Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness

Michael Kearns[1], Seth Neel[1], Aaron Roth[1] and Zhiwei Steven Wu[2]

[1]University of Pennsylvania
[2]Microsoft Research

November 14, 2017

Abstract

The most prevalent notions of fairness in machine learning are *statistical* definitions: they fix a small collection of pre-defined groups, and then ask for parity of some statistic of the classifier (like classification rate or false positive rate) across these groups. Constraints of this form are susceptible to (intentional or inadvertent) *fairness gerrymandering*, in which a classifier appears to be fair on each individual group, but badly violates the fairness constraint on one or more structured *subgroups* defined over the protected attributes (such as certain combinations of protected attribute values). We propose instead to demand statistical notions of fairness across exponentially (or infinitely) many subgroups, defined by a structured class of functions over the protected attributes. This interpolates between statistical definitions of fairness, and recently proposed individual notions of fairness, but it raises several computational challenges. It is no longer clear how to even check or *audit* a fixed classifier to see if it satisfies such a strong definition of fairness. We prove that the computational problem of auditing subgroup fairness for both equality of false positive rates and statistical parity is equivalent to the problem of weak agnostic learning — which means it is computationally hard in the worst case, even for simple structured subclasses. However, it also suggests that common heuristics for learning can be applied to successfully solve the auditing problem in practice. We then derive an algorithm that provably converges to the best fair distribution over classifiers in a given class, given access to oracles which can optimally solve the agnostic learning problem and the auditing problem. The algorithm is based on a formulation of subgroup fairness as fictitious play in a two-player zero-sum game between a Learner (the primal player) and an Auditor (the dual player). Finally, we implement our algorithm using linear regression as a heuristic oracle, and show that we can effectively both audit and learn fair classifiers on real datasets.

1 Introduction

As machine learning is being deployed in increasingly consequential domains (including policing [Rudin, 2013], criminal sentencing [Barry-Jester et al., 2015], and lending [Koren, 2016]), the problem of ensuring that learned models are *fair* has become urgent.

Approaches to fairness in machine learning can coarsely be divided into two kinds: *statistical* and *individual* notions of fairness. Statistical notions typically fix a small number of protected demographic groups $\mathcal{G}$ (such as racial groups), and then ask for (approximate) parity of some statistical measure across all of these groups. One popular statistical measure asks for equality of false positive or negative rates across all groups in $\mathcal{G}$ (this is also sometimes referred to as an *equal opportunity* constraint [Hardt et al., 2016]). Another asks for equality of classification rates (also known as *statistical parity*). These statistical notions of fairness are the kinds of fairness definitions most common in the literature (see e.g. Kamiran and Calders [2012], Hajian and Domingo-Ferrer [2013], Kleinberg et al. [2017], Hardt et al. [2016], Friedler et al. [2016], Zafar et al. [2017], Chouldechova [2017]).

One main attraction of statistical definitions of fairness is that they can in principle be obtained and checked without making any assumptions about the underlying population, and hence lead to more immediately actionable algorithmic approaches. On the other hand, individual notions of fairness ask for the algorithm to satisfy some guarantee which binds at the individual, rather than group, level. This often has the semantics that "individuals who are similar" should be treated "similarly" [Dwork et al., 2012], or "less qualified individuals should not be favored over more qualified individuals" [Joseph et al., 2016]. Individual notions of fairness can have attractively strong semantics, but their main drawback is that they seemingly require more assumptions to be made about the setting under consideration.

The semantics of statistical notions of fairness would be significantly stronger if they were defined over a large number of *subgroups*, thus permitting a rich middle ground between fairness only for a small number of coarse pre-defined groups, and the strong assumptions needed for fairness at the individual level. Consider the kind of "fairness gerrymandering" that can occur when we only look for unfairness over a small number of pre-defined groups:

Example 1.1. *Imagine a setting with two binary features, corresponding to race (say black and white) and gender (say male and female), both of which are distributed independently and uniformly at random in a population. Consider a classifier that labels an example positive if and only if it corresponds to a black man, or a white woman. Then the classifier will appear to be equitable when one considers either protected attribute alone: it labels both men and women as positive 50% of the time, and labels both black and white individuals as positive 50% of the time. But if one looks at any* conjunction *of the two attributes (such at black women), then it is apparent that the classifier maximally violates the statistical parity fairness constraint. Similarly, embedding the same classification rule into the datapoints whose true label is negative results in the same example for equal opportunity fairness.*

We remark that the issue raised by this toy example is not merely hypothetical. In our experiments in Section 5, we show that similar violations of fairness on subgroups of the pre-defined groups can result from the application of standard machine learning methods applied to real datasets. To avoid such problems, we would like to be able to satisfy a fairness constraint not just for the small number of protected groups defined by single protected attributes, but for a combinatorially large or even infinite collection of structured subgroups definable over protected attributes.

In this paper, we consider the problem of *auditing* binary classifiers for equal opportunity and statistical parity, and the problem of *learning* classifiers subject to these constraints, when the number of protected groups is large. There are exponentially many ways of carving up a population into subgroups, and we cannot necessarily identify a small number of these *a priori* as the only ones we need to be concerned about. At the same time, we cannot insist on any notion of statistical fairness for *every* subgroup of the population: for example, any imperfect classifier could be accused of being unfair to the subgroup of individuals defined ex-post as the set of individuals it misclassified. This simply corresponds to "overfitting" a fairness constraint. We note that the individual fairness definition of Joseph et al. [2016] (when restricted to the binary classification setting) can be viewed as asking for equalized false positive rates across the singleton subgroups, containing just one individual each[1] — but naturally, in order to achieve this strong definition of fairness, Joseph et al. [2016] have to make structural assumptions about the form of the ground truth. It is, however, sensible to ask for fairness for large *structured* subsets of individuals: so long as these subsets have a bounded VC dimension, the *statistical* problem of learning and auditing fair classifiers is easy, so long as the dataset is sufficiently large. This can be viewed as an interpolation between equal opportunity fairness and the individual "weakly meritocratic" fairness definition from Joseph et al. [2016], that does not require making any assumptions about the ground truth. Our investigation focuses on the computational challenges, both in theory and in practice.

1.1 Our Results

Briefly, our contributions are:

- Formalization of the problem of auditing and learning classifiers for fairness with respect to rich classes of subgroups $\mathcal{G}$.
- Results proving (under certain assumptions) the computational equivalence of auditing $\mathcal{G}$ and (weak) agnostic learning of $\mathcal{G}$. While these results imply theoretical intractability of auditing for some natural classes $\mathcal{G}$, they also suggest that practical machine learning heuristics can be applied to the auditing problem.
- A provably convergent algorithm for learning classifiers that are fair with respect to $\mathcal{G}$ based on formulation as fictitious play in a two-player zero-sum game between a Learner (the primal player) and an Auditor (the dual player).
- An implementation and extensive empirical evaluation of this algorithm demonstrating its effectiveness on a real dataset in which subgroup fairness is a concern.

In more detail, we start by studying the computational challenge of simply *checking* whether a given classifier satisfies equal opportunity and statistical parity. Doing this in time linear in the number of protected groups is simple: for each protected group, we need only estimate a single expectation. However, when there are many different protected attributes which can be combined to define the "protected groups", the number of potential protected groups is combinatorially large[2].

We model the problem by specifying a class of functions $\mathcal{G}$ defined over a set of d protected attributes. $\mathcal{G}$ defines a set of protected subgroups. Each function $g_i \in \mathcal{G}$ corresponds to a protected subgroup $\mathcal{G}_i = \{x : g_i(x) = 1\}$[3]. The first result of this paper is that for both equal

[1] It also asks for equalized false negative rates, and that the false positive rate is smaller than the true positive rate. Here, the randomness in the "rates" is taken entirely over the randomness of the classifier.

[2] For example, as discussed in a recent Propublica investigation [Angwin and Grassegger, 2017], Facebook policy protects groups against hate speech if the group is definable as a *conjunction* of protected attributes. Under the Facebook schema, "race" and "gender" are both protected attributes, and so the Facebook policy protects "black women" as a distinct class, separately from black people and women. When there are d protected attributes, there are 2^d protected groups. As a statistical estimation problem, this is not a large obstacle — we can estimate 2^d expectations to error ε so long as our data set has size $O(d/\varepsilon^2)$, but there is now a computational problem.

[3] For example, in the case of Facebook's policy, the protected attributes include "race, sex, gender identity, religious affiliation, national origin, ethnicity, sexual orientation and serious disability/disease" [Angwin and Grassegger, 2017], and $\mathcal{G}$ represents the class of boolean conjunctions. In other words, a group defined by individuals having any *subset* of values for the protected attributes is protected.

opportunity and statistical parity, the computational problem of *checking* whether a classifier D violates statistical fairness with respect to the set of protected groups $\mathcal{G}$ is equivalent to the problem of *agnostically learning* $\mathcal{G}$, in a strong sense [Kearns et al., 1994]. This equivalence has two implications:

1. First, it allows us to import *computational hardness* results from the learning theory literature. Agnostic learning turns out to be computationally hard in the worst case, even for extremely simple classes of functions $\mathcal{G}$ (like boolean conjunctions and linear threshold functions). As a result, we can conclude that auditing a classifier D for statistical fairness violations with respect to a class $\mathcal{G}$ is also computationally hard. This means we should not expect to find a polynomial time algorithm that is always guaranteed to solve the auditing problem.

2. However, in practice, various learning heuristics (like logistic regression, SVMs, backpropagation for neural networks, etc.) are commonly used to learn accurate classifiers which are known to be hard to learn in the worst case. The equivalence we show between agnostic learning and auditing is *distribution specific* — that is, if on a particular data set, a heuristic learning algorithm can solve the agnostic learning problem (on an appropriately defined subset of the data), it can be used also to solve the auditing problem on the same data set.

These results appear in Section 3.

Next, we consider the problem of *learning* a classifier that equalizes false positive rates across all (possibly infinitely many) sub-groups, defined by a class of functions $\mathcal{G}$. As per the reductions described above, this problem is computationally hard in the worst case. However, under the assumption that we have efficient oracles which solve the *agnostic learning* problem and the *auditing* problem, we give and analyze an algorithm for this problem based on the *fictitious play* dynamic from game theory. We prove that the optimal fair classifier can be found as the equilibrium of a zero-sum game, in which the action space of the "Learner" player corresponds to classifiers, and the action space of the "Auditor" player corresponds to subgroups implicitly defined by $\mathcal{G}$. The best response problem for the two players corresponds to agnostic learning and auditing, respectively. This means that given a heuristic algorithm capable of solving the agnostic learning problem and the auditing problem, both players can implement the fictitious play dynamic, which requires repeatedly best responding to their opponent's empirical history of play. This dynamic is known to converge to Nash equilibrium in zero-sum games — and hence, in our setting, to the optimal fair classifier. The derivation of the algorithm (and its guarantees) appears in Section 4.

Finally, we implement our algorithm and demonstrate its practicality by efficiently learning classifiers that approximately equalize false positive rates across any group definable by a linear threshold function on 18 protected attributes in the "Communities and Crime" dataset. We use simple, fast regression algorithms as heuristics to implement agnostic learning oracles, and (via our reduction from agnostic learning to auditing) auditing oracles. Our results suggest that it is possible in practice to learn fair classifiers with respect to a large class of subgroups that still achieve non-trivial error. We also implement the algorithm of Agarwal et al. [2017] to learn a classifier that approximately equalizes false positive rates on the same dataset on the 36 groups defined just by the 18 individual protected attributes. We then audit this learned classifier with respect to all linear threshold functions on the 18 protected attributes, and find a subgroup on which the fairness constraint is substantially violated, despite fairness being achieved on all marginal attributes. This shows that phenomenon like Example 1.1 can arise in real learning problems. Full details are contained in Section 5.

