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-New York City

April 16, 2018

Abstract

The most prevalent notions of fairness in machine learning are *statistical* definitions: they fix a small collection of high-level, pre-defined groups (such as race or gender), and then ask for approximate parity of some statistic of the classifier (like positive 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 two algorithms that provably converge to the best fair distribution over classifiers in a given class, given access to oracles which can optimally solve the agnostic learning problem. The algorithms are based on a formulation of subgroup fairness as a two-player zero-sum game between a Learner (the primal player) and an Auditor (the dual player). Both algorithms compute an equilibrium of this game. We obtain our first algorithm by simulating play of the game by having Learner play an instance of the no-regret *Follow the Perturbed Leader* algorithm, and having Auditor play best response. This algorithm provably converges to an approximate Nash equilibrium (and thus to an approximately optimal subgroup-fair distribution over classifiers) in a polynomial number of steps. We obtain our second algorithm by simulating play of the game by having both players play *Fictitious Play*, which enjoys only provably asymptotic convergence, but has the merit of simplicity and faster per-step computation. We implement the Fictitious Play version 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 have attractively strong semantics, but their main drawback is that achieving them seemingly requires 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, in the sense that 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 as black women), then it is apparent that the classifier maximally violates the statistical parity fairness constraint. Similarly, if examples have a binary label that is also distributed uniformly at random, and independently from the features, the classifier will satisfy equal opportunity fairness with respect to either protected attribute alone, even though it maximally violates it with respect to conjunctions of two attributes.*

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.
- Provably convergent algorithms for learning classifiers that are fair with respect to $\mathcal{G}$, based on a formulation as a two-player zero-sum game between a Learner (the primal player) and an Auditor (the dual player). We provide two different algorithms, both of which are based on solving for the equilibrium of this game. The first provably converges in a polynomial number of steps and is based on simulation of the game dynamics when the Learner uses *Follow the Perturbed Leader* and the Auditor uses best response; the second is only guaranteed to converge asympotically but is computationally simpler, and involves both players using *Fictitious Play*.
- An implementation and extensive empirical evaluation of the Fictitious Play 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, their number is combinatorially large[2].

[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.

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 $g \in \mathcal{G}$ corresponds to the protected subgroup $\{x : g_i(x) = 1\}$[3]. The first result of this paper is that for both equal opportunity and statistical parity, the computational problem of *checking* whether a classifier or decision-making algorithm D violates statistical fairness with respect to the set of protected groups $\mathcal{G}$ is equivalent to the problem of *agnostically learning* $\mathcal{G}$ [Kearns et al., 1994], in a strong and distribution-specific sense. 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 boosting, 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 or negative 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 an efficient oracles which solves the *agnostic learning* problem, we give and analyze algorithms for this problem based on a game-theoretic formulation. We first prove that the optimal fair classifier can be found as the equilibrium of a two-player, zero-sum game, in which the (pure) strategy space of the "Learner" player corresponds to classifiers in $\mathcal{H}$, and the (pure) strategy space of the "Auditor" player corresponds to subgroups defined by $\mathcal{G}$. The best response problems for the two players correspond to agnostic learning and auditing, respectively. We show that both problems can be solved with a single call to a *cost sensitive classification oracle*, which is equivalent to an agnostic learning oracle. We then draw on extant theory for learning in games and no-regret algorithms to derive two different algorithms based on simulating game play in this formulation. In the first, the Learner employs the well-studied *Follow the Perturbed Leader (FTPL)* algorithm on an appropriate linearization of its best-response problem, while the Auditor approximately best-responds to the distribution over classifiers of the Learner at each step. Since FTPL has a no-regret guarantee, we obtain an algorithm that provably converges in a polynomial number of steps.

While it enjoys strong provable guarantees, this first algorithm is randomized (due to the noise added by FTPL), and the best-response step for the Auditor is polynomial time but computationally expensive. We thus propose a second algorithm that is deterministic, simpler and faster per step, based on both players adopting the Fictitious Play learning dynamic. This algorithm has weaker theoretical guarantees: it has provable convergence only asymptotically, and not in a polynomial number of steps — but is more practical and converges rapidly in practice. The derivation of these algorithms (and their guarantees) appear in Section 4.

[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.

Finally, we implement the Fictitious Play 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

Independent of our work, Hébert-Johnson et al. [2017] also consider a related and complementary notion of fairness that they call "multicalibration". In settings in which one wishes to train a real-valued predictor, multicalibration can be considered the "calibration" analogue for the definitions of subgroup fairness that we give for false positive rates, false negative rates, and classification rates. For a real-valued predictor, calibration informally requires that for every value $v \in [0,1]$ predicted by an algorithm, the fraction of individuals who truly have a positive label in the subset of individuals on which the algorithm predicted v should be approximately equal to v. Multicalibration asks for approximate calibration on every set defined implicitly by some circuit in a set $\mathcal{G}$. Hébert-Johnson et al. [2017] give an algorithmic result that is analogous to the one we give for learning subgroup fair classifiers: a polynomial time algorithm for learning a multi-calibrated predictor, given an agnostic learning algorithm for $\mathcal{G}$. In addition to giving a polynomial-time algorithm, we also give a practical variant of our algorithm (which is however only guaranteed to converge in the limit) that we use to conduct empirical experiments on real data.

Thematically, the most closely related piece of prior 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 particular 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* sub-groups, 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 (and 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 agnostic learning oracle is sufficient to have the both players play "fictitious play", and that there is a transformation of

the best response problem such that an agnostic learning algorithm is enough to efficiently implement follow the perturbed leader.

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 and Preliminaries

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 $(x,y,D(X))$ that results from sampling $(x,x',y) \sim \mathcal{P}$, and providing x, the true label y, and the classification $D(X) = D(x,x')$ but not the unprotected attributes 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 parameter $\gamma \in [0,1]$. For each $g \in \mathcal{G}$, define*

$$\alpha_{SP}(g,\mathcal{P}) = \Pr_{\mathcal{P}}[g(x) = 1] \quad \textit{and,} \quad \beta_{SP}(g,D,\mathcal{P}) = |\mathrm{SP}(D) - \mathrm{SP}(D,g)|,$$

where $\mathrm{SP}(D) = \Pr_{\mathcal{P},D}[D(X) = 1]$ *and* $\mathrm{SP}(D,g) = \Pr_{\mathcal{P},D}[D(X) = 1 | g(x) = 1]$ *denote the overall acceptance rate of D and the acceptance rate of D on group g respectively. We say that D satisfies* γ-statistical parity (SP) Fairness *with respect to $\mathcal{P}$ and $\mathcal{G}$ if for every $g \in \mathcal{G}$*

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

We will sometimes refer to $\mathrm{SP}(D)$ *as the* SP base rate.

