An Empirical Study of Rich Subgroup Fairness for Machine Learning

Michael Kearns
Department of Computer and Information Sciences, University of Pennsylvania
mkearns@cis.upenn.edu

Seth Neel
Department of Statistics, University of Pennsylvania
sethneel@wharton.upenn.edu

Aaron Roth
Department of Computer and Information Sciences, University of Pennsylvania
aaroth@cis.upenn.edu

Zhiwei Steven Wu
Department of Computer Science and Engineering, University of Minnesota
zsw@umn.edu

ABSTRACT

Kearns, Neel, Roth, and Wu [ICML 2018] recently proposed a notion of *rich subgroup fairness* intended to bridge the gap between statistical and individual notions of fairness. Rich subgroup fairness picks a statistical fairness constraint (say, equalizing false positive rates across protected groups), but then asks that this constraint hold over an exponentially or infinitely large collection of *subgroups* defined by a class of functions with bounded VC dimension. They give an algorithm guaranteed to learn subject to this constraint, under the condition that it has access to oracles for perfectly learning absent a fairness constraint. In this paper, we undertake an extensive empirical evaluation of the algorithm of Kearns et al. On four real datasets for which fairness is a concern, we investigate the basic convergence of the algorithm when instantiated with fast heuristics in place of learning oracles, measure the tradeoffs between fairness and accuracy, and compare this approach with the recent algorithm of Agarwal, Beygelzeimer, Dudik, Langford, and Wallach [ICML 2018], which implements weaker and more traditional marginal fairness constraints defined by individual protected attributes. We find that in general, the Kearns et al. algorithm converges quickly, large gains in fairness can be obtained with mild costs to accuracy, and that optimizing accuracy subject only to marginal fairness leads to classifiers with substantial subgroup unfairness. We also provide a number of analyses and visualizations of the dynamics and behavior of the Kearns et al. algorithm. Overall we find this algorithm to be effective on real data, and rich subgroup fairness to be a viable notion in practice.

CCS CONCEPTS

• **Computing methodologies → Machine learning**;

KEYWORDS

Algorithmic Bias, Subgroup Fairness, Fairness Auditing, Fair Classification

ACM Reference Format:
In *FAT* '19: Conference on Fairness, Accountability, and Transparency (FAT* '19), January 29–31, 2019, Atlanta, GA, USA.*

FAT '19, January 29–31, 2019, Atlanta, GA, USA*

DOI: 10.1145/3287560.3287592

1 INTRODUCTION

The most common definitions of fairness in machine learning are statistical in nature. They proceed by fixing a small number of "protected subgroups" (such as racial or gender groups), and then ask that some statistic of interest be approximately equalized across groups. Standard choices for these statistics include positive classification rates [3], false positive or false negative rates [4, 8, 13] and positive predictive value [4, 13] — see [2] for more examples. These definitions are pervasive in large part because they are easy to check, although there are interesting computational challenges in learning subject to these constraints in the worst case — see e.g. [16].

Unfortunately, these statistical definitions are not very meaningful to individuals: because they are constraints only over *averages* taken over large populations, they promise essentially nothing about how an individual person will be treated. Dwork et al. [7] enumerate a "catalogue of evils" which show how definitions of this sort can fail to provide meaningful guarantees. Kearns et al. [10] identify a particularly troubling failure of standard statistical definitions of fairness, which can arise naturally without malicious intent, called "fairness gerrymandering". They illustrate the idea with the following toy example shown in Figure 1, described as follows.

Suppose individuals each have two sensitive attributes: race (say blue and green) and gender (say male and female). Suppose that these two attributes are distributed independently and uniformly at random, and are uncorrelated with a binary label that is also distributed uniformly at random. If we view gender and race as defining classes of people that we wish to protect, we could take a standard statistical fairness definition from the literature — say the equal odds condition of [8], which asks to equalize false positive rates across protected groups, and instantiate it with the four protected groups: "Men", "Women", "blue people", and "green people". The following classifier satisfies this condition, although only by

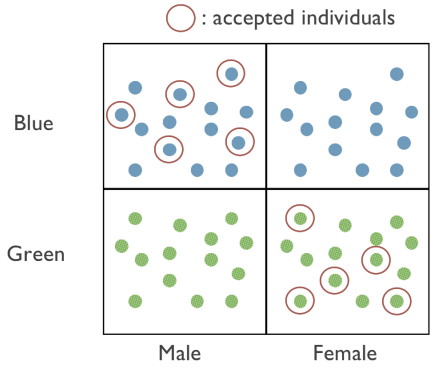

Figure 1: Fairness Gerrymandering: A Toy Example [10]

"cheating" and packing its unfairness into structured subgroups of the protected populations: it labels a person as positive only if they are a blue man or a green woman. This equalizes false positive rates across the four specified groups, but of course not over the finer-grained subgroups defined by the intersections of the two protected attributes.

Kearns et al. [10] also proposed an approach to the problem of fairness gerrymandering: rather than asking for statistical definitions of fairness that hold over a small number of coarsely defined groups, ask for them to hold over a combinatorially or infinitely large collection of subgroups defined by a set of functions $\mathcal{G}$ of the protected attributes (Hébert-Johnson et al. [9] independently made a similar proposal). For example, we could ask to equalize false positive rates across every subgroup that can be defined as the intersection or conjunction of d protected attributes, for which there are 2^d such groups. Kearns et al. [10] showed that as long as the class of functions defining these subgroups has bounded VC dimension, the statistical learning problem of finding the best (distribution over) classifiers in $\mathcal{H}$ subject to the constraint of equalizing the positive classification rate, the false positive rate, or the false negative rate over every subgroup defined over $\mathcal{G}$ is solvable whenever the dataset size is sufficiently large relative to the VC dimension of $\mathcal{G}$ and $\mathcal{H}$. Taking inspiration from the technique of Agarwal et al. [1], they were able to show that even with combinatorially many subgroup fairness constraints, the computational problem of learning the optimal fair classifier is once again solvable efficiently whenever the learner has access to a black-box classifier (oracle) which can solve the *unconstrained* learning problems over $\mathcal{G}$ and $\mathcal{H}$ respectively. Similarly, given access to an oracle for $\mathcal{G}$, they were able to efficiently solve the problem of *auditing* for rich subgroup fairness: finding the $g \in \mathcal{G}$ that corresponds to the subgroup for whom the statistical fairness constraint was most violated.