1.2 Further Related Work

Thematically, the most closely related piece of work is Zhang and Neill [2016], who also aim to audit classification algorithms for discrimination in subgroups that have not been pre-defined. Our work differs from theirs in a number of important ways. First, we audit the algorithm for common measures of statistical unfairness, whereas Zhang and Neill [2016] design a new measure compatible with their algorithmic technique. Second, we give a formal analysis of our algorithm. Finally, we audit with respect to subgroups defined by a class of functions $\mathcal{G}$, which we can take to have bounded VC dimension, which allows us to give formal out-of-sample guarantees. Zhang and Neill [2016] attempt to audit with respect to *all possible* subgroups, which introduces a severe multiple-hypothesis testing problem, and risks overfitting. Most importantly we give actionable algorithms for learning subgroup fair classifiers, whereas Zhang and Neill [2016] restrict attention to auditing.

Technically, the most closely related piece of work (from which we take inspiration for our algorithm in Section 4) is Agarwal et al. [2017], who show that given access to an agnostic learning oracle for a class $\mathcal{H}$, there is an efficient algorithm to find the lowest-error distribution over classifiers in $\mathcal{H}$ subject to equalizing false positive rates across polynomially many subgroups. Their algorithm can be viewed as solving the same zero-sum game that we solve, but in which the "subgroup" player plays gradient descent over his pure strategies, one for each sub-group. This ceases to be an efficient or practical algorithm when the number of subgroups is large, as is our case. Our main insight is that an auditing algorithm (which we show can also be implemented using an agnostic learning oracle) is sufficient to have the dual player play "fictitious play", which is a dynamic also known to converge to Nash equilibrium.

There is also other work showing computational hardness for fair learning problems. Most notably, Woodworth et al. [2017] show that finding a linear threshold classifier that approximately minimizes hinge loss subject to equalizing false positive rates across populations is computationally hard (assuming that refuting a random k-XOR formula is hard). In contrast, we show that even *checking* whether a classifier satisfies a false positive rate constraint on a particular data set is computationally hard (if the number of subgroups on which fairness is desired is too large to enumerate).

2 Model

We model each individual as being described 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. Note that in our formulation, an auditing algorithm not only may not see the unprotected attributes x', it may not even be aware of their existence. For example, x' may represent proprietary features or consumer data purchased by a credit scoring company.

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 decision making algorithm, and let $D(X)$ denote the (possibly randomized) decision induced by D on individual (X, y). We restrict attention in this paper to the case in which D makes a binary classification decision: $D(X) \in \{0,1\}$. Thus we alternately refer to D as a classifier. When *auditing* a fixed classifier D, it will be helpful to make reference to the distribution over examples (X, y) together with their induced classification $D(X)$. Let $P_{\text{audit}}(D)$ denote the induced *target joint distribution* over the tuple $(x, x', y, D(X))$ that results from sampling $(x, x', y) \sim \mathcal{P}$, and then appending the classification $D(X)$. Note that the randomness here is over both the randomness of $\mathcal{P}$, and the potential randomness of the classifier D.

We will be concerned with learning and auditing classifiers D satisfying two common statistical fairness constraints: equality of classification rates (also known as statistical parity), and equality of false positive rates (also known as equal opportunity). Auditing for equality of false negative rates is symmetric and so we do not explicitly consider it. 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.

Definition 2.1 (Statistical Parity (SP) Subgroup Fairness). *Fix any classifier D, distribution $\mathcal{P}$, collection of group indicators $\mathcal{G}$, and parameters $\alpha, \beta \in [0,1]$. We say that D satisfies (α,β)-*statistical parity (SP) Fairness *with respect to $\mathcal{P}$ and $\mathcal{G}$ if for every $g \in \mathcal{G}$ such that* $\min(\Pr[g(x) = 1], \Pr[g(x) = 0]) \geq \alpha$ *we have:*

$$|\Pr[D(X) = 1 | g(x) = 1] - \Pr[D(X) = 1]| \leq \beta$$

We will sometimes refer to the SP base rate, *which we write as: $b_{SP} = b_{SP}(D,P) = \Pr[D(X) = 1]$.*

Remark 2.2. *Note that our definition has two approximation parameters, both of which are important. We are not required to be "fair" to a group g if it (or its complement) represent only a small fraction of the total probability mass. The parameter α governs how "small" a fraction of the population we are allowed to ignore. Similarly, we do not require that the probability of a positive classification in every subgroup is exactly equal to the base rate, but instead allow deviations of magnitude β. Both of these approximation parameters are important from a statistical estimation perspective. We can never hope from a finite sample of data to precisely estimate any of the probabilities made reference to in the definition of statistical parity fairness — we can only hope to estimate them up to some finite precision β. Moreover, in order to obtain reasonable statistical estimates of the classification probability on a group g, we need sufficiently many samples of members of that group, which will be impossible if α can be taken to be* 0.

Definition 2.3 (False Positive (FP) Subgroup Fairness). *Fix any classifier D, distribution $\mathcal{P}$, collection of group indicators $\mathcal{G}$, and parameters $\alpha, \beta \in [0,1]$. We say D satisfies (α,β)-*False Positive (FP) Fairness *with respect to $\mathcal{P}$ and $\mathcal{G}$ if for every $g \in \mathcal{G}$ such that* $\min(\Pr[g(x) = 1, y = 0], \Pr[g(x) = 0, y = 0]) \geq \alpha$ *we have:*

$$|\Pr[D(X) = 1 | g(x) = 1, y = 0] - \Pr[D(X) = 1 | y = 0]| \leq \beta$$

We will sometimes refer to FP-base rate, which we write as: $b_{FP} = b_{FP}(D,P) = \Pr[D(X) = 1 | y = 0]$.

Remark 2.4. *This definition is symmetric to the definition of statistical parity fairness, except that the α parameter is now used to exclude groups g such that* negative examples *($y = 0$) from those g (or its complement) have probability mass less than α. This is again necessary from a statistical perspective: we cannot accurately estimate false positive rates on a group g without having observed sufficiently many negative examples from that group.*

For both statistical parity and false positive fairness, if the algorithm D fails to satisfy the (α,β)-fairness condition, then we say that D is (α,β)-*unfair* for either statistical parity or false positive rates, with respect to $\mathcal{P}$ and $\mathcal{G}$. We call any subgroup g which witnesses this unfairness an (α,β)-*unfair certificate* for $(D,\mathcal{P})$.

An *auditing algorithm* for a notion of fairness is given sample access to $P_{\text{audit}}(D)$ for some classifier D. It will either deem D to be fair with respect to $\mathcal{P}$, or will else produce a certificate of unfairness.

Definition 2.5 (Auditing). *Fix a notion of fairness (either statistical parity or false-positive fairness), a collection of group indicators $\mathcal{G}$ over the protected features, and any $\alpha, \beta, \alpha', \beta' \in (0,1]$ such that $\alpha' \leq \alpha$ and $\beta' \leq \beta$. A collection of classifiers $\mathcal{H}$ is $(\alpha,\beta,\alpha',\beta')$-*(efficiently) auditable *under distribution $\mathcal{P}$ for groups $\mathcal{G}$ if there exists an auditing algorithm A such that for every classifier $D \in \mathcal{H}$, when given access the distribution $P_{\text{audit}}(D)$, A runs in time* $\text{poly}(1/\alpha, 1/\alpha', 1/\beta, 1/\beta', 1/\delta)$*, and with probability* $(1-\delta)$*, outputs an (α',β')-unfair certificate for D whenever D is (α,β)-unfair with respect to $\mathcal{P}$ and $\mathcal{G}$.*

As we will show, our definition of auditing is closely related to weak agnostic learning.

Definition 2.6 (Weak Agnostic Learning Kalai et al. [2008])**.** *Let Q be a distribution over $\mathcal{X}\times\{0,1\}$ and let $\varepsilon,\gamma\in(0,1/2)$ such that $\varepsilon\geq\gamma$. We say that the function class $\mathcal{G}$ is (ε,γ)-*weakly agnostically learnable *under distribution Q if there exists an algorithm L such that when given sample access to Q, L runs in time $\mathrm{poly}(1/\gamma,1/\delta)$, and with probability $1-\delta$, outputs a hypothesis $h\in\mathcal{G}$ such that*

$$\min_{f\in\mathcal{G}} err(f,Q)\leq 1/2-\varepsilon \implies err(h,Q)\leq 1/2-\gamma.$$

where $err(h,Q)=\Pr_{(x,y)\sim Q}[h(x)\neq y]$.

2.1 Generalization Error

In this section, we observe that the error rate of a classifier D, as well as the degree to which it violates (α,β)-fairness (for both statistical parity and false positive rates) can be accurately approximated with the empirical estimates for these quantities on a dataset (drawn i.i.d. from the underlying distribution $\mathcal{P}$) so long as the dataset is sufficiently large. Once we establish this fact, since our main interest is in the computational problem of auditing and learning, in the rest of the paper, we assume that we have direct access to the underlying distribution and do not make further reference to sample complexity issues.

A standard VC dimension bound (see, e.g. Kearns and Vazirani [1994]) states:

Theorem 2.7. *Fix a class of functions $\mathcal{H}$. For any distribution $\mathcal{P}$, let $S\sim\mathcal{P}^m$ be a dataset consisting of m examples (X_i,y_i) sampled i.i.d. from $\mathcal{P}$. Then for any $0<\delta<1$, with probability $1-\delta$, for every $h\in\mathcal{H}$, we have:*

$$|err(h,\mathcal{P})-err(h,S)|\leq O\left(\sqrt{\frac{\mathrm{VCDIM}(\mathcal{H})\log m+\log(1/\delta)}{m}}\right)$$

where $err(h,S)=\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}[h(X_i)\neq y_i]$.

The above theorem implies that so long as $m\geq\tilde{O}(\mathrm{VCDIM}(\mathcal{H})/\varepsilon^2)$, then minimizing error over the empirical sample S suffices to minimize error up to an additive ε term on the true distribution $\mathcal{P}$. Below, we give two analogous statements for fairness constraints:

Theorem 2.8 (SP Uniform Convergence)**.** *Fix a class of functions $\mathcal{H}$ and a class of group indicators $\mathcal{G}$. For any distribution $\mathcal{P}$, let $S\sim\mathcal{P}^m$ be a dataset consisting of m examples (X_i,y_i) sampled i.i.d. from $\mathcal{P}$. Then for any $0<\delta<1$, with probability $1-\delta$, for every $h\in\mathcal{H}$ and $g\in\mathcal{G}$ with $\Pr_{(X,y)\sim\mathcal{P}}[g(x)=1]\geq\alpha$, we have:*

$$\left|\Pr_{(X,y)\sim\mathcal{P}}[h(X)=1|g(x)=1]-\Pr_{(X,y)\sim S}[h(X)=1|g(x)=1]\right|\leq\tilde{O}\left(\frac{1}{\alpha}\sqrt{\frac{(\mathrm{VCDIM}(\mathcal{H})+\mathrm{VCDIM}(\mathcal{G}))\log m+\log(1/\delta)}{m}}\right)$$

Similarly:

Theorem 2.9 (FP Uniform Convergence)**.** *Fix a class of functions $\mathcal{H}$ and a class of group indicators $\mathcal{G}$. For any distribution $\mathcal{P}$, let $S\sim\mathcal{P}^m$ be a dataset consisting of m examples (X_i,y_i) sampled i.i.d. from $\mathcal{P}$. Then for any $0<\delta<1$, with probability $1-\delta$, for every $h\in\mathcal{H}$ and $g\in\mathcal{G}$ with $\Pr_{(X,y)\sim\mathcal{P}}[g(x)=1,y=0]\geq\alpha$, we have:*

$$\left|\Pr_{(X,y)\sim\mathcal{P}}[h(X)=1|g(x)=1,y=0]-\Pr_{(X,y)\sim S}[h(X)=1|g(x)=1,y=0]\right|\leq$$

$$\tilde{O}\left(\frac{1}{\alpha}\sqrt{\frac{(\mathrm{VCDIM}(\mathcal{H})+\mathrm{VCDIM}(\mathcal{G}))\log m+\log(1/\delta)}{m}}\right)$$

These theorems together imply that for both statistical parity and false positive rate fairness, the degree to which a group g violates the constraint of (α, β) fairness can be estimated up to error ε, so long as $m \geq \tilde{O}((\text{VCDIM}(\mathcal{H}) + \text{VCDIM}(\mathcal{G}))/(\alpha\varepsilon)^2)$. The proofs can be found in Appendix A.