Remark 2.2. *Note that our definition references two approximation parameters, both of which are important. We are allowed to ignore 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 up to β. Both of these approximation parameters are necessary from a statistical estimation perspective. We control both of them with a single parameter γ.

Definition 2.3 (False Positive (FP) Subgroup Fairness)**.** *Fix any classifier D, distribution $\mathcal{P}$, collection of group indicators $\mathcal{G}$, and parameter $\gamma \in [0,1]$. For each $g \in \mathcal{G}$, define*

$$\alpha_{FP}(g,\mathcal{P}) = \Pr_{\mathcal{P}}[g(x)=1, y=0] \quad \textit{and,} \quad \beta_{FP}(g,D,\mathcal{P}) = |\mathrm{FP}(D) - \mathrm{FP}(D,g)|$$

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

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

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

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

Remark 2.4. *This definition is symmetric to the definition of statistical parity fairness, except that the parameter α is now used to exclude any group g such that* negative examples *($y=0$) from g (or its complement) have probability mass less than α. This is again necessary from a statistical estimation perspective.*

For either statistical parity and false positive fairness, if the algorithm D fails to satisfy the γ-fairness condition, then we say that D is *γ-unfair* with respect to $\mathcal{P}$ and $\mathcal{G}$. We call any subgroup g which witnesses this unfairness an *γ-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 Algorithm)**.** *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 $\delta, \gamma, \gamma' \in (0,1)$ such that $\gamma' \leq \gamma$. A (γ,γ')-auditing algorithm for $\mathcal{G}$ with respect to distribution $\mathcal{P}$ is an algorithm $\mathcal{A}$ such that for any classifier D, when given access the distribution $P_{\text{audit}}(D)$, $\mathcal{A}$ runs in time* $\mathrm{poly}(1/\gamma', \log(1/\delta))$*, and with probability $1-\delta$, outputs a γ'-unfair certificate for D whenever D is γ-unfair with respect to $\mathcal{P}$ and $\mathcal{G}$. If D is γ'-fair, $\mathcal{A}$ will output "fair".*

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

Definition 2.6 (Weak Agnostic Learning [Kearns et al., 1994, Kalai et al., 2008])**.** *Let Q be a distribution over $\mathcal{X} \times \{0,1\}$ and let $\varepsilon, \varepsilon' \in (0,1/2)$ such that $\varepsilon \geq \varepsilon'$. We say that the function class $\mathcal{G}$ is* $(\varepsilon,\varepsilon')$-weakly agnostically learnable *under distribution Q if there exists an algorithm L such that when given sample access to Q, L runs in time* $\mathrm{poly}(1/\varepsilon', 1/\delta)$*, and with probability $1-\delta$, outputs a hypothesis $h \in \mathcal{G}$ such that*

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

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

Cost-Sensitive Classification. In this paper, we will also give reductions to *cost-sensitive classification (CSC)* problems. Formally, an instance of a CSC problem for the class $\mathcal{H}$ is given by a set of n tuples $\{(X_i, c_i^0, c_i^1)\}_{i=1}^n$ such that c_i^ℓ corresponds to the cost for predicting label ℓ on point X_i. Given such an instance as input, a CSC oracle finds a hypothesis $\hat{h} \in \mathcal{H}$ that minimizes the total cost across all points:

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

A crucial property of a CSC problem is that the solution is invariant to translations of the costs.

Claim 2.7. *Let $\{(X_i, c_i^0, c_i^1)\}_{i=1}^n$ be a CSC instance, and $\{(\tilde{c}_i^0, \tilde{c}_i^1)\}$ be a set of new costs such that there exist $a_1, a_2, \ldots, a_n \in \mathbb{R}$ such that $\tilde{c}_i^\ell = c_i^\ell + a_i$ for all i and ℓ. Then*

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

Remark 2.8. *We note that cost-sensitive classification is polynomially equivalent to agnostic learning Zadrozny et al. [2003]. We give both definitions above because when describing our results for auditing, we wish to directly appeal to known hardness results for weak agnostic learning, but it is more convenient to describe our algorithms via oracles for cost-sensitive classification.*

Follow the Perturbed Leader. We will make use of the *Follow the Perturbed Leader (FTPL)* algorithm as a no-regret learner for online linear optimization problems [Kalai and Vempala, 2005]. To formalize the algorithm, consider $\mathcal{S} \subset \{0,1\}^d$ to be a set of "actions" for a learner in an online decision problem. The learner interacts with an adversary over T rounds, and in each round t, the learner (randomly) chooses some action $a^t \in \mathcal{S}$, and the adversary chooses a loss vector $\ell^t \in [-M, M]^d$. The learner incurs a loss of $\langle \ell^t, a^t \rangle$ at round t.

FTPL is a simple algorithm that in each round perturbs the cumulative loss vector over the previous rounds $\overline{\ell} = \sum_{s<t} \ell^s$, and chooses the action that minimizes loss with respect to the perturbed cumulative loss vector. We present the full algorithm in Algorithm 1, and its formal guarantee in Theorem 2.9.

Algorithm 1 Follow the Perturbed Leader (FTPL) Algorithm

Input: Loss bound M, action set $\mathcal{S} \in \{0,1\}^d$

Initialize: Let $\eta = (1/M)\sqrt{\frac{1}{\sqrt{d}T}}$, $\mathcal{D}_U$ be the uniform distribution over $[0,1]^d$, and let $a^1 \in \mathcal{S}$ be arbitrary.

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

Play action a^t; Observe loss vector ℓ^t and suffer loss $\langle \ell^t, a^t \rangle$.

Update:

$$a^{t+1} = \operatorname*{argmin}_{a\in\mathcal{S}} \left[\eta \sum_{s\le t} \langle \ell^s, a \rangle + \langle \xi^t, a \rangle \right]$$

where ξ^t is drawn independently for each t from the distribution $\mathcal{D}_U$.

Theorem 2.9 (Kalai and Vempala [2005])**.** *For any sequence of loss vectors $\ell^1, \ldots, \ell^T$, the FTPL algorithm has regret*

$$\mathbb{E}\left[\sum_{t=1}^{T} \langle \ell^t, a^t \rangle\right] - \min_{a\in\mathcal{S}} \sum_{t=1}^{T} \langle \ell^t, a \rangle \le 2d^{5/4} M \sqrt{T}$$

where the randomness is taken over the perturbations ξ^t across rounds.

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 γ-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 (or equivalently, that the empirical data defines the distribution of interest), and do not make further reference to sample complexity or overfitting issues.