While the work of Kearns et al. [10] is satisfying from a theocratical point of view, it leaves open a number of pressing empirical questions. For example, their theory is built for an idealized setting with perfect learning oracles — in practice heuristic oracles may fail. Moreover, perhaps rich subgroup fairness is asking for too much in practice — maybe enforcing combinatorially many constraints leads to an untenable tradeoff with error. Finally, perhaps enforcing combinatorially many constraints is not necessary — perhaps on real data, it is enough to call upon the algorithm of [1] for enforcing statistical fairness constraints on the small number of groups defined by the marginal protected attributes, and rich subgroup fairness will follow incidentally. Put another way: Is the so-called *fairness gerrymandering* problem only a theoretical curiosity, or does it arise organically when standard classifiers are optimized subject to marginal statistical fairness constraints?

In this paper, we conduct an extensive set of experiments to answer these questions. We study the algorithm from [10] — instantiated with fast heuristic learning oracles — when used to train a linear classifier subject to approximately equalizing false positive rates across a rich set of subgroups defined by linear threshold functions. On four real datasets, we characterize:

(1) The basic convergence properties of the algorithm — although this algorithm has provable guarantees when instantiated with learning *oracles* for $\mathcal{G}$ and $\mathcal{H}$, when these oracles are (necessarily) replaced with heuristics, the guarantees of the algorithm become heuristic as well. We find that the algorithm typically converges (Subsection 3.2), and provides a controllable trade-off between fairness and accuracy despite its heuristic guarantees (Subsection 3.3). We visualize the optimization trajectory of the algorithm (Subsection 3.5), and *discrimination heatmaps* showing the evolution of the subgroup discrimination of the algorithm over time (Subsection 3.4).
(2) The trade-off between subgroup fairness and accuracy. We find that for each dataset, there are appealing compromises between error and subgroup fairness. Thus achieving rich subgroup fairness may be possible in practice without a severe loss in predictive accuracy (Subsection 3.3).
(3) The subgroup (unfairness) that can result when one applies more standard approaches, that either ignore fairness constraints all together, or equalize false positive rates only across a small number of subgroups defined by individual protected attributes. By *auditing* the models produced by these standard approaches with the rich subgroup auditor of [10], we find that often subgroup fairness constraints are violated, even by algorithms which are explicitly equalizing false positive rates across the groups defined on the marginal protected attributes.

In light of these findings, we submit that rich subgroup fairness constraints are both important, and can be satisfied at reasonable cost: both in terms of computation, and in terms of accuracy. We hope that algorithms like that of [10] which can be used to satisfy rich subgroup fairness become part of the standard toolkit for fair machine learning.

1.1 Further Related Work

While Kearns at al. [10] propose and study rich sub-group fairness for false positive and negative constraints, Hébert-Johnson et al. study the analogous notion for *calibration* constraints, which they call *multi-calibration* [9]. Kim et al. extend this style of analysis to *accuracy* constraints (asking that a classifier be equally accurate on a combinatorially large collection of subgroups) [11]. Kim et al. also extend it to metric fairness constraints [12], converting

the *individual* metric fairness constraint of [7] into a statistical constraint that asks that *on average*, individuals in (combinatorially many) subgroups should be treated differently only in proportion to the average difference between individuals in the subgroups, as measured with respect to some similarity metric.

2 DEFINITIONS

We begin with some definitions, following the notation in [10]. We study the classification of individuals defined by a tuple $((x, x'), y)$, where $x \in \mathcal{X}$ denotes a vector of *protected attributes*, $x' \in \mathcal{X}'$ denotes a vector of *unprotected attributes*, and $y \in \{0, 1\}$ denotes a label. We will write $X = (x, x')$ to denote the joint feature vector. We assume that points (X, y) are drawn i.i.d. from an unknown distribution $\mathcal{P}$. Let D be a binary classifier, and let $D(X) \in \{0, 1\}$ denote the (possibly randomized) classification induced by D on individual (X, y).

We will be concerned with learning and auditing classifiers D satisfying a common statistical fairness constraint: equality of false positive rates (also known as equal opportunity). The techniques in Agarwal et al. [1] and Kearns et al. [10] also apply equally well to equality of false negative rates and equality of classification rates (also known as statistical parity).[1]

Each fairness constraint is defined with respect to a set of protected groups. We define sets of protected groups via a family of indicator functions $\mathcal{G}$ for those groups, defined over protected attributes. Each $g : \mathcal{X} \to \{0, 1\} \in \mathcal{G}$ has the semantics that $g(x) = 1$ indicates that an individual with protected features x is in group g. We now formally define false positive subgroup fairness.

Definition 2.1 (False Positive Subgroup Fairness)**.** Fix any classifier D, distribution $\mathcal{P}$, collection of group indicators $\mathcal{G}$, and parameter $\gamma \in [0, 1]$. For each $g \in \mathcal{G}$, define

$$\alpha_{FP}(g, \mathcal{P}) = \Pr_{\mathcal{P}}[g(x) = 1, y = 0],$$

$$\beta_{FP}(g, D, \mathcal{P}) = |\mathrm{FP}(D) - \mathrm{FP}(D, g)|$$

where $\mathrm{FP}(D) = \Pr_{D,\mathcal{P}}[D(X) = 1 \mid y = 0]$ and $\mathrm{FP}(D, g) = \Pr_{D,\mathcal{P}}[D(X) = 1 \mid g(x) = 1, y = 0]$ denote the overall false-positive rate of D and the false-positive rate of D on group g respectively.

We say D satisfies γ-*False Positive (FP) Fairness* with respect to $\mathcal{P}$ and $\mathcal{G}$ if for every $g \in \mathcal{G}$

$$\alpha_{FP}(g, \mathcal{P}) \cdot \beta_{FP}(g, D, \mathcal{P}) \leq \gamma.$$

We will sometimes refer to $\mathrm{FP}(D)$ FP-base rate.

Since we do not consider other measures in this paper, we refer to this notion as simply "subgroup fairness." Given a fixed subgroup $g \in \mathcal{G}$ we will refer to the quantity $\alpha_{FP}(g, \mathcal{P}) \cdot \beta_{FP}(g, D, \mathcal{P})$ as the subgroup fairness wrt g, or alternately the γ-unfairness of g. The notion of subgroup fairness imposes a statistical constraint on combinatorially many groups definable by the protected attributes. This is in contrast to more common statistical fairness definitions, defined on coarse groups definable by a single protected attribute. Given a protected attribute x_i and a value for that attribute a, define the function $g_{i,a}(x) = \mathbf{1}\{x_i = a\}$ denoting the set of individuals who have that particular value of their protected attribute. In contrast to subgroup fairness, we refer to a classifier D as *marginally* fair if it satisfies false positive subgroup fairness with respect to the functions $\{g_{i,a}\}$ for each protected attribute x_i and realization a.