3 Equivalence of Auditing and Weak Agnostic Learning

In this section, we give a reduction from the problem of auditing both statistical parity and false positive rate fairness, to the problem of agnostic learning, and vice versa. This has two implications. The main implication is that, from a worst-case analysis point of view, auditing is computationally hard in almost every case (since it inherits this pessimistic state of affairs from agnostic learning). However, worst-case hardness results in learning theory have not prevented the successful practice of machine learning, and there are many heuristic algorithms that in real-world cases successfully solve "hard" agnostic learning problems. Our reductions also imply that these heuristics can be used successfully as auditing algorithms.

We make the following mild assumption on the class of group indicators $\mathcal{G}$, to aid in our reductions. It is satisfied by most natural classes of functions, but is in any case essentially without loss of generality (since learning negated functions can be simulated by learning the original function class on a dataset with flipped class labels).

Assumption 3.1. *We assume the set of group indicators $\mathcal{G}$ satisfies closure under negation: for any $g \in \mathcal{G}$, we also have $\neg g \in \mathcal{G}$.*

The following two distributions will be useful for describing our results:

- P^D: the marginal distribution on $(x, D(X))$.
- $P^D_{y=0}$: the conditional distribution on $(x, D(X))$, conditioned on $y = 0$.

We will think about these as the target distributions for a learning problem: i.e. the problem of learning to predict $D(X)$ from only the protected features x. We will relate the ability to agnostically learn on these distributions, to the ability to audit D given access to the original distribution $P_{\text{audit}}(D)$.

3.1 Statistical Parity Fairness

We give our reduction first for statistical parity fairness. The reduction for false positive rate fairness will follow as a corollary, since auditing for false positive rate fairness can be viewed as auditing for statistical parity fairness on the subset of the data restricted to $y = 0$.

Theorem 3.2. *Fix any distribution P over individual data points, any set of group indicators $\mathcal{G}$, and any classifier D. Suppose that the (statistical parity) base rate $b_{SP}(D, P) = 1/2$. The following two relationships hold:*

- *If D is $(\gamma, \gamma, \alpha', \beta')$-auditable for statistical parity fairness under distribution P and group indicator set $\mathcal{G}$, then the class $\mathcal{G}$ is $(\gamma, \alpha'\beta')$-weakly agnostically learnable under P^D.*
- *If $\mathcal{G}$ is $(\alpha\beta, \gamma)$-weakly agnostically learnable under distribution P^D, then D is $(\alpha, \beta, \gamma, \gamma)$-auditable for statistical parity fairness under P and group indicator set $\mathcal{G}$.*

We will prove Theorem 3.2 in two steps. First, we show that any unfair certificate f for D has non-trivial error for predicting the decision made by D from the sensitive attributes.

Lemma 3.3. *Suppose that the base rate $b_{SP}(D, P) \leq 1/2$ and there exists a function f such that $\Pr[f(x) = 1] = \alpha$ and $|\Pr[D(X) = 1 \mid f(x) = 1] - b_{SP}(D, P)| > \beta$. Then*

$$\max\{\Pr[D(X) = f(x)], \Pr[D(X) = \neg f(x)]\} \geq b_{SP} + \alpha\beta$$

Proof. Let $b = b_{SP}(D,p)$ denote the base rate. Given that $|\Pr[D(X)=1 \mid f(x)=1]-b| > \beta$, we know either $\Pr[D(X)=1 \mid f(x)=1] > b+\beta$ or $\Pr[D(X)=1 \mid f(x)=1] < b-\beta$.

In the first case, we know $\Pr[D(X)=1 \mid f(x)=0] < b$, and so $\Pr[D(X)=0 \mid f(x)=0] > 1-b$. It follows that

$$\begin{aligned}\Pr[D(X)=f(x)] &= \Pr[D(X)=f(x)=1]+\Pr[D(X)=f(x)=0]\\ &= \Pr[D(X)=1 \mid f(x)=1]\Pr[f(x)=1]+\Pr[D(X)=0 \mid f(x)=0]\Pr[f(x)=0]\\ &> \alpha(b+\beta)+(1-\alpha)(1-b) = (\alpha-1)b+(1-\alpha)(1-b)+b+\alpha\beta = (1-\alpha)(1-2b)+b+\alpha\beta\end{aligned}$$

In the second case, we have $\Pr[D(X)=0 \mid f(x)=1] > (1-b)+\beta$ and $\Pr[D(X)=1 \mid f(x)=0] > b$. We can then bound

$$\begin{aligned}\Pr[D(X)=f(x)] &= \Pr[D(X)=1 \mid f(x)=0]\Pr[f(x)=0]+\Pr[D(X)=0 \mid f(x)=1]\Pr[f(x)=1]\\ &> (1-\alpha)b+\alpha(1-b+\beta) = \alpha(1-2b)+b+\alpha\beta.\end{aligned}$$

In both cases, we have $(1-2b) \geq 0$ by our assumption on the base rate. Since $\alpha \in [0,1]$, we know

$$\max\{\Pr[D(X)=f(x)], \Pr[D(X)=\neg f(x)]\} \geq b+\alpha\beta$$

□

In the next step, we show that if there exists any function f that accurately predicts the decisions made by the algorithm D, then either f or $\neg f$ can serve as an unfair certificate of D.

Lemma 3.4. *Suppose that the base rate $b_{SP}(D,P) \geq 1/2$ and there exists a function f such that $\Pr[D(X)=f(x)] \geq b_1(D,P)+\gamma$ for some value $\gamma \in (0,1/2)$. Then there exists a function g such that $\Pr[g(x)=1] \geq \gamma$ and $|\Pr[D(X)=1 \mid g(x)=1]-b_{SP}(D,p)| > \gamma$, where $g \in \{f, \neg f\}$.*

Proof. Again recall that

$$\begin{aligned}\Pr[D(X)=f(x)] &= \Pr[D(X)=f(x)=1]+\Pr[D(X)=f(x)=0]\\ &= \Pr[D(X)=1 \mid f(x)=1]\Pr[f(x)=1]+\Pr[D(X)=0 \mid f(x)=0]\Pr[f(x)=0]\end{aligned}$$

Since $\Pr[D(X)=f(x)] \geq b+\gamma$ and $\Pr[f(x)=1]+\Pr[f(x)=0]=1$,

$$\max\{\Pr[D(X)=1 \mid f(x)=1], \Pr[D(X)=0 \mid f(x)=0]\} \geq b+\gamma.$$

Thus, we must have either $\Pr[D(X)=1 \mid f(x)=1] \geq b+\gamma$ or $\Pr[D(X)=1 \mid f(x)=0] \leq b-\gamma$. Next, we observe that $\min\{\Pr[f=1], \Pr[f=0]\} \geq \gamma$. This follows since $\Pr[D(x,x')=f] = \Pr[D(x,x')=f=1]+\Pr[D(x,x')=f=0] \leq \Pr[f=1]+\Pr[D=0]$. But $\Pr[f=1] \geq \Pr[D(x,x')=f]-\Pr[D=0] \geq b+\gamma-b=\gamma$. The same result holds for $\Pr[f=0]$. □

Proof of Theorem 3.2. Suppose that the class $\mathcal{G}$ satisfies $\min_{f\in\mathcal{G}} err(f, P^D) \leq 1/2-\gamma$. Then by Lemma 3.4, there exists some $g \in \mathcal{G}$ such that $\Pr[g(x)=1] \geq \gamma$ and $|\Pr[D(X)=1 \mid g(x)=1]-b_1(D,p)| > \gamma$. By the assumption of auditability, we can then use the auditing algorithm to find a group $g' \in \mathcal{G}$ that is an (α',β')-unfair certificate of D. By Lemma 3.3, we know that either g' or $\neg g'$ predicts D with an accuracy of at least $1/2+\alpha'\beta'$.

In the reverse direction: consider the auditing problem on the classifier D. We can treat each pair $(x, D(X))$ as a labelled example and learn a hypothesis in $\mathcal{G}$ that approximates the decisions made by D. Suppose that D is (α,β)-unfair. Then by Lemma 3.3, we know that there exists some $g \in \mathcal{G}$ such that $\Pr[D(X)=g(x)] \geq 1/2+\alpha\beta$. Therefore, the weak agnostic learning algorithm from the hypothesis of the theorem will return some g' with $\Pr[D(X)=g'(x)] \geq 1/2+\gamma$. By Lemma 3.4, we know g' or $\neg g'$ is a (γ,γ)-unfair certificate for D. □

Remark 3.5. *The assumption in Theorem 3.2 that the base rate is exactly $1/2$ can be relaxed. In the appendix, we extend the result by showing that our reduction holds as long as the base rate by b_{SP} lies in the interval $[1/2-\varepsilon, 1/2+\varepsilon]$ for sufficiently small ε.*

3.2 False Positive Fairness

A corollary of the above reduction is an analogous equivalence between auditing for false-positive fairness and agnostic learning. This is because a false-positive fairness constraint can be viewed as a statistical parity fairness constraint on the subset of the data such that $y = 0$. Therefore, Theorem 3.2 implies the following:

Corollary 3.6. *Fix any distribution P over individual data points, any set of group indicators $\mathcal{G}$, and any classifier D. Suppose that the (false-positive) base rate $b_{FP}(D, P) = 1/2$. The following two relationships hold:*

- *If D is $(\gamma, \gamma, \alpha', \beta')$-auditable for false-positive fairness under distribution P and group indicator set $\mathcal{G}$, then the class $\mathcal{G}$ is $(\gamma, \alpha'\beta')$-weakly agnostically learnable under $P^D_{y=0}$.*
- *If $\mathcal{G}$ is $(\alpha\beta, \gamma)$-weakly agnostically learnable under distribution $P^D_{y=0}$, then D is $(\alpha, \beta, \gamma, \gamma)$-auditable for false-positive fairness under distribution P and group indicator set $\mathcal{G}$.*

3.3 Worst-Case Intractability of Auditing

While we shall see in subsequent sections that the equivalence given above has positive algorithmic and experimental consequences, from a purely theoretical perspective the reduction of agnostic learning to auditing has strong negative worst-case implications. More precisely, we can import a long sequence of formal intractability results for agnostic learning to obtain:

Theorem 3.7. *Under standard complexity-theoretic intractability assumptions, for $\mathcal{G}$ the classes of conjunctions of boolean attributes, linear threshold functions, or bounded-degree polynomial threshold functions, there exist distributions P such that the auditing problem cannot be solved in polynomial time, for either statistical parity or false positive fairness.*

The proof of this theorem follows from Theorem 3.2, Corollary 3.6, and the following negative results from the learning theory literature. Feldman et al. [2012] show a strong negative result for weak agnostic learning for conjunctions: given a distribution on labeled examples from the hypercube such that there exists a monomial (or conjunction) consistent with $(1-\varepsilon)$-fraction of the examples, it is NP-hard to find a halfspace that is correct on $(1/2+\varepsilon)$-fraction of the examples, for arbitrary constant $\varepsilon > 0$. Diakonikolas et al. [2011] show that under the Unique Games Conjecture, no polynomial-time algorithm can find a degree-d polynomial threshold function (PTF) that is consistent with $(1/2+\varepsilon)$ fraction of a given set of labeled examples, even if there exists a degree-d PTF that is consistent with a $(1-\varepsilon)$ fraction of the examples. Diakonikolas et al. [2011] also show that it is NP-Hard to find a degree-2 PTF that is consistent with a $(1/2+\varepsilon)$ fraction of a given set of labeled examples, even if there exists a halfspace (degree-1 PTF) that is consistent with a $(1-\varepsilon)$ fraction of the examples.

While Theorem 3.7 shows that certain natural subgroup classes $\mathcal{G}$ yield intractable auditing problems in the worst case, in the rest of the paper we demonstrate that effective heuristics for this problem on specific (non-worst case) distributions can be used to derive an effective and practical learning algorithm for subgroup fairness.