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

Theorem 2.10. *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:*

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

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 $m \geq \tilde{O}(\text{VCDIM}(\mathcal{H})/\varepsilon^2)$, 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 2.11 (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 $g \in \mathcal{G}$*

$$\left|\alpha_{SP}(g,\mathcal{P}_S)\,\beta_{SP}(g,h,\mathcal{P}_S) - \alpha_{SP}(g,\mathcal{P})\,\beta_{SP}(g,h,\mathcal{P})\right| \leq \tilde{O}\left(\sqrt{\frac{(\text{VCDIM}(\mathcal{H}) + \text{VCDIM}(\mathcal{G}))\log m + \log(1/\delta)}{m}}\right)$$

where $\mathcal{P}_S$ denotes the empirical distribution over the realized sample S.

Similarly:

Theorem 2.12 (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 $g \in \mathcal{G}$, we have:*

$$\left|\alpha_{FP}(g,\mathcal{P})\,\beta_{FP}(g,D,\mathcal{P}) - \alpha_{FP}(g,\mathcal{P})\,\beta_{FP}(g,D,\mathcal{P})\right| \leq \tilde{O}\left(\sqrt{\frac{(\text{VCDIM}(\mathcal{H}) + \text{VCDIM}(\mathcal{G}))\log m + \log(1/\delta)}{m}}\right)$$

where $\mathcal{P}_S$ denotes the empirical distribution over the realized sample S.

These theorems together imply that for both SP and FP subgroup fairness, the degree to which a group g violates the constraint of γ-fairness can be estimated up to error ε, so long as $m \geq \tilde{O}((\text{VCDIM}(\mathcal{H}) + \text{VCDIM}(\mathcal{G}))/\varepsilon^2)$. The proofs can be found in Appendix B.

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, and we exploit this in the development of our algorithmic results and their experimental evaluation.

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 $\neg g \in \mathcal{G}$.*

Recalling that $X = (x, x')$ and the following notions will be useful for describing our results:

- $\mathrm{SP}(D) = \Pr_{\mathcal{P},D}[D(X) = 1]$ and $\mathrm{FP}(D) = \Pr_{D,\mathcal{P}}[D(X) = 1 \mid y = 0]$.
- $\alpha_{SP}(g, \mathcal{P}) = \Pr_{\mathcal{P}}[g(x) = 1]$ and $\alpha_{FP}(g, \mathcal{P}) = \Pr_{\mathcal{P}}[g(x) = 1, y = 0]$.
- $\beta_{SP}(g, D, \mathcal{P}) = |\mathrm{SP}(D) - \mathrm{SP}(D, g)|$ and $\beta_{FP}(g, D, \mathcal{P}) = |\mathrm{FP}(D) - \mathrm{FP}(D, g)|$.
- 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 SP subgroup fairness. The reduction for FP subgroup fairness will follow as a corollary, since auditing for FP subgroup 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 $\mathcal{P}$, and any set of group indicators $\mathcal{G}$. Then for any $\gamma, \varepsilon > 0$, the following relationships hold:*

- *If there is a $(\gamma/2, (\gamma/2 - \varepsilon))$ auditing algorithm for $\mathcal{G}$ for all D such that $\mathrm{SP}(D) = 1/2$, then the class $\mathcal{G}$ is $(\gamma, \gamma/2 - \varepsilon)$-weakly agnostically learnable under P^D.*
- *If $\mathcal{G}$ is $(\gamma, \gamma - \varepsilon)$-weakly agnostically learnable under distribution P^D for all D such that $\mathrm{SP}(D) = 1/2$, then there is a $(\gamma, (\gamma - \varepsilon)/2)$ auditing algorithm for $\mathcal{G}$ for SP fairness under $\mathcal{P}$.*

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 $\mathrm{SP}(D) \leq 1/2$ and there exists a function f such that*

$$\alpha_{SP}(g, \mathcal{P})\, \beta_{SP}(g, D, \mathcal{P}) = \gamma.$$

Then

$$\max\{\Pr[D(X) = f(x)], \Pr[D(X) = \neg f(x)]\} \geq \mathrm{SP}(D) + \gamma.$$

Proof. To simplify notations, let $b = \mathrm{SP}(D)$ denote the base rate, $\alpha = \alpha_{SP}$ and $\beta = \beta_{SP}$. First, observe that either $\Pr[D(X) = 1 \mid f(x) = 1] = b + \beta$ or $\Pr[D(X) = 1 \mid f(x) = 1] = b - \beta$ holds.

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

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

which recovers our bound. □

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 unfairness certificate for D.

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

$$\alpha_{SP}(g, \mathcal{P})\, \beta_{SP}(g, D, \mathcal{P}) \geq \gamma/2,$$

where $g \in \{f, \neg f\}$.

Proof. Let $b = \mathrm{SP}(D)$. We can expand $\Pr[D(X) = f(x)]$ as follows:

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

This means

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

Suppose that $(\Pr[D(X) = 1 \mid f(x) = 1] - b)\Pr[f(x) = 1] \geq \gamma/2$, then our claim holds with $g = f$. Suppose not, then we must have

$$(\Pr[D(X) = 0 \mid f(x) = 0] - b)\Pr[f(x) = 0] = ((1-b) - \Pr[D(X) = 1 \mid f(x) = 0])\Pr[f(x) = 0] \geq \gamma/2$$

Note that by our assumption $b \geq (1-b)$. This means

$$(b - \Pr[D(X) = 1 \mid f(x) = 0])\Pr[f(x) = 0] \geq ((1-b) - \Pr[D(X) = 1 \mid f(x) = 0])\Pr[f(x) = 0] \geq \gamma/2$$

which implies that our claim holds with $g = \neg f$. □

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 $\Pr[g(x) = 1] |\Pr[D(X) = 1 \mid g(x) = 1] - \mathrm{SP}(D)| \geq \gamma/2$. By the assumption of auditability, we can then use the auditing algorithm to find a group $g' \in \mathcal{G}$ that is an $(\gamma/2 - \varepsilon)$-unfair certificate of D. By Lemma 3.3, we know that either g' or $\neg g'$ predicts D with an accuracy of at least $1/2 + (\gamma/2 - \varepsilon)$.