If the algorithm D fails to satisfy the γ-subgroup fairness condition, then we say that D is γ-*unfair* with respect to $\mathcal{P}$ and $\mathcal{G}$. We call any subgroup g which witnesses this unfairness a γ-*unfair certificate* for $(D, \mathcal{P})$.

An *auditing algorithm* for a notion of fairness is given sample access to points from the underlying distribution, as well as the classification outcomes provided by D. It will either deem D to be fair with respect to $\mathcal{P}$, or else produces a certificate of unfairness.

The algorithms of Agarwal et al. [1] and Kearns et al. [10] studied in this paper both assume access to oracles which can solve *cost-sensitive classification (CSC)* problems. Formally, an instance of a CSC problem for the class $\mathcal{H}$ is given by a set of n tuples $\{(X_i, c_i^0, c_i^1)\}_{i=1}^n$ such that c_i^ℓ corresponds to the cost for predicting label ℓ on point X_i. Given such an instance as input, a CSC oracle finds a hypothesis $\hat{h} \in \mathcal{H}$ that minimizes the total cost across all points:

$$\hat{h} \in \operatorname*{argmin}_{h \in \mathcal{H}} \sum_{i=1}^{n} [h(X_i)c_i^1 + (1 - h(X_i))c_i^0] \tag{1}$$

Following both [1] and Kearns et al. [10], in all of the experiments in this paper we take the classes $\mathcal{H}$ and $\mathcal{G}$ to be linear threshold functions, and we use a linear regression heuristic for both auditing and learning. The heuristic finds a linear threshold function as follows:

- Train two linear regression models r_0, r_1 to predict c_0 and c_1 respectively.
- Given a new point x, predict the cost of classifying x as 0 and 1 using our regression models: these are $r_0(x)$ and $r_1(x)$ respectively.
- Output the prediction $\hat{y}$ corresponding to lower predicted cost: $\hat{y} = \operatorname{argmin}_{i \in \{0,1\}} r_i(x)$.

We leave the precise descriptions of the algorithm from [10] — which we will refer to as the SUBGROUP algorithm — to the appendix. We refer the reader to [10] for details about its derivation and guarantees.[2] At this point we remark only that the algorithm operates by expressing the optimization problem to be solved (minimize error, subject to subgroup fairness constraints) as solving for the equilibrium in a two player zero-sum game, between a *Learner* and an *Auditor*. The *Learner* has the set of hypothesis $\mathcal{H}$ as its action (pure strategy) space, and the *Auditor* has the set of subgroups $\mathcal{G}$ as its action space. The best response problem for the Auditor corresponds to the auditing problem: finding the subgroup $g \in \mathcal{G}$ for which the strategy of the learner violates the fairness constraints the most. The best response problem for the Learner corresponds

[1] or more generally to any fairness constraint that can be expressed as a linear equality on the conditional moments $\mathbb{E}\left[t(X, y, D(X) | \varepsilon(X, y))\right]$, where $\varepsilon(X, y)$ is an event defined with respect to (X, y), and $t : X \times \{0, 1\} \times \{0, 1\} \to [0, 1]$ [1]. Equality of false positive rate is a particular instantiation of this kind of constraint where ε is the event $y = 0$, and $t = \mathbf{1}\{D(X) = 1\}$.

[2] [10] actually give two algorithms, one of which employs no-regret learning techniques and converges in a polynomial number of rounds, but is randomized; and the other of which is known to converge only in the limit (but is conjectured to converge quickly), and is deterministic. We focus on the deterministic algorithm in this paper, because it is more amenable to implementation, despite its weaker theoretical guarantees. We find that it performs well in practice despite its weaker theory.

to solving a weighted (but unconstrained) empirical risk minimization problem. The best response problem for both players can be expressed as solving a cost sensitive classification problem. The algorithm SUBGROUP essentially simulates the *fictitious play* of this game, which proceeds over rounds, and in each round t both players best respond to their opponent's empirical history of play:

- Learner plays h_t in $\mathcal{H}$ that minimizes objective function balancing error and unfairness on subgroups $g_1, \ldots, g_{t-1}$ found by Auditor so far;
- Auditor finds subgroup g_t in $\mathcal{G}$ on which the uniform distribution over $h_1, \ldots, h_t$ violates γ-fairness the most.

This can be done efficiently assuming access to oracles which solve the cost sensitive classification problem over $\mathcal{G}$ and $\mathcal{H}$ respectively.

3 EMPIRICAL EVALUATION

In this section, we describe an extensive empirical investigation of the SUBGROUP algorithm on four datasets in which fairness is a potential concern. Among the questions of primary interest are the following:

- Does the SUBGROUP algorithm work in practice, despite the use of imperfect heuristics for the Learner and Auditor?
- Is the notion of subgroup fairness interesting empirically, in that there are palatable trade-offs between accuracy and subgroup fairness (as opposed to it being too strong a constraint, and thus resulting in a very steep error increase for even weak subgroup fairness)?

We will answer these questions strongly in the affirmative, which is perhaps the overarching message of our results. We also carefully compare subgroup fairness to standard marginal fairness, and show that optimizing for the latter in general does poorly on the former — thus something like the SUBGROUP algorithm is actually necessary to achieve subgroup fairness.

More generally, aside from performance, we provide a number of empirical analyses that elucidate the underlying behavior and convergence properties of the SUBGROUP algorithm, and discuss its strengths and weaknesses.

3.1 Datasets

We ran experiments on 3 datasets from the UCI Machine Learning Repository [6]: **Communities and Crime** [14], **Adult**, and **Student** [5], and the **Law School** dataset from the Law School Admission Council's National Longitudinal Bar Passage Study [15]. These datasets were selected due to their potential fairness concerns, including:

- Data points representing individual people (or in the case of Communities and Crimes, small U.S. communities of people);
- The presence of features capturing properties often associated with possible discrimination, including race, gender, and age;
- Potential sensitivity of the predictions being made, such as violent crime, income, or performance in school.

The properties of these datasets are summarized in Table 1, including the number of instances, the prediction being made, the overall number of features (which varies from 10 to 128), the number of protected features in the subgroup class (which varies from 3 to 18), the nature of the protected features, and the baseline (majority class) error rate.