4 A Learning Algorithm Subject to Fairness Constraints $\mathcal{G}$

In this section, we give a general algorithm for training a (randomized) classifier that satisfies approximate false-positive fairness simultaneously for all protected subgroups specified by a family of group indicator functions $\mathcal{G}$. All of our techniques also apply to a statistical parity or false negative rate constraint.

Let $\mathcal{H}$ be a hypothesis class over both the protected and unprotected attributes, and let $\mathcal{G}$ be a collection of group indicators over the protected attributes. We assume that $\mathcal{H}$ contains a constant classifier (which implies that there is at least one fair classifier to be found, for any distribution). For our convergence theorem, we will also assume that $\mathcal{H}$ and $\mathcal{G}$ are finite — but this is not necessary for our algorithm to be well defined (in fact, in Section 5, we run our algorithm with an infinite hypothesis class $\mathcal{H}$ and observe that it still converges). Our goal will be to find the distribution over classifiers from $\mathcal{H}$ that minimizes classification error subject to the fairness constraint over $\mathcal{G}$. Note that because minimizing classification error is NP-hard even for simple classes $\mathcal{H}$ (even absent a fairness constraint), our goal will not be to design an algorithm that runs in polynomial time in the worst case. Instead, our ultimate goal is to design a practical algorithm. We design and analyze our algorithm imagining that we have oracles capable of solving (unconstrained) learning and auditing problems. These will ultimately be implemented in practice with heuristic learning algorithms. We will design an iterative algorithm that provably converges to the optimal randomized classifier, while only needing to explicitly represent a concise state (linear in the number of timesteps the algorithm has run, but independent of the number of hypotheses $|\mathcal{H}|$ or the number of protected groups $|\mathcal{G}|$).

More formally, let $p = (p_h)_{h\in\mathcal{H}}$ denote a probability distribution over $\mathcal{H}$. Fix a distribution over examples P. It will be useful to have notation for the false positive rate of a fixed classifier h on a group g:

$$\mathrm{FP}(h,g) = \mathop{\mathbb{E}}_{P}\left[h(x,x') \mid y = 0,\ g(x) = 1\right]$$

and for the overall false-positive rate of h (over all groups g):

$$\mathrm{FP}(h) = \mathop{\mathbb{E}}_{P}\left[h(x,x') \mid y = 0\right].$$

For any $\alpha \in (0,1)$, let $\mathcal{G}_\alpha = \{g \in \mathcal{G} \mid \Pr_P[g(x) = 1, y = 0] \geq \alpha\}$. We consider the following "Fair ERM" problem:

$$\begin{aligned} &\min_{p\in\Delta_{\mathcal{H}}} \mathop{\mathbb{E}}_{h\sim p}\left[err(h,P)\right] \\ \text{such that} \quad &\mathop{\mathbb{E}}_{h\sim p}\left[\mathrm{FP}(h,g)\right] - \mathop{\mathbb{E}}_{h\sim p}\left[\mathrm{FP}(h)\right] = 0 \qquad \forall g \in \mathcal{G}_\alpha \end{aligned}$$

where $err(h,P) = \Pr_P[h(x,x') \neq y]$. Observe that this is a linear program (with one variable for every classifier $h \in \mathcal{H}$), and that the LP is feasible: the constant classifiers that labels all points 1 or 0 satisfy all subgroup fairness constraints.

Overview of our solution. We design an algorithm to solve the above "Fair ERM" linear program via the following steps:

1. First, we derive the partial Lagrangian of the Fair ERM LP, and note that computing an approximately optimal solution to the LP is equivalent to finding an approximate minmax solution for a zero-sum game, in which the objective function is the value of the Lagrangian. The actions of the primal or "Learner" player correspond to classifiers $h \in \mathcal{H}$, and the actions of the dual or "Auditor" player correspond to groups $g \in \mathcal{G}$.
2. In order to reason about convergence, we restrict the set of dual variables to lie in a bounded set: a scaling of the unit ℓ_1 ball. In particular, it is useful in the analysis to view the set of vertices in the ℓ_1 ball as a finite set of pure strategies for the Auditor player. We show that as a function of the restriction we choose, the approximate minmax solution of the constrained game continues to give an approximately optimal solution to the fair ERM problem.
3. We observe that given a mixed strategy for the Auditor, the best response problem of the Learner corresponds to an agnostic learning problem. Similarly, given a mixed strategy for the Learner, the best response problem of the Auditor corresponds to an auditing

problem. Hence, if we have oracles for solving agnostic learning and auditing problems, we can compute best responses for both players, in response to arbitrary mixed strategies of their opponents.

4. Finally, we note that the ability to compute best responses for each player is sufficient to implement the “fictitious play” dynamics, in which both players repeatedly best respond to the mixed strategy corresponding to the empirical play history of their opponent. In zero-sum games, fictitious play dynamics are known to converge to Nash equilibrium, which achieve the minmax value of the game. In our case, this corresponds to the Learner converging to the optimal solution to the fair ERM problem. Besides implementing agnostic learning and auditing oracles heuristically, simulating this dynamic requires only remembering the set of classifiers h_t played by the Learner, and each of the groups g_t played by the Auditor at each of the T rounds so far.

4.1 Lagrangian with Restricted Dual Space

For each constraint, corresponding to a group $g \in \mathcal{G}_\alpha$, we introduce a dual variable λ_g. The partial Lagrangian of the linear program is the following:

$$\mathcal{L}(p, \lambda) = \mathop{\mathbb{E}}_{h \sim p}[err(h, P)] + \sum_{g \in \mathcal{G}} \lambda_g \left(\mathop{\mathbb{E}}_{h \sim p}[\mathrm{FP}(h, g)] - \mathop{\mathbb{E}}_{h \sim p}[\mathrm{FP}(h)] \right)$$

By Sion’s minmax theorem [Sion, 1958], we have

$$\min_{p \in \Delta_\mathcal{H}} \max_{\lambda \in \mathbb{R}^{|\mathcal{G}|}} \mathcal{L}(p, \lambda) = \max_{\lambda \in \mathbb{R}^{|\mathcal{G}|}} \min_{p \in \Delta_\mathcal{H}} \mathcal{L}(p, \lambda) = \mathrm{OPT}$$

where OPT denotes the optimal objective value in the fair ERM problem. Similarly, the quantity $\arg\min_{p \in \Delta_\mathcal{H}} \max_{\lambda \in \mathbb{R}^{|\mathcal{G}|}} \mathcal{L}(p, \lambda)$ corresponds to an optimal feasible solution to the fair ERM linear program. Thus, finding an optimal solution for the fair ERM problem reduces to computing a minmax solution for the Lagrangian. Our algorithm will compute such a minmax solution by iteratively optimizing over both the primal variables p and dual variables λ. In order to guarantee convergence in our optimization, we will restrict the dual space to the following bounded set:

$$\Lambda = \{\lambda \in \mathbb{R}^{|\mathcal{G}|} \mid \|\lambda\|_1 \le C\}.$$

where C will be a parameter of our algorithm. Since Λ is a compact and convex set, the minmax condition continues to hold [Sion, 1958]:

$$\min_{p \in \Delta_\mathcal{H}} \max_{\lambda \in \Lambda} \mathcal{L}(p, \lambda) = \max_{\lambda \in \Lambda} \min_{p \in \Delta_\mathcal{H}} \mathcal{L}(p, \lambda) \tag{1}$$

If we knew an upper bound C on the ℓ_1 norm of the optimal dual solution, then this restriction on the dual solution would not change the minmax solution of the program. We do not in general know such a bound. However, we now show that even though we restrict the dual variables to lie in a bounded set, any approximate minmax solution to Equation (1) is also an approximately optimal and approximately feasible solution to the original fair ERM problem.

Theorem 4.1. *Let $(\hat{p}, \hat{\lambda})$ be a ν-approximate minmax solution to the Lagrangian problem in the sense that*

$$\mathcal{L}(\hat{p}, \hat{\lambda}) \le \min_{p \in \Delta_\mathcal{H}} \mathcal{L}(p, \hat{\lambda}) + \nu \qquad \text{and,} \qquad \mathcal{L}(\hat{p}, \hat{\lambda}) \ge \max_{\lambda \in \Lambda} \mathcal{L}(\hat{p}, \lambda) - \nu.$$

Then $err(\hat{p}, P) \le \mathrm{OPT} + 2\nu$ and for any $g \in \mathcal{G}_\alpha$,

$$\left| \mathop{\mathbb{E}}_{h \sim \hat{p}}[\mathrm{FP}(h, g)] - \mathop{\mathbb{E}}_{h \sim \hat{p}}[\mathrm{FP}(h)] \right| \le \frac{1 + 2\nu}{C}.$$

Proof. Let p^* be the optimal feasible solution for our constrained optimization problem. Let $\hat{g} \in \arg\max_{g\in\mathcal{G}_\alpha} |\mathbb{E}_{h\sim\hat{p}}[\text{FP}(h,g)] - \mathbb{E}_{h\sim\hat{p}}[\text{FP}(h)]|$, and let $\lambda' \in \mathbb{R}^{|\mathcal{G}_\alpha|}$ be the best response to $\hat{p}$ with

$$\lambda'_{\hat{g}} = C\,\text{sgn}\left(\mathop{\mathbb{E}}_{h\sim\hat{p}}[\text{FP}(h,\hat{g})] - \mathop{\mathbb{E}}_{h\sim\hat{p}}[\text{FP}(h)]\right)$$

and zeros in all other coordinates.

By the definition of a ν-approximate minmax solution, we know that $\mathcal{L}(\hat{p},\hat{\lambda}) \geq \mathcal{L}(\hat{p},\lambda') - \nu$. This implies that

$$\mathcal{L}(\hat{p},\hat{\lambda}) \geq err(\hat{p},P) + C\left|\mathop{\mathbb{E}}_{h\sim\hat{p}}[\text{FP}(h,\hat{g})] - \mathop{\mathbb{E}}_{h\sim\hat{p}}[\text{FP}(h)]\right| - \nu \tag{2}$$

Since the optimal solution p^* is feasible, we know that $\mathcal{L}(p^*,\hat{\lambda}) = err(p^*,P)$. Note that

$$\mathcal{L}(\hat{p},\hat{\lambda}) \leq \min_{p\in\Delta_\mathcal{H}} \mathcal{L}(p,\hat{\lambda}) + \nu \leq \mathcal{L}(p^*,\hat{\lambda}) + \nu \tag{3}$$

Combining Equations (2) and (3), we get

$$err(\hat{p},P) + C\left|\mathop{\mathbb{E}}_{h\sim\hat{p}}[\text{FP}(h,\hat{g})] - \mathop{\mathbb{E}}_{h\sim\hat{p}}[\text{FP}(h)]\right| \leq \mathcal{L}(\hat{p},\hat{\lambda}) + \nu \leq \mathcal{L}(p^*,\hat{\lambda}) + 2\nu = err(p^*,P) + 2\nu$$

Note that $C\left|\mathbb{E}_{h\sim\hat{p}}[\text{FP}(h,\hat{g})] - \mathbb{E}_{h\sim\hat{p}}[\text{FP}(h)]\right| \geq 0$, so we must have $err(\hat{p},P) \leq err(p^*,P) + 2\nu = \text{OPT} + 2\nu$. Furthermore, since $err(\hat{p},P), err(p^*,P) \in [0,1]$, we know

$$C\left|\mathop{\mathbb{E}}_{h\sim\hat{p}}[\text{FP}(h,\hat{g})] - \mathop{\mathbb{E}}_{h\sim\hat{p}}[\text{FP}(h)]\right| \leq 1 + 2\nu,$$

which implies that constraint violation satisfies $\left|\mathbb{E}_{h\sim\hat{p}}[\text{FP}(h,\hat{g})] - \mathbb{E}_{h\sim\hat{p}}[\text{FP}(h)]\right| \leq (1+2\nu)/C$. □