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 γ-unfair. Then by Lemma 3.3, we know that there exists some $g \in \mathcal{G}$ such that $\Pr[D(X) = g(x)] \geq 1/2 + \gamma$. Therefore, the weak agnostic learning algorithm from the hypothesis of the theorem will return some g' with $\Pr[D(X) = g'(x)] \geq 1/2 + (\gamma - \varepsilon)$. By Lemma 3.4, we know g' or $\neg g'$ is a $(\gamma - \varepsilon)/2$-unfair certificate for D. □

3.2 False Positive Fairness

A corollary of the above reduction is an analogous equivalence between auditing for FP subgroup fairness and agnostic learning. This is because a FP 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 3.5. *Fix any distribution $\mathcal{P}$, and any set of group indicators $\mathcal{G}$. The following two relationships hold:*

- *If there is a $(\gamma/2, (\gamma/2 - \varepsilon))$ auditing algorithm for $\mathcal{G}$ for all D such that $\mathrm{FP}(D) = 1/2$, then the class $\mathcal{G}$ is $(\gamma, \gamma/2 - \varepsilon)$-weakly agnostically learnable under $P^D_{y=0}$.*
- *If $\mathcal{G}$ is $(\gamma, \gamma - \varepsilon)$–weakly agnostically learnable under distribution $P^D_{y=0}$ for all D such that $\mathrm{FP}(D) = 1/2$, then there is a $(\gamma, (\gamma - \varepsilon)/2)$ auditing algorithm for FP subgroup fairness for $\mathcal{G}$ under distribution $\mathcal{P}$.*

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 3.6. *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 3.5, 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 3.6 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 present an algorithm for training a (randomized) classifier that satisfies false-positive subgroup 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 S denote a set of n labeled examples $\{z_i = (x_i, x_i'), y_i)\}_{i=1}^n$, and let $\mathcal{P}$ denote the empirical distribution over this set of examples. Let $\mathcal{H}$ be a hypothesis class defined 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).

Our goal will be to find the distribution over classifiers from $\mathcal{H}$ that minimizes classification error subject to the fairness constraint over $\mathcal{G}$. We will design an iterative algorithm that, when given access to a CSC oracle, computes an optimal randomized classifier in polynomial time.

Let D denote a probability distribution over $\mathcal{H}$. Consider the following *Fair ERM (Empirical Risk Minimization)* problem:

$$\min_{D \in \Delta_{\mathcal{H}}} \mathop{\mathbb{E}}_{h \sim D} [err(h, \mathcal{P})] \tag{2}$$

$$\text{such that } \forall g \in \mathcal{G} \qquad \alpha_{FP}(g, \mathcal{P})\, \beta_{FP}(g, D, \mathcal{P}) \leq \gamma. \tag{3}$$

where $err(h, \mathcal{P}) = \Pr_{\mathcal{P}}[h(x, x') \neq y]$, and the quantities α_{FP} and β_{FP} are defined in Definition 2.3. We will write OPT to denote the objective value at the optimum for the Fair ERM problem, that is the minimum error achieved by a γ-fair distribution over the class $\mathcal{H}$.

Observe that the optimization is feasible for any distribution $\mathcal{P}$: the constant classifiers that labels all points 1 or 0 satisfy all subgroup fairness constraints. At the moment, the number of decision variables and constraints may be infinite (if $\mathcal{H}$ and $\mathcal{G}$ are infinite hypothesis classes), but we will address this momentarily.

Assumption 4.1 (Cost-Sensitive Classification Oracle)**.** *We assume our algorithm has access to the cost-sensitive classication oracles* CSC($\mathcal{H}$) *and* CSC($\mathcal{G}$) *over the classes $\mathcal{H}$ and $\mathcal{G}$.*

Our main theoretical result is an computationally efficient oracle-based algorithm for solving the Fair ERM problem.

Theorem 4.2. *Fix any $\nu, \delta \in (0, 1)$. Then given an input of n data points and accuracy parameters ν, δ and access to oracles* CSC($\mathcal{H}$) *and* CSC($\mathcal{G}$)*, there exists an algorithm runs in polynomial time, and with probability at least $1 - \delta$, output a randomized classifier $\hat{D}$ such that $err(\hat{D}, \mathcal{P}) \leq \text{OPT} + \nu$, and for any $g \in \mathcal{G}$, the fairness constraint violations satisfies*

$$\alpha_{FP}(g, \mathcal{P})\, \beta_{FP}(g, \hat{D}, \mathcal{P}) \leq \gamma + O(\nu).$$

Overview of our solution. We present our solution in steps:

- **Step 1: Fair ERM as LP.** First, we rewrite the Fair ERM problem as a linear program with finitely many decision variables and constraints even when $\mathcal{H}$ and $\mathcal{G}$ are infinite. To do this, we take advantage of the fact that Sauer's Lemma lets us bound the number of labellings that any hypothesis class $\mathcal{H}$ of bounded VC dimension can induce on any fixed dataset. The LP has one variable for each of these possible labellings, rather than one variable for each hypothesis. Moreover, again by Sauer's Lemma, we have one constraint for each of the finitely many possible subgroups induced by $\mathcal{G}$ on the fixed dataset, rather than one for each of the (possibly infinitely many) subgroups definable over arbitrary datasets. This step is important — it will guarantee that strong duality holds.
- **Step 2: Formulation as Game.** We then derive the partial Lagrangian of the LP, and note that computing an approximately optimal solution to this LP is equivalent to finding an approximate minmax solution for a corresponding zero-sum game, in which the payoff function U is the value of the Lagrangian. The pure strategies of the primal or "Learner" player correspond to classifiers $h \in \mathcal{H}$, and the pure strategies of the dual or "Auditor" player correspond to subgroups $g \in \mathcal{G}$. Intuitively, the Learner is trying to minimize the sum of the prediction error and a fairness penalty term (given by the Lagrangian), and the Auditor is trying to penalize the fairness violation of the Learner by first identifying the subgroup with the greatest fairness violation and putting all the weight on the dual

variable corresponding to this subgroup. In order to reason about convergence, we restrict the set of dual variables to lie in a bounded set: C times the probability simplex. C is a parameter that we have to set in the proof of our theorem to give the best theoretical guarantees — but it is also a parameter that we will vary in the experimental section.

- **Step 3: Best Responses as CSC.** We observe that given a mixed strategy for the Auditor, the best response problem of the Learner corresponds to a CSC problem. Similarly, given a mixed strategy for the Learner, the best response problem of the Auditor corresponds to an auditing problem (which can be represented as a CSC problem). Hence, if we have oracles for solving CSC problems, we can compute best responses for both players, in response to arbitrary mixed strategies of their opponents.
- **Step 4: FTPL for No-Regret.** Finally, we show that the ability to compute best responses for each player is sufficient to implement dynamics known to converge quickly to equilibrium in zero-sum games. Our algorithm has the Learner play *Follow the Perturbed Leader (FTPL)* Kalai and Vempala [2005], which is a no-regret algorithm, against an Auditor who at every round best responds to the learner's mixed strategy. By the seminal result of Freund and Schapire [1996], the average plays of both players converge to an approximate equilibrium. In order to implement this in polynomial time, we need to represent the loss of the learner as a low-dimensional linear optimization problem. To do so, we first define an appropriately translated CSC problem for any mixed strategy λ by the Auditor, and cast it as a linear optimization problem.