Some methodological notes:

- We note that two of the datasets (Law School and Adult) were initially much larger but were extremely imbalanced with respect to the predicted label, making sensible error comparisons numerically difficult. We thus randomly downsampled these two datasets to obtain approximately balanced prediction problems on each.
- All categorical variables have been preprocessed with a one-hot encoding.
- The SUBGROUP algorithm has two input parameters: the maximum allowed subgroup fairness violation. γ, and a tuning parameter C which represents (in the theoretical derivation in [10]) an upper bound on the magnitude of the dual variables needed to express the fairness constrained empirical risk minimization problem. We view γ as an important control variable allowing us to explore the tradeoff between fairness and accuracy, and thus will vary it in our experiments. On the other hand, C is more of a nuisance parameter, and thus for consistency and simplicity we set $C = 10$ in all experiments. Experimentation with larger values of C did not reveal qualitatively different findings on the datasets investigated.
- We emphasize that *all results are reported in-sample on the datasets*, and thus we are treating the empirical distributions of the datasets as the "true" distributions of interest. We do this because our primary interest is simply in examining the performance and behavior of the SUBGROUP algorithm on the actual data or distributions, and not in generalization per se. As noted in [10], theoretical generalization bounds for both error and subgroup fairness can be obtained by standard methods, and will depend on (e.g.) the VC dimension of the Learner's model class $\mathcal{H}$ and the Auditor's subgroup class $\mathcal{G}$. As usual, we would expect empirical generalization to often be considerably better than the worst-case theory.

3.2 Empirical Convergence of SUBGROUP

We begin with an examination of the convergence properties of the SUBGROUP algorithm on the four datasets. Kearns et al. [10] had already reported preliminary convergence results for the Communities and Crime dataset, showing that their algorithm converges quickly, and that varying the input γ provides an appealing trade-off between error and fairness. In addition to replicating those findings for Communities and Crime, we also find that they are not an optimistic anomaly. For example, for the Law School dataset, in Figure 2 we plot both the error ε_t (panel (a)) and the fairness violation γ_t (panel (b)) as a function of the iteration t, for values of the input γ ranging from 0 to 0.03. We see that the algorithm converges relatively quickly (on the order of thousands of iterations), and that increasing the input γ generally yields decreasing error and increasing fairness violation (typically saturating the input γ), as suggested by the idealized theory.

But on other datasets the empirical convergence does not match the idealized theory as cleanly, presumably due to the use of imperfect Learner and Auditor heuristics. In panels (c) and (d) of Figure 2

Dataset	Size	Prediction	#Features	#Protected	Protected Feature Types	Baseline
Communities and Crime	1994	High Violent Crime ?	128	18	Race	0.3
Law School	2053	Pass Bar Exam ?	10	4	Race, Income, Age, Gender	0.49
Student	396	Course Performance ?	30	5	Age, Gender, Relationship, Alcohol Use	0.47
Adult	2021	Income >= $50K ?	14	3	Age, Race, Gender	0.50

Table 1: Description of Data Sets.

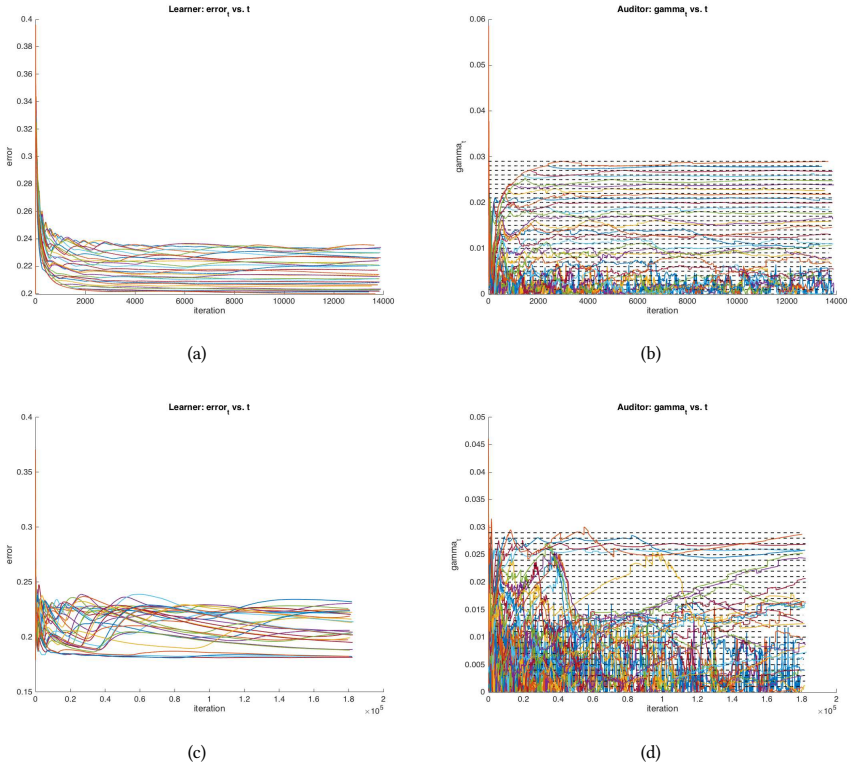

Figure 2: Error ε_t and fairness violation γ_t for Law School dataset (panels (a) and (b)) and Adult data set (panels (c) and (d)), for values of input γ ranging from 0 to 0.03. Dashed horizontal lines on γ_t plots correspond to varying values of γ.

we again plot ε_t and γ_t, but now for the Adult dataset. Even after approximately 180,000 iterations, the algorithm does not appear to have converged, with ε_t still showing long-term oscillatory behavior, γ_t exhibiting extremely noisy dynamics (especially at smaller input γ values), and there being no clear systematic, monotonic relationship between the input γ and error acheived. But despite this departure from the theory, it remains the case that varying γ still yields a diverse set of $\langle \varepsilon_t, \gamma_t \rangle$ pairs, as we will see in the next section. In this sense, even in the absence of convergence the algorithm can be viewed as a valuable search tool for models trading off accuracy and fairness.

Overall, we found rather similar convergent behavior on the Communities and Crime and Law School datasets, and less convergent behavior on the Adult and Student datasets.

3.3 Subgroup Pareto Frontiers and Comparison to Marginal Fairness

Regardless of convergence, for plots such as those in Figure 2, it is natural to take the $\langle \varepsilon_t, \gamma_t \rangle$ pairs across all t and all input γ, and compute the undominated or Pareto frontier of these pairs. This frontier represents the accuracy-fairness tradeoff achieved by the SUBGROUP algorithm on a given data set, which is arguably its most important output. The choice of where one wants to be on the frontier is a policy question that should be made by domain experts and stakeholders, and dependent on the stakes involved (e.g. online advertising vs. criminal sentencing).