4.2 Computing Approximate Minmax Solutions

To compute an approximate minmax solution, we will first view Equation (1) as the following two player zero-sum matrix game. The Learner (or the minimization player) has pure strategies corresponding to $\mathcal{H}$, and the Auditor (or the maximization player) has pure strategies corresponding to the set of vertices Λ_{pure} in Λ — more precisely, each vertex or pure strategy consists of a choice of a $g \in \mathcal{G}_\alpha$, along with the coefficient $+C$ or $-C$ that the corresponding g-fairness constraint will have in the Lagrangian. In the case of finite $\mathcal{G}$, we can write

$$\Lambda_{\text{pure}} = \{\pm C e_g \mid g \in \mathcal{G}_\alpha\}$$

where e_g is the standard basis vector in coordinate G in $|G|$-dimensional space. (This is strictly for notational convenience — our algorithm will never need to actually represent such vectors explicitly.) Note that any vector in Λ can be written as a convex combination of maximization player's pure strategies, or in other words: as a mixed strategy for the Auditor. For any pair of actions $(h,\lambda) \in \mathcal{H} \times \Lambda_{\text{pure}}$, the payoff is defined as

$$U(h,\lambda) = err(h,P) + \sum_{g\in\mathcal{G}_\alpha} \lambda_g (\text{FP}(h,g) - \text{FP}(h)).$$

Claim 4.2. *Let $p \in \Delta_\mathcal{H}$ and $\lambda \in \Lambda$ such that (p,λ) is a ν-approximate minmax strategy in the zero-sum game defined above. Then (p,λ) is also a ν-approximate minmax solution for Equation (1).*

Our problem reduces to finding an approximate equilibrium for this game. To compute an approximate equilibrium for the zero-sum game, we use a simple iterative procedure known as *fictitious play* [Brown, 1949]. It proceeds in rounds, and in every round each player chooses a best response to his opponent's empirical history of play across previous rounds, by treating it as the mixed strategy that randomizes uniformly over the empirical history. The seminal result of Robinson [1951] shows that the empirical mixed strategy of the two players converges to Nash equilibrium in any finite, bounded zero sum game. To formally describe fictitious play in our setting, we will first characterize the best response subroutine for the two players.

Learner's best response. Fix any mixed strategy (dual solution) $\overline{\lambda} \in \Lambda$ of the Auditor. The Learner's best response is given by:

$$\min_{h\in\mathcal{H}} err(h,P) + \sum_{g\in\mathcal{G}_\alpha} \overline{\lambda}_g \left(\mathrm{FP}(h,g) - \mathrm{FP}(h)\right) \tag{4}$$

Note that it suffices for the Learner to optimize over deterministic classifiers $h \in \mathcal{H}$, rather than distributions over classifiers. This is because the Learner is solving a linear optimization problem over the simplex, and so always has an optimal solution at a vertex (i.e. a single classifier $h \in \mathcal{H}$).

We can reduce this problem to one that can be solved with a single call to a *cost-sensitive classification oracle*. A *cost-sensitive classification oracle* for the class $\mathcal{H}$ does the following: given as input a set of n points $\{(X_i, c_i^0, c_i^1)\}$ such that c_i^ℓ corresponds to the cost for predicting label ℓ on point X_i, the algorithm outputs an hypothesis $\hat{h}$ such that

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

We can find a best response for the Learner by making a call to the cost sensitive classification oracle, in which we assign costs to each example (X_i, y_i) as follows:

- if $y_i = 1$, then $c_i^0 = \frac{1}{n}$ and $c_i^1 = 0$;
- otherwise, $c_i^0 = 0$ and $c_i^1 = \frac{1}{n} + \frac{1}{n}\sum_{g\in\mathcal{G}_\alpha} \overline{\lambda}_g \left(\frac{\mathbf{1}[g(x_i)=1]}{\Pr[g(x)=1 \& y=0]} - \frac{1}{\Pr[y=0]}\right)$

Remark 4.3. *In our reduction, the cost of correctly classifying a point is always 0, and the only thing that varies is the cost of misclassifying a point. Note that the solution to a cost-sensitive classification problem is invariant to translation of the costs, and so the costs can always be translated so that they are non-negative. In this case, a cost-sensitive classification problem is simply a weighted agnostic learning problem, which is equivalent to a standard agnostic learning problem in which some points have been duplicated in proportion to their weights. Hence, our use of a cost-sensitive classification oracle is computationally equivalent to an agnostic learning oracle. We describe it directly as a cost-sensitive classification oracle, as this is easier to implement, as we do in Section 5.*

Claim 4.4. *Given the costs specified above, the cost-sensitive classification oracle returns an hypothesis that solves the Learner's best response problem in Equation* (4).

Auditor's best response. Next, we show the best response for the Auditor is to play a pure strategy corresponding to the one subgroup in which the learner's randomized classifier induces the largest disparity in the false-positive rates. In our algorithm, we assume an *auditing oracle* that can identify such a subgroup.

Lemma 4.5. *Fix any $\overline{p} \in \Delta_{\mathcal{H}}$. A best response of the maximization player against $\overline{p}$ is $\lambda' \in \Lambda_{pure}$ with $\lambda'_{g(\overline{p})} = C\, sgn\left[\mathbb{E}_{h\sim\overline{p}}[\mathrm{FP}(h, g(\overline{p}))] - \mathbb{E}_{h\sim\overline{p}}[\mathrm{FP}(h)]\right]$, where*

$$g(\overline{p}) = \arg\max_{g\in\mathcal{G}_\alpha} \left| \mathop{\mathbb{E}}_{h\sim\overline{p}}[\mathrm{FP}(h,g)] - \mathop{\mathbb{E}}_{h\sim\overline{p}}[\mathrm{FP}(h)] \right|$$

Proof. Note that

$$\underset{\lambda\in\Lambda_{\text{pure}}}{\operatorname{argmax}} \mathcal{L}(\overline{p},\lambda) = \underset{\lambda\in\Lambda_{\text{pure}}}{\operatorname{argmax}} \mathop{\mathbb{E}}_{h\sim\overline{p}}[err(h,P)] + \sum_{g\in\mathcal{G}} \lambda_g \left(\mathop{\mathbb{E}}_{h\sim\overline{p}}[\mathrm{FP}(h,g)] - \mathop{\mathbb{E}}_{h\sim\overline{p}}[\mathrm{FP}(h)] \right)$$

$$= \underset{\lambda\in\Lambda_{\text{pure}}}{\operatorname{argmax}} \sum_{g\in\mathcal{G}} \lambda_g \left(\mathop{\mathbb{E}}_{h\sim\overline{p}}[\mathrm{FP}(h,g)] - \mathop{\mathbb{E}}_{h\sim\overline{p}}[\mathrm{FP}(h)] \right)$$

Note that this is a linear optimization problem over a scaling of the ℓ_1 ball, and so has a solution at a vertex, which corresponds to a single group $g \in \mathcal{G}_\alpha$. Thus, there is always a best response λ' that puts all its weight on the group that maximizes $|\mathbb{E}_{h\sim\overline{p}}[\mathrm{FP}(h,g)] - \mathbb{E}_{h\sim\overline{p}}[\mathrm{FP}(h)]|$. □

We can now describe the full algorithm, and state a theorem regarding its convergence.

Algorithm 1 FairFictPlay: Fair Fictitious Play

Input: distribution P over the labelled data points, auditing oracle $\mathcal{A}$ and cost-sensitive classification oracle $\mathcal{C}$, dual bound C, and number of rounds T

Initialize: set h^0 to be some classifier in $\mathcal{H}$, set λ^0 to be the zero vector. Let $\overline{p}$ and $\overline{\lambda}$ be the point distributions that put all their mass on h^0 and λ^0 respectively.

For $t = 1, \ldots, T$:

Learner best responds: Call the oracle $\mathcal{C}$ to find a classifier

$$h^t \in \underset{h\in\mathcal{H}}{\operatorname{argmin}}\; err(h,P) + \sum_{g\in\mathcal{G}_\alpha} \overline{\lambda}_g \left(\mathrm{FP}(h,g) - \mathrm{FP}(h)\right)$$

Auditor best responds: Call the oracle $\mathcal{A}$ to find a subgroup

$$g(\overline{p}) \in \underset{g\in\mathcal{G}_\alpha}{\operatorname{argmax}} \left| \mathop{\mathbb{E}}_{h\sim\overline{p}}[\mathrm{FP}(h,g)] - \mathop{\mathbb{E}}_{h\sim\overline{p}}[\mathrm{FP}(h)] \right|$$

Let $\lambda^t \in \Lambda_{\text{pure}}$ with $\lambda^t_{g(\overline{p})} = C \operatorname{sgn}\left[\mathbb{E}_{h\sim\overline{p}}[\mathrm{FP}(h,g(\overline{p}))] - \mathbb{E}_{h\sim\overline{p}}[\mathrm{FP}(h)]\right]$

Update the empirical play distributions:

Let $\overline{p}$ be the uniform distribution over the set of classifiers $\{h^0, \ldots, h^{t-1}\}$

Let $\overline{\lambda} = \frac{\sum_{t'<t} \lambda^{t'}}{t}$ be the auditor's empirical dual vector

Output: the final empirical distribution $\overline{p}$ over classifiers

Theorem 4.6. *Fix any constant C. Suppose we run Fair Fictitious Play with dual bound C for T rounds. Then the randomized classifier $\overline{p}$ it outputs satisfies $err(\overline{p},P) \le \mathrm{OPT} + 2\nu$ and for any $g \in \mathcal{G}_\alpha$,*

$$\left| \mathop{\mathbb{E}}_{h\sim\overline{p}}[\mathrm{FP}(h,g)] - \mathop{\mathbb{E}}_{h\sim\hat{p}}[\mathrm{FP}(h)] \right| \le \frac{1+2\nu}{C},$$

where $\nu = O\left(T^{\frac{-1}{|\mathcal{G}|+|\mathcal{H}|-2}}\right)$

Proof. By the result of Robinson [1951], the output $(\overline{p}, \overline{\lambda})$ form a ν-approximate equilibrium of the Lagrangian game, where

$$\nu = O\left(T^{\frac{-1}{|\mathcal{G}|+|\mathcal{H}|-2}}\right).$$

Our result then directly follows from Theorem 4.1. □

Remark 4.7. *The theoretical convergence rate in Theorem 4.6 is quite slow, because the best known convergence rate for fictitious play (proven by Robinson [1951]) has an exponential dependence on the number of actions of each player. However, fictitious play is known to converge much more quickly in practice. In fact*, Karlin's conjecture *is that fictitious play enjoys a polynomial convergence rate*[4] *even in the worst case. We note that if we assumed more about the auditor and learner player — namely that they could play* no regret *algorithms, then we would obtain convergence in a polynomial number of rounds. But our weaker assumptions of an auditing and learning oracle are closer to practice: as we show in Section 5, our algorithm is practical and converges quickly on real data.*

5 Experimental Evaluation

5.1 Description of Data

We conclude our results with an extensive analysis of our algorithm's behavior and performance on a real data set, and a demonstration that our methods are necessary empirically to avoid fairness gerrymandering.

The dataset we use for our experimental valuation is known as the "Communities and Crime" (C&C) dataset, available at the UC Irvine Data Repository[5]. Each record in this dataset describes the aggregate demographic properties of a different U.S. community; the data combines socio-economic data from the 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCR. The total number of records is 1994, and the number of features is 122. The variable to be predicted is the rate of violent crime in the community.

While there are larger and more recent datasets in which subgroup fairness is a potential concern, there are properties of the C&C dataset that make it particularly appealing for the initial experimental evaluation of our proposed algorithm. Foremost among these is the relatively high number of sensitive or protected attributes, and the fact that they are real-valued (since they represent aggregates in a community rather than specific individuals). This means there is a very large number of protected sub-groups that can be defined over them. There are distinct continuous features measuring the percentage or per-capita representation of multiple racial groups (including white, black, Hispanic, and Asian) in the community, each of which can vary independently of the others. Similarly, there are continuous features measuring the average per capita incomes of different racial groups in the community, as well as features measuring the percentage of each community's police force that falls in each of the racial groups. Thus restricting to features capturing race statistics and a couple of related ones (such as the percentage of residents who do not speak English well), we obtain an 18-dimensional space of real-valued protected attributes. We note that the C&C dataset has numerous other features that arguably could or should be protected as well (such as gender features), which would raise the dimensionality of the protected subgroups even further. [6]

We convert the real-valued rate of violent crime in each community to a binary label indicating whether the community is in the 70th percentile of that value, indicating that it is a relatively high-crime community. Thus the strawman baseline that always predicts 0 (lower crime) has error approximately 30% or 0.3 on this classification problem. We chose the 70th percentile since it seems most natural to predict the highest crime rates. This choice also

[4]A strong version of Karlin's conjecture, in which ties can be broken adversarially, has recently been disproven Daskalakis and Pan [2014]. But as far as the we know, fictitious play with fixed tie-breaking rules may still converge to equilibrium at a polynomial rate.

