C.; Fioretto, F.; Van Hentenryck, P. Differentially Private and Fair Deep Learning: A Lagrangian Dual Approach. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Volume 35, pp. 9932–9939.
81. Foulds, J.R.; Pan, S. (Eds.) *Bulletin of the IEEE Technical Committee on Data Engineering*; IEEE Computer Society: Washington, DC, USA, 2020; Volume 43.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.