It is also of interest to compare the subgroup fairness achieved by the SUBGROUP algorithm (which is explicitly optimizing under a subgroup fairness constraint) with an algorithm only optimizing under weaker and more traditional marginal fairness constraints. To this end, we also implemented a version of the algorithm from

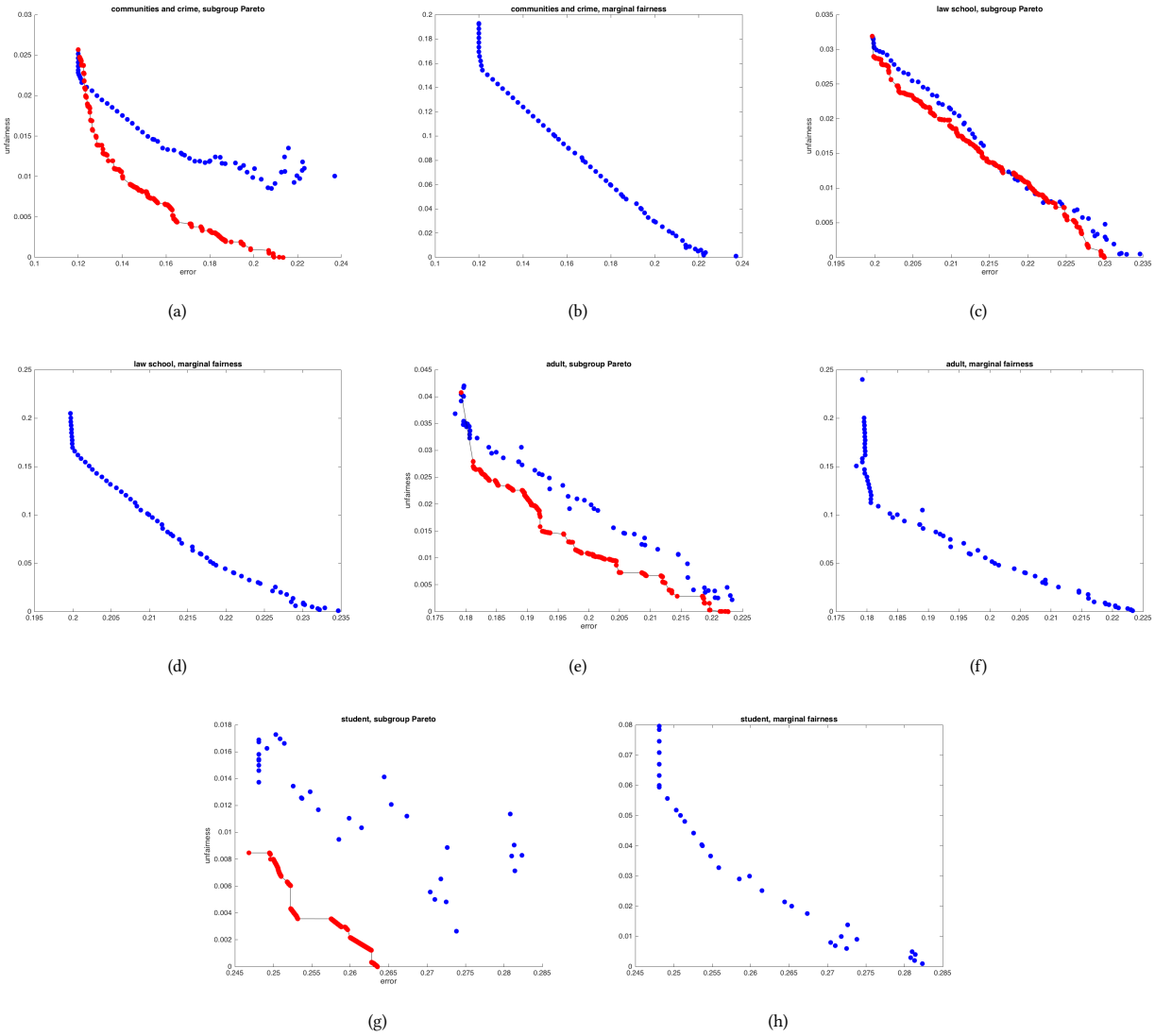

Figure 3: Left column: The red points show the Pareto frontier of error (x axis) and subgroup fairness violation (y axis) for the SUBGROUP **algorithm across all four data sets, while the blue points show the error and subgroup fairness violation for the models achieved by the** MARGINAL **algorithm. Right column: The error and marginal fairness violation for the** MARGINAL **algorithm across all four data sets. Ordering of datasets is Communities and Crime, Law School, Adult, and Student.**

[1] — which we will refer to as the MARGINAL algorithm — for marginal fairness.[3] From a theoretical perspective, *a priori* we would expect models trained for marginal fairness to fare poorly on subgroup fairness. But it is an empirical question — perhaps on some datasets, demanding marginal fairness already suffices to enforce subgroup fairness as well. Thus the high-level question is whether the SUBGROUP framework and algorithm are worth the added analytical and computational overhead.

[3]Since some of the protected attributes are continuous rather than discrete, and the MARGINAL algorithm only handles discrete attributes, in order to run the marginal fairness algorithm we create sensitive groups by thresholding on the mean of each sensitive attribute.

In the left column of Figure 3, we show the SUBGROUP algorithm Pareto frontiers for subgroup fairness on all four datasets, and also the pairs achieved by the MARGINAL algorithm. In the right column, we also separately show the marginal fairness frontier achieved by the MARGINAL algorithm. Before discussing the particulars of each dataset, we first make the following general observations:

- For most datasets, the SUBGROUP algorithm yields a Pareto curve that frequently lies well below the straight line connecting its endpoints (which we can think of as an empirical form of strong convexity), and thus there are non-trivial

tradeoffs between accuracy and fairness to consider. On some of these curves there are regions of steep descent where subgroup unfairness can be reduced significantly with negligible increase in error.