4.1 Rewriting the Fair ERM Problem

To rewrite the Fair ERM problem, we note that even though both $\mathcal{G}$ and $\mathcal{H}$ can be infinite sets, the sets of possible labellings on the data set S induced by these classes are finite. More formally, we will write $\mathcal{G}(S)$ and $\mathcal{H}(S)$ to denote the set of all labellings on S that are induced by $\mathcal{G}$ and $\mathcal{H}$ respectively, that is

$$\mathcal{G}(S) = \{(g(x_1), \ldots, g(x_n)) \mid g \in \mathcal{G}\} \qquad \text{and,} \qquad \mathcal{H}(S) = \{(h(X_1), \ldots, h(X_n)) \mid h \in \mathcal{H}\}$$

We can bound the cardinalities of $\mathcal{G}(S)$ and $\mathcal{H}(S)$ using Sauer's Lemma.

Lemma 4.3 (Sauer's Lemma (see e.g. Kearns and Vazirani [1994])). *Let S be a data set of size n. Let $d_1 = \text{VCDIM}(\mathcal{H})$ and $d_2 = \text{VCDIM}(\mathcal{G})$ be the VC-dimensions of the two classes. Then*

$$|\mathcal{H}(S)| \leq O\left(n^{d_1}\right) \qquad \text{and} \qquad |\mathcal{G}(S)| \leq O\left(n^{d_2}\right).$$

Given this observation, we can then consider an equivalent optimization problem where the distribution D is over the set of labellings in $\mathcal{H}(S)$, and the set of subgroups are defined by the labellings in $\mathcal{G}(S)$. We will view each g in $\mathcal{G}(S)$ as a Boolean function.

To simplify notations, we will define the following "fairness violation" functions for any $g \in \mathcal{G}$ and any $h \in \mathcal{H}$:

$$\Phi_+(h,g) \equiv \alpha_{FP}(g,P)\,(\text{FP}(h) - \text{FP}(h,g)) - \gamma \tag{4}$$
$$\Phi_-(h,g) \equiv \alpha_{FP}(g,\mathcal{P})\,(\text{FP}(h,g) - \text{FP}(h)) - \gamma \tag{5}$$

Moreover, for any distribution D over $\mathcal{H}$, for any sign $\bullet \in \{+,-\}$

$$\Phi_\bullet(D,g) = \mathop{\mathbb{E}}_{h\sim D}[\Phi_\bullet(h,g)].$$

Claim 4.4. *For any $g \in \mathcal{G}$, $h \in \mathcal{H}$, and any $\nu > 0$,*

$$\max\{\Phi_+(D,g), \Phi_-(D,g)\} \leq \nu \quad \text{if and only if} \quad \alpha_{FP}(g,\mathcal{P})\,\beta_{FP}(g,D,\mathcal{P}) \leq \gamma + \nu.$$

Thus, we will focus on the following equivalent optimization problem.

$$\min_{D \in \Delta_{\mathcal{H}(S)}} \mathop{\mathbb{E}}_{h \sim D} [err(h, \mathcal{P})] \tag{6}$$

$$\text{such that for each } g \in \mathcal{G}(S): \qquad \Phi_+(D, g) \leq 0 \tag{7}$$

$$\Phi_-(D, g) \leq 0 \tag{8}$$

For each pair of constraints (7) and (8), corresponding to a group $g \in \mathcal{G}(S)$, we introduce a pair of dual variables λ_g^+ and λ_g^-. The partial Lagrangian of the linear program is the following:

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

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

$$\min_{D \in \Delta_{\mathcal{H}(S)}} \max_{\lambda \in \mathbb{R}_+^{2|\mathcal{G}(S)|}} \mathcal{L}(p, \lambda) = \max_{\lambda \in \mathbb{R}_+^{2|\mathcal{G}(S)|}} \min_{D \in \Delta_{\mathcal{H}(S)}} \mathcal{L}(p, \lambda) = \text{OPT}$$

where OPT denotes the optimal objective value in the fair ERM problem. Similarly, the distribution $\arg\min_D \max_\lambda \mathcal{L}(D, \lambda)$ 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 algorithms will both compute such a minmax solution by iteratively optimizing over both the primal variables D and dual variables λ. In order to guarantee convergence in our optimization, we will restrict the dual space to the following bounded set:

$$\Lambda = \{\lambda \in \mathbb{R}_+^{2|\mathcal{G}(S)|} \mid \|\lambda\|_1 \leq C\}.$$

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]:

$$\min_{D \in \Delta_{\mathcal{H}(S)}} \max_{\lambda \in \Lambda} \mathcal{L}(D, \lambda) = \max_{\lambda \in \Lambda} \min_{D \in \Delta_{\mathcal{H}(S)}} \mathcal{L}(D, \lambda) \tag{9}$$

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 can show that even though we restrict the dual variables to lie in a bounded set, any approximate minmax solution to Equation (9) is also an approximately optimal and approximately feasible solution to the original fair ERM problem.

Theorem 4.5. *Let $(\hat{D}, \hat{\lambda})$ be a ν-approximate minmax solution to the Λ-bounded Lagrangian problem in the sense that*

$$\mathcal{L}(\hat{D}, \hat{\lambda}) \leq \min_{D \in \Delta_{\mathcal{H}(S)}} \mathcal{L}(D, \hat{\lambda}) + \nu \quad \textit{and,} \quad \mathcal{L}(\hat{D}, \hat{\lambda}) \geq \max_{\lambda \in \Lambda} \mathcal{L}(\hat{D}, \lambda) - \nu.$$

Then $err(\hat{D}, \mathcal{P}) \leq \text{OPT} + 2\nu$ and for any $g \in \mathcal{G}(S)$,

$$\alpha_{FP}(g, \mathcal{P})\, \beta_{FP}(g, \hat{D}, \mathcal{P}) \leq \gamma + \frac{1 + 2\nu}{C}.$$

4.2 Zero-Sum Game Formulation

To compute an approximate minmax solution, we will first view Equation (9) 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 either is the all zero vector or consists of a choice of a $g \in \mathcal{G}(S)$, along with the sign + or − that the corresponding g-fairness constraint will have in the Lagrangian. More formally, we write

$$\Lambda_{\text{pure}} = \{\lambda \in \Lambda \text{ with } \lambda_g^\bullet = C \mid g \in \mathcal{G}(S), \bullet \in \{\pm\}\} \cup \{\mathbf{0}\}$$

Even though the number of pure strategies scales linearly with $|\mathcal{G}(S)|$, 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 the 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

$$U(h, \lambda) = err(h, \mathcal{P}) + \sum_{g \in \mathcal{G}(S)} \left(\lambda_g^+ \Phi_+(h, g) + \lambda_g^- \Phi_-(h, g)\right).$$

Claim 4.6. *Let $D \in \Delta_{\mathcal{H}(S)}$ and $\lambda \in \Lambda$ such that (p, λ) is a ν-approximate minmax equilibrium in the zero-sum game defined above. Then (p, λ) is also a ν-approximate minmax solution for Equation (9).*

Our problem reduces to finding an approximate equilibrium for this game. A key step in our solution is the ability to compute best responses for both players in the game, which we now show can be solved by the cost-sensitive classication (CSC) oracles.