[5]http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime

[6]Experiments on other sets of protected features and other datasets where fairness is a concern will be reported on in later versions.

demonstrates that although our formal reductions required a 0.5 base rate, our proposed algorithm can perform well in practice without this condition.

As in the theoretical sections of the paper, our main interest and emphasis is on the effectiveness of our proposed algorithm **FairFictPlay** on a given dataset, including:

- Whether the algorithm in fact converges, and does so in a feasible amount of computation. Recall that formal convergence is only guaranteed under the assumption of oracles that do not exist in practice, and even then is only guaranteed asymptotically.
- Whether the randomized classifier learned by the algorithm has nontrivial accuracy, as well as strong subgroup fairness properties.
- Whether the algorithm permits nontrivial tuning of the trade-off between accuracy and subgroup fairness.

As discussed in Section 2.1, we note that all of these issues can be investigated entirely in-sample, without concern for generalization performance. Thus for simplicity, despite the fact that our algorithm enjoys all the usual generalization properties depending on the VC dimension of the Learner's hypothesis space and the Auditor's subgroup space (see Theorems 2.9 and 2.8), we report all results here on the full C&C dataset of 1994 points, treating it as the true distribution of interest.

5.2 Algorithm Implementation

The main details in the implementation of **FairFictPlay** are the identification of the model classes for Learner and Auditor, the implementation of the cost sensitive classification oracle and auditing oracle, and the identification of the protected features for Auditor. For our experiments, at each round Learner chooses a linear threshold function over all 122 features. We implement the cost sensitive classification oracle via a two stage regression procedure. In particular, the inputs to the cost sensitive classification oracle are cost vectors c_0, c_1, where the i^{th} element of c_k is the cost of predicting k on datapoint i. We train two linear regression models r_0, r_1 to predict c_0 and c_1 respectively, using all 122 features. Given a new point x, we predict the cost of classifying x as 0 and 1 using our regression models: these predictions are $r_0(x)$ and $r_1(x)$ respectively. Finally we output the prediction $\hat{y}$ corresponding to lower predicted cost: $\hat{y} = \text{argmin}_{i \in \{0,1\}} r_i(x)$.

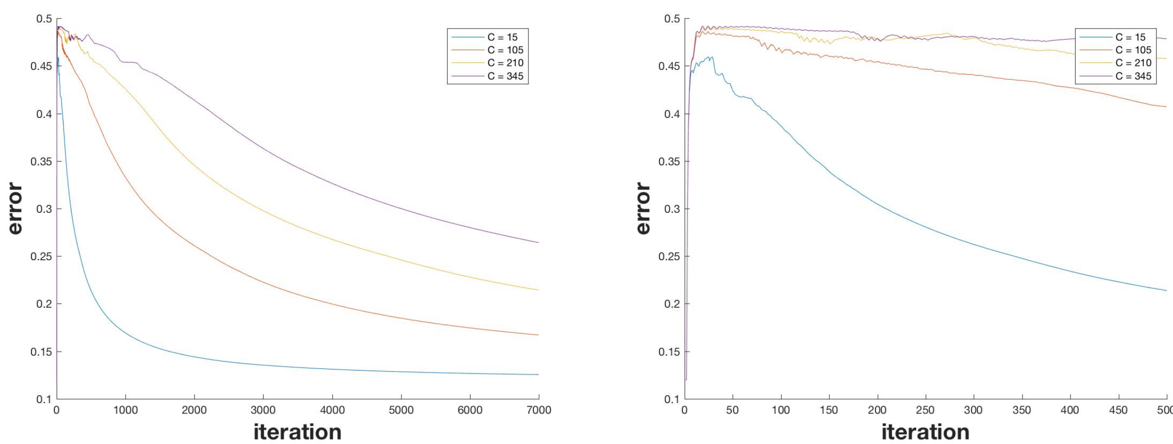
Auditor's model class consists of all linear threshold functions over just the 18 aforementioned protected race-based attributes. As per the algorithm, at each iteration t Auditor attempts to find a subgroup on which the false positive rate is substantially different than the base rate, given the Learner's randomized classifier so far. We implement the auditing oracle by treating it as a weighted regression problem in which the goal is find a linear function (which will be taken to define the subgroup) that on the negative examples, can predict the Learner's probabilistic classification on each point. We use the same regression subroutine as Learner does, except that Auditor only has access to the 18 sensitive features, rather than all 122.

Recall that in addition to the choices of protected attributes and model classes for Learner and Auditor, **FairFictPlay** has a parameter C, which is a bound on the norm of the dual variables for Auditor (the dual player). The theory of Section 4 tells us that if C is larger than some threshold, the dual player is unconstrained and can find their optimal minmax strategy for the game, enforcing perfect subgroup fairness, with Learner achieving whatever accuracy is possible given perfect subgroup fairness.

More generally however, if C is less than the perfect subgroup fairness threshold, there is still a well-defined game in which Auditor's strategy space is constrained, and thus Auditor has less influence on the Learner's objective. Thus in theory, smaller values of C trade weaker subgroup fairness constraints for higher accuracy. More precisely, for any finite C, we know from Theorem 4.6 that at equilibrium the false-positive rate disparity of Learner's randomized classifier on any subgroup is at most on the order of $1/C$, and has error at most that of the optimal perfectly fair classifier (infinite C), and possibly less due to the relaxed fairness constraints. In our experiments we therefore run **FairFictPlay** for a wide range of values of C (from 0 to 500, with greater resolution at smaller values due to higher sensitivity there), in the hope of sketching a Pareto curve of models giving a spectrum of empirically optimal trade-offs between accuracy and subgroup fairness.

5.3 Results

Particularly in light of the gaps between the idealized theory and the actual implementation, the most basic questions about **FairFictPlay** are whether it converges at all, and if so, whether it converges to "interesting" models — that is, models with both nontrivial classification error (much better than the 30% or 0.3 baserate), and nontrivial subgroup fairness (much better than ignoring fairness altogether). We shall see that at least for the C&C dataset, the answers to these questions is strongly affirmative.

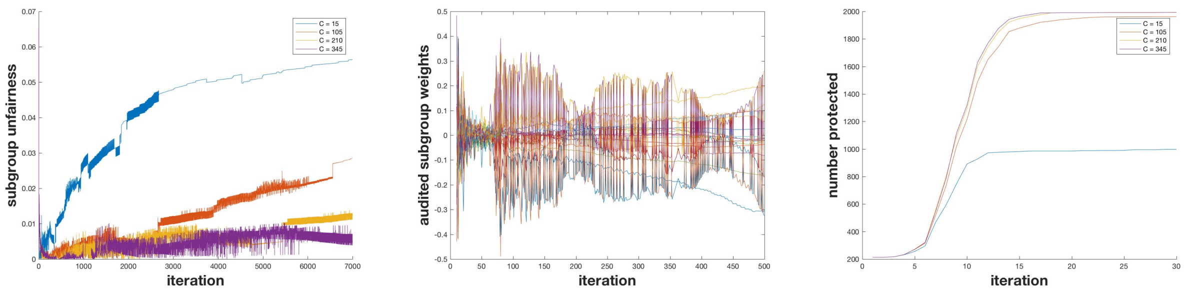


(a) (b)

Figure 1: Evolution of the error of Learner's probabilistic classifier across iterations, for varying choices of C. (a) Evolution to approximate convergence. (b) Early iterations, showing initial rise in error as subgroup fairness is enforced by Auditor.

We begin by examining the evolution of the error of Learner's model. In the left panel of Figure 1 we show the error of the model found by Learner vs. iteration for several values of C. We see that the error is indeed converging smoothly, and furthermore that C is clearly influencing the asymptotic error as well as the rate of convergence — smaller values of C, in which the Auditor's power is more limited, lead to faster convergence to lower error. In the right panel, we zoom in on just the early iterations in order to highlight the influence of the Auditor. All values of C result in an initial model that is simply the optimal model on the data absent any fairness considerations (which is approximately 0.11). Successive iterations of the game, however, begin to impose Auditor's subgroup fairness constraints on the Learner, resulting in a rapid rise of the error. Larger values of C cause a higher and longer rise in the error before its decline to its asymptotic error.

Figure 1 measures the impact of Auditor on Learner only indirectly, via the influence on the error of the model learned. In Figure 2 we more directly measure the quality and amount of fairness enforced by the Auditor for the same values of C. One natural direct measure of the Auditor's impact is to define the unfairness $u(g_t)$ of the Learner's model H_{t-1} on the subgroup g_t chosen by Auditor as follows: $u(g_t)$ is the fraction of datapoints falling in g_t, multiplied by the absolute difference between the FP rate of H_{t-1} on g_t and the background FP rate. This effectively measures the Auditor's ability to find a large subgroup with high FP disparity compared to the base rate, and is analogous to the quantity $\alpha\beta$ in the definition of (α, β)-fairness. Large values of $u(g_t)$ mean that the Auditor has found a relatively highly discriminated subgroup with relatively large mass, whereas small values mean that the Auditor found only relatively small discrimination in a subgroup with relatively little mass.



(a) (b) (c)

Figure 2: (a) Evolution of $u(g_t)$ for Auditor for varying C. (b) Evolution of Auditor's weights on the 18 protected attributes, for $C = 105$. (c) Cumulative population protected by Auditor for varying C. See text for details.

In Figure 2(a), we show the evolution of $u(g_t)$ for the same values of C as in Figure 1. While not easily seen in the plot, for all values of C there is an initial spike in $u(g_t)$, corresponding to the first models chosen by Learner obeying no or few fairness constraints. These spikes diminish rapidly as Auditor enforces fairness, and the process settles into a gradual balance between error and fairness. We also see that for smaller values of C, for which the Auditor's influence is more restricted due to the limited weights in the Lagrangian, $u(g_t)$ is uniformly higher than for larger values of C. For all values of C, $u(g_t)$ generally increases with t — the fairness constraints on the Learner grow with time — but appear to asymptote to values that depend on C in the expected fashion.

Figures 1 and 2(a) together demonstrate that **FairFictPlay** does indeed learn models providing a nontrivial balance of error and subgroup fairness. For example, examining the red curves in these figures corresponding to the choice $C = 105$, we see that at iteration 7000, **FairFictPlay** has found a model with error approximately 0.167 (much better than the strawman error 0.3), and for which the Auditor is unable to find subgroups with unfairness above 0.02 (much lower than the unfairness being found for $C = 15$ at the same iteration, which is close to 0.06, or for the unconstrained model at round 1, which is roughly the same.)

While the analysis so far confirms that the Auditor indeed is enforcing fairness constraints on the Learner, it does not provide a sense of the extent to which the Auditor is finding multiple "different" subgroups over time, and is thus truly utilizing the power of its rich space of models over the restricted attributes (in contrast to the more traditional approach of ensuring fairness with respect to just one or a few pre-defined protected groups). Perhaps the most basic way of measuring the diversity of subgroups found by Auditor is to look at the actual evolution of the subgroup models — in this case, the weight vectors of the linear classifiers found. In Figure 2(b), we show these weights over time for just the choice $C = 105$ (plots for other values of C are similar), and from iteration 10 to 1000 since the earliest weights are large and would dominate the scale. It is clear from this figure that the Auditor indeed explores a diverse set of linear classifiers over time, as many of the 18 protected attribute weights change sign repeatedly, a few in sometimes oscillatory fashion. But a more intuitive measure would be to simply look at how many datapoints are "touched" by the Auditor over time — that is, how many datapoints have fallen in one or more of the subgroups found by the Auditor so far, which we can view as measuring the number of "protected" datapoints. As the Auditor gradually finds more, and more diverse, subgroups to add to the Lagrangian of the Learner, this quantity should increase. We illustrate this phenomenon in Figure 2(c), which shows the cumulative number of datapoints touched by Auditor at each iteration, again for the same selection of values of C. For all C the y value starts at 213, since the very first linear classifier found by the Auditor defines a subgroup of 213 datapoints. For the smallest value of C ($C = 15$), the value eventually climbs to roughly half the dataset, while for the larger values of C the entire dataset is eventually touched by the Auditor. Note that this does not indicate that fairness is obtained for every datapoint or subgroup, just that the Auditor has presented Learner with fairness constraints involving every datapoint. Note that these values are reached very early in the process (the first 30 iterations), despite the fact that the other plots clearly indicate that the balance between accuracy and fairness continues long after.