- While the MARGINAL algorithm performs well with respect to marginal fairness (right column) as expected, it fares much worse than the SUBGROUP algorithm on subgroup fairness for three of the datasets. Thus marginal fairness is not just theoretically, but also empirically a weaker notion, and generally will not imply subgroup fairness "for free".
- Nevertheless, there are a handful of points in which the MARGINAL algorithm produces models that actually lie below (and thus dominate) the SUBGROUP Pareto curve by a small amount. While this is not possible under the idealized theory — subgroup fairness is a strictly stronger notion than marginal fairness — it can again be explained by the use of imperfect learning heuristics by both algorithms.
- Focusing just on the MARGINAL marginal fairness curves in the right column, we see that each of them begins with a steep drop, meaning that in every case, the marginal unfairness of the unconstrained error-optimal model can be significantly improved with little or no increase in error.
- By matching points between the MARGINAL marginal and subgroup fairness plots, we find that with the exception of the Student data set, there is a systematic relationship between marginal and subgroup unfairness: asking the MARGINAL algorithm to reduce marginal unfairness also causes it to reduce subgroup unfairness — but not by as much as the SUBGROUP algorithm achieves.

Together these observations let us conclude that subgroup fairness is a strong but achievable notion in practice (at least on these datasets), and that the SUBGROUP algorithm appears to be an effective tool for its investigation.

It is also worth commenting on the differences across datasets, and focusing not just on the qualitative shapes of the Pareto curves but their actual numerical specifics — especially since in real applications, these will matter to stakeholders. For instance, the actual range of error values spanned by the SUBGROUP Pareto curves ranges from nearly 10% (Communities and Crime) to less than 2% (Student). So perhaps for Communities and Crime, the tradeoff is starker from an accuracy perspective. We now provide some brief commentary on each dataset.

Communities and Crime (panels (a) and (b)): This is the dataset with perhaps the cleanest and most convex SUBGROUP Pareto curve, with steep drops in subgroup unfairness possible for minimal error increase at the beginning. In particular are able to reduce the initial γ-unfairness from 0.026 to less than 0.005 while only increasing the error from 0.12 to 0.16. This is a meaningful reduction in unfairness – e.g. reducing a 26% percent difference in false positive rate on a subgroup comprising 10% of the population, to a less than 5% false positive rate disparity on a subgroup of the same size. Eventually the Pareto curve flattens out, resulting in increasing accuracy costs for reduced unfairness. While the MARGINAL subgroup unfairness curve matches the SUBGROUP Pareto curve on the far left (for all datasets), since this corresponds to minimizing error unconstrained by any fairness notion, the outperformance by SUBGROUP grows rapidly as we make stronger fairness demands.

Law School (panels (c) and (d)): Here the SUBGROUP Pareto curve appears to be approximately linear, thus providing a constant tradeoff between accuracy and subgroup fairness. Interestingly, this is the one dataset in which asking for marginal fairness appears to also yield subgroup fairness for free, as the MARGINAL curve lies very close to the SUBGROUP curve. Since this dataset has the fewest number of features overall and the second fewest number of protected features, one might be tempted to conjecture that when the number of protected features is small, guaranteeing marginal fairness approximately guarantees rich subgroup fairness. This claim is falsified by the fact that on the Adult dataset which has similar dimensionality (see below), there is a large gap between the SUBGROUP and MARGINAL subgroup fairness curves.

Adult (panels (e) and (f)): Here we see a less smooth SUBGROUP curve, possibly corresponding to the poorer convergence properties on this dataset mentioned earlier. Nevertheless, the numerical tradeoff exhibits regions of both steep, inexpensive reduction in unfairness and flat, costly reduction. MARGINAL is again considerably worse when evaluated on subgroup fairness, but still shows a systematic relationship to marginal fairness.

Student (panels (g) and (h)): Similar to Adult, a varied SUBGROUP curve with multiple tradeoff regimes. This is also the lone dataset in which reducing marginal fairness appears to have no relationship to subgroup fairness — while the MARGINAL marginal pareto curve in panel (h) remains relatively smooth, the subgroup fairness of the corresponding models in panel (d) is now not only worse than for SUBGROUP, but shows no monotonicity. SUBGROUP is able to decrease γ-unfairness to 0 with only a 2% increase in error, while the MARGINAL algorithm only drives the subgroup unfairness to 0.002 at its best, with an over 3% increase in error from the unconstrained classifier.

Having established the efficacy of subgroup fairness and the SUBGROUP algorithm on the four datasets, we now turn to experiments and visualizations allowing us to better understand the behavior and dynamics of the algorithm.

3.4 Flattening the Discrimination Surface

Recall that in the various analyses and plots above, we rely on the Auditor of SUBGROUP to detect unfairness. This Auditor is in turn a heuristic, relying on an optimization procedure without any theoretical guarantees, which could potentially fail in practice. This means that while any detected unfairness is a lower bound on the true subgroup unfairness, it could be the case that the heuristic Auditor is simply failing to detect a larger disparity, and that the models learned by SUBGROUP look more fair than they really are.

We explore this possibility on the Communities & Crime dataset by implementing a brute force Auditor that runs alongside the SUBGROUP algorithm. To make brute force auditing computationally tractable, we designate only two attributes as protected;