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

$$\operatorname*{argmin}_{D \in \Delta_{\mathcal{H}(S)}} err(h, \mathcal{P}) + \sum_{g \in \mathcal{G}(S)} \left(\lambda_g^+ \Phi_+(D, g) + \lambda_g^- \Phi_-(D, g)\right) \tag{10}$$

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 CSC oracle. In particular, we can assign costs to each example (X_i, y_i) as follows:

- if $y_i = 1$, then $c_i^0 = 0$ and $c_i^1 = -\frac{1}{n}$;
- otherwise, $c_i^0 = 0$ and

$$c_i^1 = \frac{1}{n} + \frac{1}{n} \sum_{g \in \mathcal{G}(S)} (\lambda_g^+ - \lambda_g^-)(\Pr[g(x) = 1 \mid y = 0] - 1)\, \mathbf{1}[g(x_i) = 1] \tag{11}$$

Given a fixed set of dual variables λ, we will write $\text{LC}(\lambda) \in \mathbb{R}^n$ to denote the vector of costs for labelling each datapoint as 1. That is, $\text{LC}(\lambda)$ is the vector such that for any $i \in [n]$, $\text{LC}(\lambda)_i = c_i^1$

Remark 4.7. *Note that in defining the costs above, we have translated them from their most natural values so that the cost of labeling any example with 0 is 0. In doing so, we recall that by Claim 2.7, the solution to a cost-sensitive classification problem is invariant to translation. As we will see, this will allow us to formulate the learner's optimization problem as a low-dimensional linear optimization problem, which will be important for an efficient implementation of follow the perturbed leader. In particular, if we find a hypothesis that produces the n labels $y = (y_1, \ldots, y_n)$ for the n points in our dataset, then the cost of this labelling in the CSC problem is by construction $\langle \text{LC}(\lambda), y \rangle$.*

Auditor's best response as CSC. Fix any mixed strategy (primal solution) $p \in \Delta_{\mathcal{H}(S)}$ of the Learner. The Auditor's best response is given by:

$$\operatorname*{argmax}_{\lambda \in \Lambda} err(D,\mathcal{P}) + \sum_{g \in \mathcal{G}(S)} \left(\lambda_g^+ \Phi_+(D,g) + \lambda_g^- \Phi_-(D,g)\right) = \operatorname*{argmax}_{\lambda \in \Lambda} \sum_{g \in \mathcal{G}(S)} \left(\lambda_g^+ \Phi_+(D,g) + \lambda_g^- \Phi_-(D,g)\right) \tag{12}$$

To find the best response, consider the problem of computing $(\hat{g}, \hat{\bullet}) = \operatorname{argmax}_{(g,\bullet)} \Phi_\bullet(D,g)$. There are two cases. In the first case, p is a strictly feasible primal solution: that is $\Phi_{\hat{\bullet}}(D,\hat{g}) < 0$. In this case, the solution to (12) sets $\lambda = \mathbf{0}$. Otherwise, if p is not strictly feasible, then by the following Lemma 4.8 the best response is to set $\lambda^{\hat{\bullet}}_{\hat{g}} = C$ (and all other coordinates to 0).

Lemma 4.8. *Fix any $\overline{D} \in \Delta_{\mathcal{H}(S)}$ such that that $\max_{g \in \mathcal{G}(S)}\{\Phi_+(\overline{D},g), \Phi_-(\overline{D},g)\} > 0$. Let $\lambda' \in \Lambda$ be vector with one non-zero coordinate $(\lambda')^{\bullet'}_{g'} = C$, where*

$$(g', \bullet') = \operatorname*{argmax}_{(g,\bullet) \in \mathcal{G}(S) \times \{\pm\}} \{\Phi_\bullet(\overline{D}, g)\}$$

Then $\mathcal{L}(\overline{D}, \lambda') \geq \max_{\lambda \in \Lambda} \mathcal{L}(\overline{D}, \lambda)$.

Therefore, it suffices to solve for $\operatorname{argmax}_{(g,\bullet)} \Phi_\bullet(D,g)$. We proceed by solving $\operatorname{argmax}_g \Phi_+(D,g)$ and $\operatorname{argmax}_g \Phi_-(D,g)$ separately: both problems can be reduced to a cost-sensitive classification problem. To solve for $\operatorname{argmax}_g \Phi_+(D,g)$ with a CSC oracle, we assign costs to each example (X_i, y_i) as follows:

- if $y_i = 1$, then $c_i^0 = 0$ and $c_i^1 = 0$;
- otherwise, $c_i^0 = 0$ and

$$c_i^1 = \frac{-1}{n}\left[\mathop{\mathbb{E}}_{h \sim D}[\mathrm{FP}(h)] - \mathop{\mathbb{E}}_{h \sim D}[h(X_i)]\right] \tag{13}$$

To solve for $\operatorname{argmax}_g \Phi_-(D,g)$ with a CSC oracle, we assign the same costs to each example (X_i, y_i), except when $y_i = 0$, labeling "1" incurs a cost of

$$c_i^1 = \frac{-1}{n}\left[\mathop{\mathbb{E}}_{h \sim D}[h(X_i)] - \mathop{\mathbb{E}}_{h \sim D}[\mathrm{FP}(h)]\right]$$

4.3 Solving the Game with No-Regret Dynamics

To compute an approximate equilibrium of the zero-sum game, we will simulate the following *no-regret dynamics* between the Learner and the Auditor over rounds: over each of the T rounds, the Learner plays a distribution over the hypothesis class according to a *no-regret* learning algorithm (Follow the Perturbed Leader), and the Auditor plays an approximate best response against the Learner's distribution for that round. By the result of Freund and Schapire [1996], the average plays of both players over time converge to an approximate equilibrium of the game, as long as the Learner has low regret.

Theorem 4.9 (Freund and Schapire [1996])**.** *Let $D^1, D^2, \ldots, D^T \in \Delta_{\mathcal{H}(S)}$ be a sequence of distributions played by the Learner, and let $\lambda^1, \lambda^2, \ldots, \lambda^T \in \Lambda_{pure}$ be the Auditor's sequence of approximate best responses against these distributions respectively. Let $\overline{D} = \frac{1}{T}\sum_{t=1}^T D^t$ and $\overline{\lambda} = \frac{1}{T}\sum_{t=1}^T \lambda^t$ be the two players' empirical distributions over their strategies. Suppose that the regret of the Learner satisfies*

$$\sum_{t=1}^T \mathop{\mathbb{E}}_{h \sim D^t}\left[U(h, \lambda^t)\right] - \min_{h \in \mathcal{H}(S)} \sum_{t=1}^T U(h, \lambda^t) \leq \gamma_L T \quad \textit{and} \quad \max_{\lambda \in \Lambda} \sum_{t=1}^T \mathop{\mathbb{E}}_{h \sim D^t}[U(h,\lambda)] - \sum_{t=1}^T \mathop{\mathbb{E}}_{h \sim D^t}\left[U(h, \lambda^t)\right] \leq \gamma_A T.$$

Then $(\overline{D}, \overline{\lambda})$ is an $(\gamma_L + \gamma_A)$-approximate minimax equilibrium of the game.