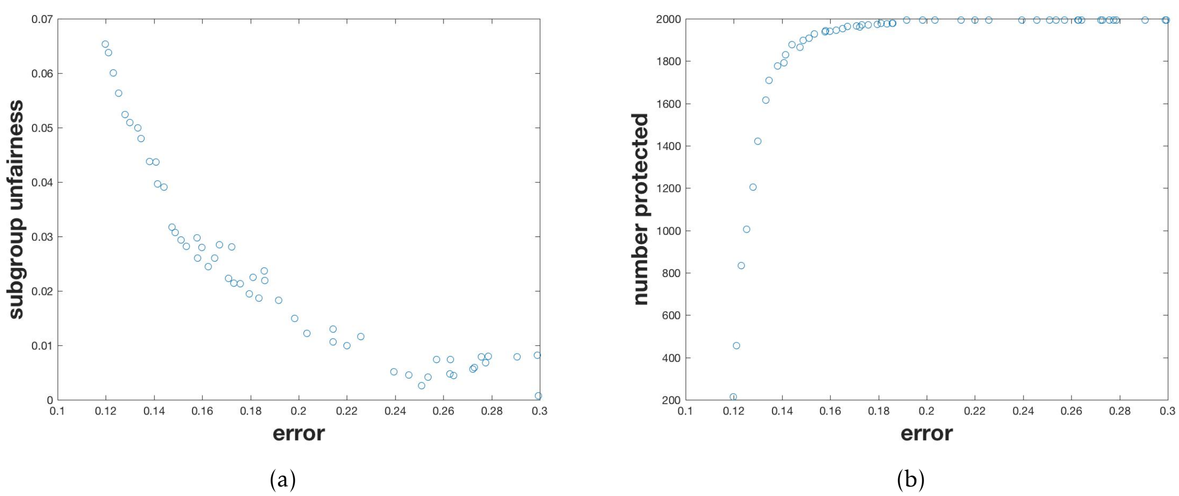


Figure 3: Pareto frontiers explored by varying C across a wide range. (a) Learner's error vs. subgroup unfairness. (b) Learner's error vs. number of protected data points. See text for details.

We also illustrate the full spectrum of error-fairness trade-offs yielded by running **FairFictPlay** with values of C ranging from 0 to 500. In Figure 3(a), each point represents both the error (x axis) and the subgroup unfairness $u(g_T)$ at $T = 7000$ for varying values of C. While not all points lie exactly on the Pareto frontier due to the differences in convergence rates, a clear menu of near-optimal trade-offs emerges. We can have error near the unconstrained optimal of 0.11, and subgroup unfairness over 0.06. At the other extreme, we can have error no better than the trivial baseline of the constant classifier of 0.3, and near-zero subgroup unfairness. In between there is an appealing curve of trade-offs, but one that also cannot be escaped without enriching the model space of the Learner or the Auditor or both. Figure 3(b) shows a similar frontier, this time between error and the number of datapoints touched by Auditor as discussed above. Again we see a menu of trade-off options, from optimal error and few subgroup members, to progressively worse error and the entire population involved in audited subgroups.

It is intuitive that one can construct (as we did in the introduction) artificial examples in which classifiers which equalize false positive rates across groups defined only with respect to individual protected binary features can exhibit unfairness in more complicated subgroups. However, it might be the case that on real-world datasets, enforcing false positive rate fairness in marginal subgroups using previously known algorithms (like Agarwal et al. [2017]) would already provide at least approximate fairness in the combinatorially many subgroups defined by a simple (e.g. linear threshold) function over the protected features. To investigate this possibility, we implemented the algorithm of Agarwal et al. [2017], which employs a similar optimization framework. In their algorithm the "primal player" plays the same weighted classification oracle we use, and the dual player plays gradient descent over a space of dimension equal to the number of protected groups. We used the same Communities and Crime dataset with the same 18 protected features. Our 18 protected attributes are real valued. In order to come up with a small number of protected groups, we threshold each real-valued attribute at its mean, and define 36 protected groups: each one corresponding to one of the protected attributes lying either above or below its mean.

We then ran the algorithm from Agarwal et al. [2017], using a learning rate of $\frac{1}{\sqrt{t}}$ at time step t in the gradient descent step. After just 13 iterations, across all 36 protected groups defined on the single protected attributes, the false positive rate disparity was already below 0.03, and the classifier had achieved non-trivial error (not far above the unconstrained optimal), thus successfully balancing accuracy with fairness on the small number of pre-defined subgroups. However, upon auditing the resulting classifier with respect to the richer class of linear threshold functions on the continuously-valued protected features, we uncovered a large subgroup whose false positive rate differed substantially from the baseline. This subgroup had weight 0.674 (consisting of well over half of the datapoints), and a false positive rate that was higher than the base rate by 0.26 — a 61% increase. While the discriminated subgroup is of course defined by a complex linear threshold function over 18 variables, the largest weights by far were on only three of these features; the subgroup can be informally interpreted as a disjunction identifying communities where the percentage of the police forces that are Black or Hispanic are relatively high, or where the percentage that is Asian is relatively low.

This simple experiment illustrates that in practice it may be easy to learn classifiers which appear fair with respect to the marginal groups given by pre-defined protected features, but may discriminate significantly against the members of a simple combinatorial subgroup.

Acknowledgements We thank Alekh Agarwal, Richard Berk, Miro Dudík, John Langford, Greg Ridgeway and Greg Yang for helpful discussions and suggestions.

References

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, and John Langford. A reductions approach to fair classification. *Fairness, Accountability, and Transparency in Machine Learning (FATML)*, 2017.

Julia Angwin and Hannes Grassegger. Facebooks secret censorship rules protect white men from hate speech but not black children. *Propublica*, 2017.

Anna Maria Barry-Jester, Ben Casselman, and Dana Goldstein. The new science of sentencing. *The Marshall Project*, August 8 2015. Retrieved 4/28/2016.

George W. Brown. Some notes on computation of games solutions, Jan 1949.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *arXiv preprint arXiv:1703.00056*, 2017.

Constantinos Daskalakis and Qinxuan Pan. A counter-example to Karlin's strong conjecture for fictitious play. In *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*, pages 11–20. IEEE, 2014.

Ilias Diakonikolas, Ryan O'Donnell, Rocco A. Servedio, and Yi Wu. Hardness results for agnostically learning low-degree polynomial threshold functions. In *Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2011, San Francisco, California, USA, January 23-25, 2011*, pages 1590–1606, 2011.

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.

Vitaly Feldman, Venkatesan Guruswami, Prasad Raghavendra, and Yi Wu. Agnostic learning of monomials by halfspaces is hard. *SIAM J. Comput.*, 41(6):1558–1590, 2012.

Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236*, 2016.

Sara Hajian and Josep Domingo-Ferrer. A methodology for direct and indirect discrimination prevention in data mining. *IEEE transactions on knowledge and data engineering*, 25(7):1445–1459, 2013.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 2016.

Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*, pages 325–333, 2016.

Adam Tauman Kalai, Yishay Mansour, and Elad Verbin. On agnostic boosting and parity learning. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing, Victoria, British Columbia, Canada, May 17-20, 2008*, pages 629–638, 2008.

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

Michael J Kearns and Umesh Virkumar Vazirani. *An Introduction to Computational Learning Theory*. MIT press, 1994.

Michael J Kearns, Robert E Schapire, and Linda M Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2-3):115–141, 1994.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *Proceedings of the 2017 ACM Conference on Innovations in Theoretical Computer Science, Berkeley, CA, USA, 2017*, 2017.

James Rufus Koren. What does that web search say about your credit? *Los Angeles Times*, July 16 2016. Retrieved 9/15/2016.

Julia Robinson. An iterative method of solving a game. *Annals of Mathematics*, pages 10–2307, 1951.

Cynthia Rudin. Predictive policing using machine learning to detect patterns of crime. *Wired Magazine*, August 2013. Retrieved 4/28/2016.

Maurice Sion. On general minimax theorems. *Pacific J. Math.*, 8(1):171–176, 1958.

Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. *arXiv preprint arXiv:1702.06081*, 2017.

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

Zhe Zhang and Daniel B Neill. Identifying significant predictive bias in classifiers. *arXiv preprint arXiv:1611.08292*, 2016.

A Generalization Bounds

Proof of Theorems 2.8 and 2.9. We give a proof of Theorem 2.8. The proof of Theorem 2.9 is identical, as false positive rates are just positive classification rates on the subset of the data for which $y = 0$.

Given a set of classifiers $\mathcal{H}$ and protected groups $\mathcal{G}$, define the following function class:

$$\mathcal{F}_{\mathcal{H},\mathcal{G}} = \{f_{h,g}(x) \doteq h(x) \wedge g(x) : h \in \mathcal{H}, g \in \mathcal{G}\}$$

We can relate the VC-dimension of $\mathcal{F}_{\mathcal{H},\mathcal{G}}$ to the VC-dimension of $\mathcal{H}$ and $\mathcal{G}$:

Claim A.1.

$$\mathrm{VCDIM}(\mathcal{F}_{\mathcal{H},\mathcal{G}}) \leq \tilde{O}(\mathrm{VCDIM}(\mathcal{H}) + \mathrm{VCDIM}(\mathcal{G}))$$

Proof. Let S be a set of size m shattered by $\mathcal{F}_{\mathcal{H},\mathcal{G}}$. Let $\pi_{\mathcal{F}_{\mathcal{H},\mathcal{G}}}(S)$ be the number of labelings of S realized by elements of $\mathcal{F}_{\mathcal{H},\mathcal{G}}$. By the definition of shattering, $\pi_{\mathcal{F}_{\mathcal{H},\mathcal{G}}}(S) = 2^m$. Now for each labeling of S by an element in $\mathcal{F}_{\mathcal{H},\mathcal{G}}$, it is realized as $(f \wedge g)(S)$ for some $f \in \mathcal{F}, g \in \mathcal{G}$. But $(f \wedge g)(S) = f(S) \wedge g(S)$, and so it can be realized as the conjunction of a labeling of S by an element of $\mathcal{F}$ and an element of $\mathcal{G}$. But since there are $\pi_{\mathcal{F}}(S)\pi_{\mathcal{G}}(S)$ such pairs of labelings, this immediately implies that $\pi_{\mathcal{F}_{\mathcal{H},\mathcal{G}}}(S) \leq \pi_{\mathcal{F}}(S)\pi_{\mathcal{G}}(S)$. Now by the Sauer-Shelah Lemma (see e.g. Kearns and Vazirani [1994]), $\pi_{\mathcal{F}}(S) = O(m^{\mathrm{VCDIM}(\mathcal{H})}), \pi_{\mathcal{G}}(S) = O(m^{\mathrm{VCDIM}(\mathcal{G})})$. Thus $\pi_{\mathcal{F}_{\mathcal{H},\mathcal{G}}}(S) = 2^m \leq O(m^{\mathrm{VCDIM}(\mathcal{H})+\mathrm{VCDIM}(\mathcal{G})})$, which implies that $m = \tilde{O}(\mathrm{VCDIM}(\mathcal{H}) + \mathrm{VCDIM}(\mathcal{G}))$, as desired. □