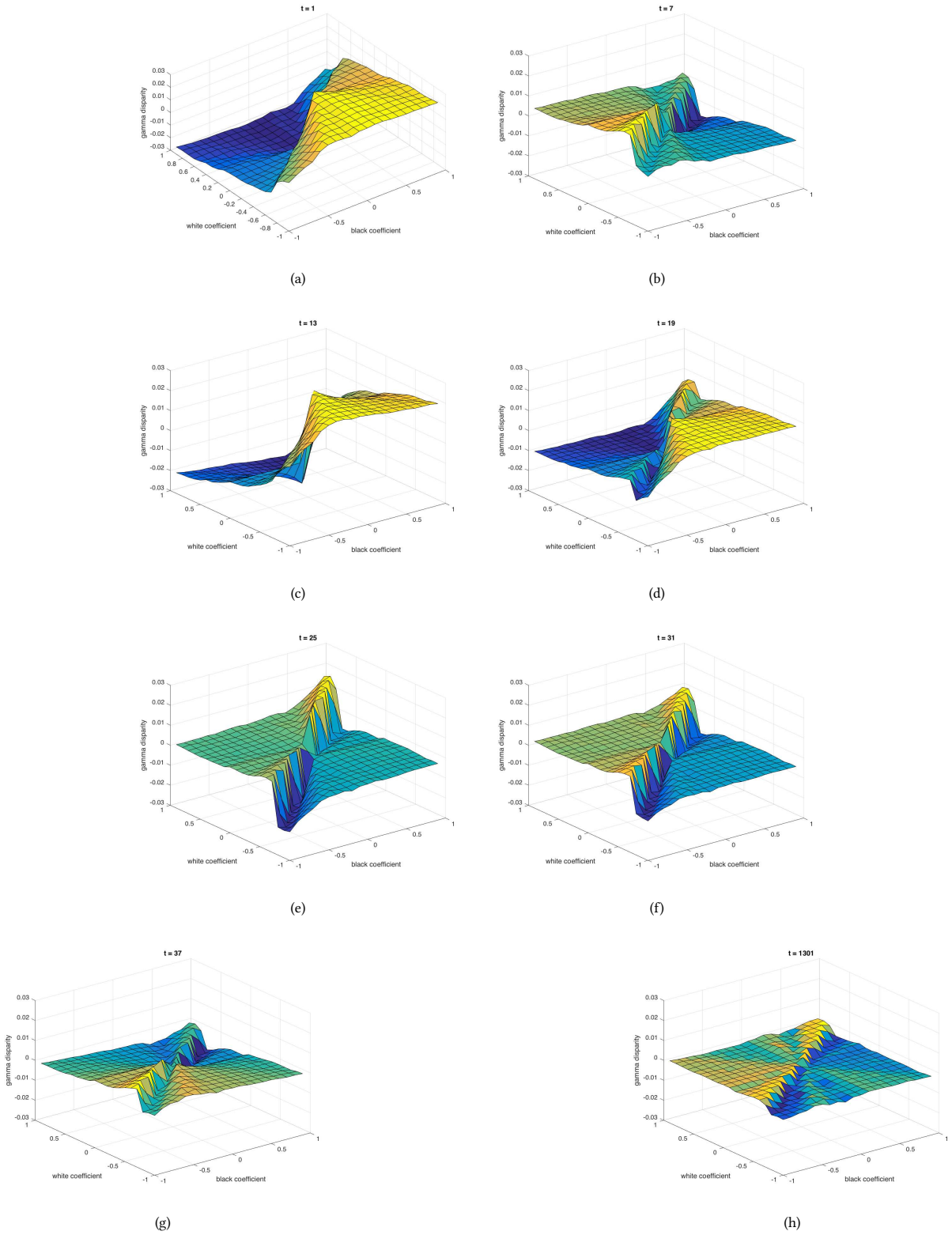

Figure 4: Evolution of discrimination surface for the SUBGROUP **algorithm from** $t = 1 \ldots 1301$**. Each point in the plane corresponds to a different subgroup over two protected attributes, and the corresponding** z **value is the current false positive discrepancy for the subgroup.**

pctwhite and *pctblack*, the percentage of each community that consists of white and black people respectively. While the SUBGROUP algorithm uses the same heuristic Auditor it always does, at each round we also perform a brute force audit as follows. Subgroups g_θ are defined by a linear threshold function θ over the 2 sensitive attributes, e.g. $(x_1, x_2) \in g_\theta$ iff $\langle \theta, (x_1, x_2) \rangle \geq 0$. We discretize $\theta \in [-1, 1]^2$ in increments of 0.1, and for the subgroup defined by each θ in the discretization we compute the γ-unfairness. Hence at each round we can take the current classifier of the Learner, and plot for each group g_θ the point $(\theta_1, \theta_2, \gamma)$.

Note that in addition to making brute force auditing tractable, restricting to two dimensions permits direct visualization of discrimination. In Figure 4, we show a sequence of "discrimination surfaces" for the SUBGROUP algorithm over the 2 protected features, with input $\gamma = 0$. The $x - y$ axes are the coefficients of θ corresponding to *whitepct* and *blackpct* respectively, and the z-axis is the γ-unfairness of the corresponding subgroup. This is our first non-heuristic view of γ-unfairness, and also shows us the entire surface of γ-unfairness, rather than just the most violated subgroup. Note that perfect subgroup fairness would correspond to an entirely flat discrimination surface at $z = 0$.

We observe first that the unconstrained classifier in $t = 1$ (panel (a)) shows a very systematic bias along the lines of our sensitive attributes. In particular groups with *whitepct* > 0 and *blackpct* < 0, e.g. communities with large numbers of white residents and relatively fewer black residents have a much higher false positive rate for being classified as violent. Conversely, majority black communities are less likely to be incorrectly labeled as violent. The mean γ-unfairness (base rate - community rate) for *whitepct* > 0, *blackpct* < 0 communities is -0.0242, whereas the mean for *whitepct* < 0, *blackpct* > 0 groups is 0.0247. The maximum γ-unfairness in $t = 1$ is 0.028, and 61.25% of the 400 subgroups have γ-unfairness > 0.02. Recall that this corresponds to e.g. a 20% disparity of the false positive rate from the base rate, for groups as large as 10% of the population. We are thus far from perfect subgroup fairness.

As the algorithm proceeds, we see this discrimination flip by $t = 7$ (panel (b)), into a regime with a higher false positive rate for predominantly black communities, and then revert again by $t = 13$. Over the early iterations these oscillations continue, growing less drastic as the γ-unfairness surface starts to flatten out noticeably by $t = 37$ (panel (g)). In panel (h) we plot $t = 1301$ and see that the surface has almost completely flattened, with maximum γ-unfairness below .0028. So over the course of the first 1300 iterations of SUBGROUP we've reduced the γ-unfairness from over 0.02 in most of the subgroups, to less than 0.0028 in *every* subgroup. Recall again that this corresponds to false positive rate disparities of at most 2.8% in subgroups that represent 10% of the population — a reduction from false positive rate disparities of 20% many similarly sized subgroups. This represents an order of magnitude improvement that results from using the classifier learned by SUBGROUP.

3.5 Understanding the Dynamics

We conclude by examining the dynamics of the SUBGROUP algorithm on the Communities and Crime dataset in greater detail. More specifically, since the algorithm is formulated as a game between a Learner who at each iteration t is trying to minimize the error ε_t, and an Auditor who is trying to minimize subgroup unfairness γ_t, we visualize the trajectories traced in $\langle \varepsilon_t, \gamma_t \rangle$ space as t increases.

The plots in Figure 5 correspond to such trajectories for input γ values of 0.001, 0.005, 0.009, and 0.022 (panels (a), (b), (c) and (d) respectively), which are denoted by the dashed lines on the γ_t axis of each figure. The 0.001 and 0.005, 0.009 values correspond to small and intermediate γ regimes, whereas 0.022 is close to (but slightly below) the subgroup unfairness of the unconstrained classifier. The trajectories are color coded from colder to warmer colors according to their iteration number to give a sense of speed of convergence.