Our Learner will play using the Follow the Perturbed Leader (FTPL), which gives a no-regret guarantee. In order to implement FPTL, we will first need to formulate the Learner's best response problem as a linear optimization problem over a low dimensional space. For each round t, let $\overline{\lambda}^t = \sum_{s<t} \lambda^s$ be the vector representing the sum of the actions played by the auditor over previous rounds, and recall that $\mathrm{LC}(\overline{\lambda}^t)$ is the cost vector given by our cost-sensitive classification reduction. Then the Learner's best response problem against $\overline{\lambda}^t$ is the following linear optimization problem

$$\min_{h \in \mathcal{H}(S)} \langle \mathrm{LC}(\overline{\lambda}^t), h \rangle.$$

To run the FTPL algorithm, the Learner will optimize a "perturbed" version of the problem above. In particular, the Learner will play a distribution D^t over the hypothesis class $\mathcal{H}(S)$ that is implicitely defined by the following sampling operation. To sample a hypothesis h from D^t, the learner solves the following randomized optimization problem:

$$\min_{h \in \mathcal{H}(S)} \langle \mathrm{LC}(\overline{\lambda}^t), h \rangle + \frac{1}{\eta} \langle \xi, h \rangle, \tag{14}$$

where η is a parameter and ξ is a noise vector drawn from the uniform distribution over $[0,1]^n$. Note that while it is intractable to explicitly represent the distribution D^t (which has support size scaling with $|\mathcal{H}(S)|$), we can sample from D^t efficiently given access to a cost-sensitive classification oracle for $\mathcal{H}$. By instantiating the standard regret bound of FTPL for online linear optimization (Theorem 2.9), we get the following regret bound for the Learner.

Lemma 4.10. *Let T be the time horizon for the no-regret dynamics. Let $D^1, \ldots, D^T$ be the sequence of distributions maintained by the Learner's FTPL algorithm with $\eta = \frac{n}{(1+C)} \sqrt{\frac{1}{\sqrt{n}T}}$, and $\lambda^1, \ldots, \lambda^T$ be the sequence of plays by the Auditor. Then*

$$\sum_{t=1}^{T} \mathop{\mathbb{E}}_{h \sim D^t} \left[U(h, \lambda^t) \right] - \min_{h \in \mathcal{H}(S)} \sum_{t=1}^{T} U(h, \lambda^t) \leq 2n^{1/4}(1+C)\sqrt{T}$$

Now we consider how the Auditor (approximately) best responds to the distribution D^t. The main obstacle is that we do not have an explicit representation for D^t. Thus, our first step is to approximate D^t with an explicitly represented sparse distribution $\hat{D}^t$. We do that by drawing m i.i.d. samples from D^t, and taking the empirical distribution $\hat{D}^t$ over the sample. The Auditor will best respond to this empirical distribution $\hat{D}^t$. To show that any best response to $\hat{D}^t$ is also an approximate best response to D^t, we will rely on the following uniform convergence lemma, which bounds the difference in expected payoff for any strategy of the auditor, when played against D^t as compared to $\hat{D}^t$.

Lemma 4.11. *Fix any $\xi, \delta \in (0,1)$ and any distribution D over $\mathcal{H}(S)$. Let $h^1, \ldots, h^m$ be m i.i.d. draws from p, and $\hat{D}$ be the empirical distribution over the realized sample. Then with probability at least $1 - \delta$ over the random draws of h^j's, the following holds,*

$$\max_{\lambda \in \Lambda} \left| \mathop{\mathbb{E}}_{h \sim \hat{D}} [U(h, \lambda)] - \mathop{\mathbb{E}}_{h \sim D} [U(h, \lambda)] \right| \leq \xi,$$

as long as $m \geq c_0 \frac{C^2(\ln(1/\delta) + d_2 \ln(n))}{\xi^2}$ for some absolute constant c_0 and $d_2 = \mathrm{VCDIM}(\mathcal{G})$.

Using Lemma 4.11, we can derive a regret bound for the Auditor in the no-regret dynamics.

Lemma 4.12. *Let T be the time horizon for the no-regret dynamics. Let $D^1, \ldots, D^T$ be the sequence of distributions maintained by the Learner's FTPL algorithm. For each D^t, let $\hat{D}^t$ be the empirical distribution over m i.i.d. draws from D^t. Let $\lambda^1, \ldots, \lambda^T$ be the Auditor's best responses against $\hat{D}^1, \ldots, \hat{D}^T$. Then with probability $1-\delta$,*

$$\max_{\lambda \in \Lambda} \sum_{t=1}^{T} \mathop{\mathbb{E}}_{h \sim D^t}[U(h,\lambda)] - \sum_{t=1}^{T} \mathop{\mathbb{E}}_{h \sim D^t}\left[U(h,\lambda^t)\right] \leq T\sqrt{\frac{c_0 C^2(\ln(T/\delta) + d_2 \ln(n))}{m}}$$

for some absolute constant c_0 and $d_2 = \mathrm{VCDIM}(\mathcal{G})$.

Finally, let $\overline{D}$ and $\overline{\lambda}$ be the average of the strategies played by the two players over the course of the dynamics. Note that $\overline{D}$ is an average of many *distributions* with large support, and so $\overline{D}$ itself has support size that is too large to represent explicitely. Thus, we will again approximate $\overline{D}$ with a sparse distribution $\hat{D}$ estimated from a sample drawn from $\overline{D}$. Note that we can efficiently *sample* from $\overline{D}$ given access to a CSC oracle. To sample, we first uniformly randomly select a round $t \in [T]$, and then use the CSC oracle to solve the sampling problem defined in (14), with the noise random variable ξ freshly sampled from its distribution. The full algorithm is described in Algorithm 2 and we present the proof for Theorem 4.2 below.