This bound, together with a standard VC-Dimension based uniform convergence theorem (see e.g. Kearns and Vazirani [1994]) implies that with probability $1-\delta$, for every $f_{h,g} \in \mathcal{F}_{\mathcal{H},\mathcal{G}}$:

$$\left|\mathbb{E}_{(X,y)\sim\mathcal{P}}[f_{h,g}(X)] - \mathbb{E}_{(X,y)\sim S}[f_{h,g}(X)]\right| \leq \tilde{O}\left(\sqrt{\frac{(\mathrm{VCDIM}(\mathcal{H}) + \mathrm{VCDIM}(\mathcal{G}))\log m + \log(1/\delta)}{m}}\right)$$

The left hand side of the above inequality can be written as:

$$\left|\Pr_{(X,y)\sim\mathcal{P}}[h(X)=1|g(x)=1]\cdot\Pr_{(X,y)\sim\mathcal{P}}[g(x)=1]-\Pr_{(X,y)\sim S}[h(X)=1|g(x)=1]\cdot\Pr_{(X,y)\sim S}[g(x)=1]\right|$$

Thus, dividing both sides by $\Pr_{(X,y)\sim\mathcal{P}}[g(x)=1]$ and noting that by assumption this quantity is at least α, we obtain:

$$\left|\Pr_{(X,y)\sim\mathcal{P}}[h(X)=1|g(x)=1]-\Pr_{(X,y)\sim S}[h(X)=1|g(x)=1]\cdot\frac{\Pr_{(X,y)\sim S}[g(x)=1]}{\Pr_{(X,y)\sim\mathcal{P}}[g(x)=1]}\right|\leq$$

$$\tilde{O}\left(\frac{1}{\alpha}\sqrt{\frac{(\mathrm{VCDIM}(\mathcal{H})+\mathrm{VCDIM}(\mathcal{G}))\log m+\log(1/\delta)}{m}}\right)$$

By the triangle inequality, we can bound

$$\left|\Pr_{(X,y)\sim\mathcal{P}}[h(X)=1|g(x)=1]-\Pr_{(X,y)\sim S}[h(X)=1|g(x)=1]\right|\leq$$

$$\left|\Pr_{(X,y)\sim\mathcal{P}}[h(X)=1|g(x)=1]-\Pr_{(X,y)\sim S}[h(X)=1|g(x)=1]\cdot\frac{\Pr_{(X,y)\sim S}[g(x)=1]}{\Pr_{(X,y)\sim\mathcal{P}}[g(x)=1]}\right|+$$
$$\left|\Pr_{(X,y)\sim S}[h(X)=1|g(x)=1]\cdot\frac{\Pr_{(X,y)\sim S}[g(x)=1]}{\Pr_{(X,y)\sim\mathcal{P}}[g(x)=1]}-\Pr_{(X,y)\sim S}[h(X)=1|g(x)=1]\right|$$

We have already bounded the first term: it remains to bound the second term. Again, invoking a standard VC-dimension bound, we obtain:

$$\Pr_{(X,y)\sim S}[g(x)=1]\leq\Pr_{(X,y)\sim\mathcal{P}}[g(x)=1]+O\left(\sqrt{\frac{\mathrm{VCDIM}(\mathcal{G})\log m+\log(1/\delta)}{m}}\right)$$

Since by assumption we have $\Pr_{(X,y)\sim\mathcal{P}}[g(x)=1]\geq\alpha$, we therefore know that with probability $1-\delta$:

$$\frac{\Pr_{(X,y)\sim S}[g(x)=1]}{\Pr_{(X,y)\sim\mathcal{P}}[g(x)=1]}\leq 1+O\left(\frac{1}{\alpha}\sqrt{\frac{\mathrm{VCDIM}(\mathcal{G})\log m+\log(1/\delta)}{m}}\right)$$

Finally, this lets us bound the 2nd term above:

$$\left|\Pr_{(X,y)\sim S}[h(X)=1|g(x)=1]\cdot\frac{\Pr_{(X,y)\sim S}[g(x)=1]}{\Pr_{(X,y)\sim\mathcal{P}}[g(x)=1]}-\Pr_{(X,y)\sim S}[h(X)=1|g(x)=1]\right|\leq O\left(\frac{1}{\alpha}\sqrt{\frac{\mathrm{VCDIM}(\mathcal{G})\log m+\log(1/\delta)}{m}}\right)$$

which completes the proof. □

B Relaxing the Base Rate Condition

Lemma B.1. *Suppose that the base rate* $b_{SP}(D,P)\in[1/2-\varepsilon,1/2+\varepsilon]$ *and there exists a function* f *such that* $\Pr[f(x)=1]=\alpha$ *and* $|\Pr[D(X)=1\mid f(x)=1]-b_{SP}(D,P)|>\beta$*. Then*

$$\max\{\Pr[D(X)=f(x)],\Pr[D(X)=\neg f(x)]\}\geq(b_{SP}-2\varepsilon)+\alpha\beta$$

Proof. Let $b = b_{SP}(D, p)$ denote the base rate. Given that $|\Pr[D(X) = 1 \mid f(x) = 1] - b| > \beta$, we know either $\Pr[D(X) = 1 \mid f(x) = 1] > b + \beta$ or $\Pr[D(X) = 1 \mid f(x) = 1] < b - \beta$.

In the first case, we know $\Pr[D(X) = 1 \mid f(x) = 0] < b$, and so $\Pr[D(X) = 0 \mid f(x) = 0] > 1 - b$. It follows that

$$\begin{aligned}
\Pr[D(X) = f(x)] &= \Pr[D(X) = f(x) = 1] + \Pr[D(X) = f(x) = 0] \\
&= \Pr[D(X) = 1 \mid f(x) = 1]\Pr[f(x) = 1] + \Pr[D(X) = 0 \mid f(x) = 0]\Pr[f(x) = 0] \\
&> \alpha(b + \beta) + (1 - \alpha)(1 - b) = (\alpha - 1)b + (1 - \alpha)(1 - b) + b + \alpha\beta = (1 - \alpha)(1 - 2b) + b + \alpha\beta
\end{aligned}$$

In the second case, we have $\Pr[D(X) = 0 \mid f(x) = 1] > (1-b)+\beta$ and $\Pr[D(X) = 1 \mid f(x) = 0] > b$. We can then bound

$$\begin{aligned}
\Pr[D(X) = f(x)] &= \Pr[D(X) = 1 \mid f(x) = 0]\Pr[f(x) = 0] + \Pr[D(X) = 0 \mid f(x) = 1]\Pr[f(x) = 1] \\
&> (1 - \alpha)b + \alpha(1 - b + \beta) = \alpha(1 - 2b) + b + \alpha\beta.
\end{aligned}$$

In both cases, we have, and $(1 - 2b) \in [-2\varepsilon, 2\varepsilon]$ based our assumption on the base rate. Since $(1 - \alpha), \alpha \in [0, 1]$, we have

$$\max\{\Pr[D(X) = f(x)], \Pr[D(X) = \neg f(x)]\} \geq b - 2\varepsilon + \alpha\beta$$

□

Lemma B.2. *Suppose that the base rate $b_{SP}(D, P) \in [1/2 - \varepsilon, 1/2 + \varepsilon]$ and there exists a function f such that $\Pr[D(X) = f(x)] \geq b_1(D, P) + \gamma$ for some value $\gamma \in (0, 1/2)$. Then there exists a function g such that $\Pr[g(x) = 1] \geq \gamma - 2\varepsilon$ and $|\Pr[D(X) = 1 \mid g(x) = 1] - b_{SP}(D, p)| > \gamma - 2\varepsilon$, where $g \in \{f, \neg f\}$.*

Proof. Let $b = b_{SP}(D, p)$ denote the base rate. Again recall that

$$\begin{aligned}
\Pr[D(X) = f(x)] &= \Pr[D(X) = f(x) = 1] + \Pr[D(X) = f(x) = 0] \\
&= \Pr[D(X) = 1 \mid f(x) = 1]\Pr[f(x) = 1] + \Pr[D(X) = 0 \mid f(x) = 0]\Pr[f(x) = 0]
\end{aligned}$$

Since $\Pr[D(X) = f(x)] \geq b + \gamma$ and $\Pr[f(x) = 1] + \Pr[f(x) = 0] = 1$,

$$\max\{\Pr[D(X) = 1 \mid f(x) = 1], \Pr[D(X) = 0 \mid f(x) = 0]\} \geq b + \gamma.$$

Thus, we must have either

$$\Pr[D(X) = 1 \mid f(x) = 1] \geq b + \gamma \qquad \text{or} \qquad \Pr[D(X) = 1 \mid f(x) = 0] \leq (1 - b) - \gamma \leq b + 2\varepsilon - \gamma.$$

Next, we observe that $\min\{\Pr[f = 1], \Pr[f = 0]\} \geq \gamma - 2\varepsilon$. This follows since $\Pr[D(x, x') = f] = \Pr[D(x, x') = f = 1] + \Pr[D(x, x') = f = 0] \leq \Pr[f = 1] + \Pr[D = 0]$. Furthermore,

$$\Pr[f = 1] \geq \Pr[D(x, x') = f] - \Pr[D = 0] \geq b + \gamma - (1 - b) = (2b - 1) + \gamma.$$

Similarly, $\Pr[D(x, x') = f = 1] + \Pr[D(x, x') = f = 0] \leq \Pr[f = 0] + \Pr[D = 1]$

$$\Pr[f = 0] \geq \Pr[D(x, x') = f] - \Pr[D = 1] \geq b + \gamma - b = \gamma.$$

This completes our proof. □

Theorem B.3 (Extension of Theorem 3.2)**.** *Fix any distribution P over individual data points, any set of group indicators $\mathcal{G}$, and any classifier D. Suppose that the (statistical parity) base rate $b_{SP}(D, P) \in [1/2 - \varepsilon, 1/2 + \varepsilon]$ for some $\varepsilon > 0$. The following two relationships hold:*

- *If D is $(\gamma - 2\varepsilon, \gamma - 2\varepsilon, \alpha', \beta')$-auditable for statistical parity fairness under distribution P and group indicator set $\mathcal{G}$, then the class $\mathcal{G}$ is $(\gamma, \alpha'\beta' - 3\varepsilon)$-weakly agnostically learnable under P^D.*

- *If $\mathcal{G}$ is $(\alpha\beta - 3\varepsilon, \gamma)$-weakly agnostically learnable under distribution P^D, then D is $(\alpha, \beta, \gamma - 3\varepsilon, \gamma - 3\varepsilon)$-auditable for statistical parity fairness under P and group indicator set $\mathcal{G}$.*

Proof. Suppose that the class $\mathcal{G}$ satisfies $\min_{f \in \mathcal{G}} err(f, P^D) \leq 1/2 - \gamma$. Then by Lemma B.2, there exists some $g \in \mathcal{G}$ such that $\Pr[g(x) = 1] \geq \gamma - 2\varepsilon$ and $|\Pr[D(X) = 1 \mid g(x) = 1] - b_1(D, p)| > \gamma - 2\varepsilon$. By the assumption of auditability, we can then use the auditing algorithm to find a group $g' \in \mathcal{G}$ that is an (α', β')-unfair certificate of D. By Lemma B.1, we know that either g' or $\neg g'$ predicts D with an accuracy of at least $1/2 + \alpha'\beta' - 3\varepsilon$.

In the reverse direction: consider the auditing problem on the classifier D. We can treat each pair $(x, D(X))$ as a labelled example and learn a hypothesis in $\mathcal{G}$ that approximates the decisions made by D. Suppose that D is (α, β)-unfair. Then by Lemma B.1, we know that there exists some $g \in \mathcal{G}$ such that $\Pr[D(X) = g(x)] \geq 1/2 + \alpha\beta - 3\varepsilon$. Therefore, the weak agnostic learning algorithm from the hypothesis of the theorem will return some g' with $\Pr[D(X) = g'(x)] \geq 1/2 + \gamma$. By Lemma B.2, we know g' or $\neg g'$ is a $(\gamma - 3\varepsilon, \gamma - 3\varepsilon)$-unfair certificate for D. $\square$