The first plot in all four trajectories corresponds to the $\langle \varepsilon_0, \gamma_0 \rangle$ of the unconstrained classifier. Furthermore, as long as the current γ_t values remain above the horizontal dashed line representing the input γ, the trajectories remain identical, as the same subgroups are being presented to the learner in each trajectory. But when γ_t falls below a given input γ, that trajectory will follow its own path going forward.

We first observe that the dynamics exhibit a fair amount of complexity and subtlety. They all begin with low error and large unfairness, and quickly follow a brief but large increase in ε_t as fairness starts to be enforced. There are steps in which both ε_t and γ_t increase, and a large early loop in trajectory space is observed. But the first three trajectories (panels (a), (b) and (c), corresponding to the three smaller values of γ) quickly settle near the input γ line, at which point begins a long, oscillatory "border war" around this line, as the Learner tries to minimize error, but is pushed back below the line by the Auditor anytime γ-fairness is violated. The idealized theory predicts that each trajectory should end at the input γ line (subgroup fairness constraint saturated), and with larger input γ (weaker fairness constraint) resulting in lower error. The empirical trajectories indeed conform nicely to the theory, with the final (red) points near the dashed lines, and further left for larger γ.

Panel (d), corresponding to a much larger input γ, diverges much earlier from the other three (on its second step), and early on sees unfairness driven far below the specified value. The dynamics then see a slow, gradual decrease of error and increase of unfairness back to the input value, with the trajectory ending up near where it began, but just slightly more fair, as specified by γ.

4 CONCLUSIONS

In this work we have established the empirical efficacy of the notion of rich subgroup fairness and the algorithm of [10] on four fairness-sensitive datasets, and the necessity of explicitly enforcing subgroup (as opposed to only marginal) fairness. There are a number of interesting directions for further experimental work we plan to pursue, including:

- Experiments with richer Learner model classes $\mathcal{H}$, while keeping the Auditor subgroup class $\mathcal{G}$ relatively simple and fixed. One conjecture is that by making the hypothesis space richer, more appealing Pareto curves may be achieved. There is also some rationale for keeping $\mathcal{G}$ simple, since we would like to have some intuitive interpretation of what the subgroups represent, while the same constraint may not hold for $\mathcal{H}$.

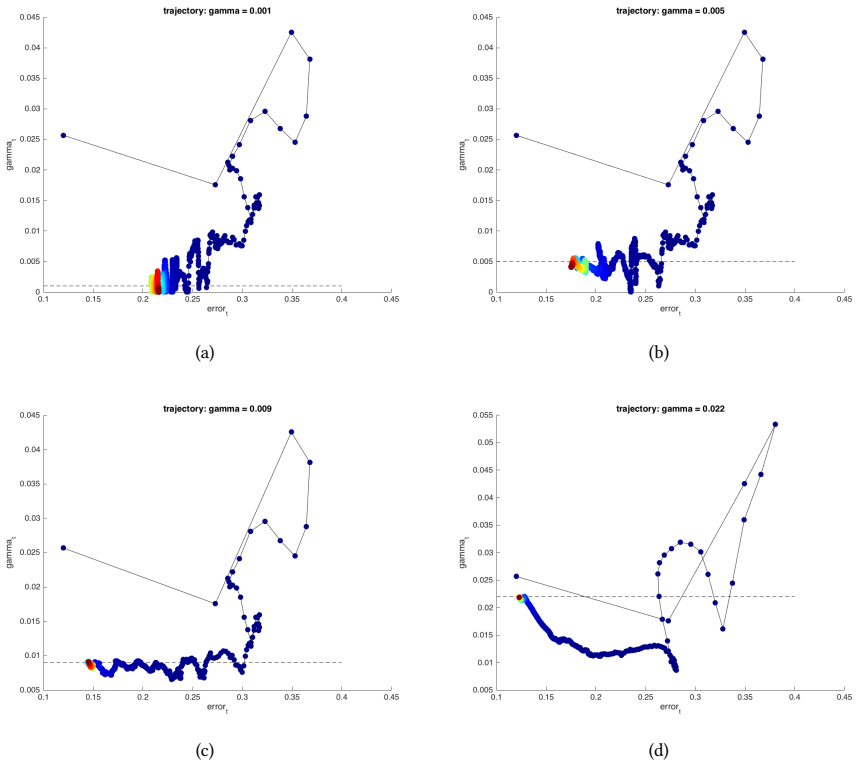

Figure 5: $\langle \varepsilon_t, \gamma_t \rangle$ trajectories for Communities and Crime, for $\gamma \in \{0.001, 0.005, 0.009, 0.022\}$.

- Implementation and experimentation with the no-regret algorithm of [10], which may have superior convergence and other properties due to its stronger theoretical guarantees.
- Experiments on the generalization performance of subgroup fairness in the form of test-set Pareto curves. While as mentioned, standard VC theory can be applied to obtain worst-case bounds, one might expect even better empirical generalization.

REFERENCES

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna M. Wallach. A reductions approach to fair classification. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *JMLR Workshop and Conference Proceedings*, pages 60–69. JMLR.org, 2018. URL http://proceedings.mlr.press/v80/agarwal18a.html.

[2] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 0(0):0049124118782533, 0. doi: 10.1177/0049124118782533. URL https://doi.org/10.1177/0049124118782533.

[3] Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.

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

[5] P. Cortez and A. Silva. Using data mining to predict secondary school student performance. *Proceedings of 5th Future Business Technology Conference*, 2008.

[6] Dua Dheeru and Efi Karra Taniskidou. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

[7] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM, 2012.

[8] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[9] Úrsula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 1944–1953, 2018. URL http://proceedings.mlr.press/v80/hebert-johnson18a.html.

[10] Michael J. Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *JMLR Workshop and Conference Proceedings*, pages 2569–2577. JMLR.org, 2018. URL http://proceedings.mlr.press/v80/kearns18a.html.

[11] Michael P Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. *arXiv preprint arXiv:1805.12317*, 2018.

[12] Michael P Kim, Omer Reingold, and Guy N Rothblum. Fairness through computationally-bounded awareness. *arXiv preprint arXiv:1803.03239*, 2018.

[13] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference, ITCS*, 2017.

[14] M.A. Redmond and A. Baveja. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research 14*, 2002.

[15] L. Wightman. Lsac national longitudinal bar passage study. 1998.

[16] Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. In *Conference on Learning Theory*, pages 1920–1953, 2017.