Algorithm 2 FairNR: Fair No-Regret Dynamics

Input: distribution $\mathcal{P}$ over n labelled data points, CSC oracles $\mathrm{CSC}(\mathcal{H})$ and $\mathrm{CSC}(\mathcal{G})$, dual bound C, and target accuracy parameter ν, δ

Initialize: Let $C = 1/\nu$, $\overline{\lambda}^0 = \mathbf{0}$, $\eta = \frac{n}{(1+C)}\sqrt{\frac{1}{\sqrt{n}T}}$,

$$m = \frac{(\ln(2T/\delta) d_2 \ln(n)) C^2 c_0 T}{\sqrt{n}(1+C)^2 \ln(2/\delta)} \quad \text{and,} \quad T = \frac{4\sqrt{n}\ln(2/\delta)}{\nu^4}$$

For $t = 1, \ldots, T$:
  Sample from the Learner's FTPL distribution:
    For $s = 1, \ldots m$:
      Draw a random vector ξ^s uniformly at random from $[0,1]^n$
      Use the oracle $\mathrm{CSC}(\mathcal{H})$ to compute $h^{(s,t)} = \mathrm{argmin}_{h \in \mathcal{H}(S)} \langle \mathrm{LC}(\overline{\lambda}^{(t-1)}), h\rangle + \frac{1}{\eta}\langle \xi^s, h\rangle$
    Let $\hat{D}^t$ be the empirical distribution over $\{h^{s,t}\}$

  Auditor best responds to $\hat{D}^t$:
    Use the oracle $\mathrm{CSC}(\mathcal{G})$ to compute $\lambda^t = \mathrm{argmax}_\lambda \mathbb{E}_{h \sim \hat{D}}[U(h,\lambda)]$
  Update: Let $\overline{\lambda}^t = \sum_{t' \leq t} \lambda^{t'}$
Sample from the average distribution $\overline{D} = \sum_{t=1}^{T} D^t$:
  For $s = 1, \ldots m$:
    Draw a random number $r \in [T]$ and a random vector ξ^s uniformly at random from $[0,1]^n$
    Use the oracle $\mathrm{CSC}(\mathcal{H})$ to compute $h^{(r,t)} = \mathrm{argmin}_{h \in \mathcal{H}(S)} \langle \mathrm{LC}(\overline{\lambda}^{(r-1)}), h\rangle + \frac{1}{\eta}\langle \xi^s, h\rangle$
  Let $\hat{D}$ be the empirical distribution over $\{h^{r,t}\}$
Output: $\hat{D}$ as a randomized classifier

Proof of Theorem 4.2. By Theorem 4.5, it suffices to show that with probability at least $1-\delta$, $(\hat{D}, \overline{\lambda})$ is a ν-approximate equilibrium in the zero-sum game. As a first step, we will rely on Theorem 4.9 to show that $(\overline{D}, \overline{\lambda})$ forms an approximate equilibrium.

By Lemma 4.10, the regret of the sequence $D^1, \ldots, D^T$ is bounded by:

$$\gamma_L = \frac{1}{T}\left[\sum_{t=1}^{T} \mathop{\mathbb{E}}_{h\sim D^t}\left[U(h,\lambda^t)\right] - \min_{h\in\mathcal{H}(S)} \sum_{t=1}^{T} U(h,\lambda^t)\right] \leq \frac{2n^{1/4}(1+C)}{\sqrt{T}}$$

By Lemma 4.12, with probability $1-\delta/2$, we have

$$\gamma_A \leq \sqrt{\frac{c_0 C^2(\ln(2T/\delta) + d_2\ln(n))}{m}}$$

We will condition on this upper-bound event on γ_A for the rest of this proof, which is the case except with probability $\delta/2$. By Theorem 4.9, we know that the average plays $(\overline{D}, \overline{\lambda})$ form an $(\gamma_L + \gamma_A)$-approximate equilibrium.

Finally, we need to bound the additional error for outputting the sparse approximation $\hat{D}$ instead of $\overline{D}$. We can directly apply Lemma 4.11, which implies that except with probability $\delta/2$, the pair $(\hat{D}, \overline{\lambda})$ form a R-approximate equilibrium, with

$$R \leq \gamma_A + \gamma_L + \frac{\sqrt{c_0 C^2(\ln(2/\delta) + d_2\ln(n))}}{\sqrt{m}}$$

Note that $R \leq \nu$ as long as we have $C = 1/\nu$,

$$m = \frac{(\ln(2T/\delta)d_2\ln(n))\,C^2 c_0 T}{\sqrt{n}(1+C)^2\ln(2/\delta)} \quad \text{and,} \quad T = \frac{4\sqrt{n}\ln(2/\delta)}{\nu^4}$$

This completes our proof. □

5 Experimental Evaluation

We now describe an experimental evaluation of our proposed algorithmic framework on a dataset in which fairness is a concern, due to the preponderance of racial and other sensitive features. We also demonstrate that for this dataset, our methods are empirically necessary to avoid fairness gerrymandering.

While the no-regret-based algorithm described in the last section enjoys provably polynomial time convergence, for the experiments we instead implemented a simpler yet effective algorithm based on *Fictitious Play* dynamics. We first describe and discuss this modified algorithm.

5.1 Solving the Game with Fictitious Play

Like the algorithm given in the last section, the algorithm we implemented works by simulating a game dynamic that converges to Nash equilibrium in the zero-sum game that we derived, corresponding to the Fair ERM problem. Rather than using a no-regret dynamic, we instead use a simple iterative procedure known as *Fictitious Play* [Brown, 1949]. Fictitious Play dynamics has the benefit of being more practical to implement: at each round, both players simply need to compute a single best response to the empirical play of their opponents, and this optimization requires only a single call to a CSC oracle. In contrast, the FTPL dynamic we gave in the previous section requires making many calls to a CSC oracle per round — a

computationally expensive process — in order to find a sparse approximation to the Learner's mixed strategy at that round. Fictitious Play also has the benefit of being deterministic, unlike the randomized sampling required in the FTPL no-regret dynamic, thus eliminating a source of experimental variance.

The disadvantage is that Fictitious Play is only known to converge to equilibrium in the limit Robinson [1951], rather than in a polynomial number of rounds (though it is conjectured to converge quickly under rather general circumstances; see Daskalakis and Pan [2014] for a recent discussion). Nevertheless, this is the algorithm that we use in our experiments — and as we will show, it performs well on real data, despite the fact that it has weaker theoretical guarantees compared to the algorithm we presented in the last section.

Fictitious play 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. Pseudocode for the implemented algorithm is given below.

Algorithm 3 FairFictPlay: Fair Fictitious Play

Input: distribution $\mathcal{P}$ over the labelled data points, CSC oracles CSC($\mathcal{H}$) and CSC($\mathcal{G}$) for the classes $\mathcal{H}(S)$ and $\mathcal{G}(S)$ respectively, 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 $\overline{D}$ and $\overline{\lambda}$ be the point distributions that put all their mass on h^0 and λ^0 respectively.

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

Compute the empirical play distributions:

Let $\overline{D}$ 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

Learner best responds: Use the oracle CSC($\mathcal{H}$) to compute $h^t = \text{argmin}_{h \in \mathcal{H}(S)} \langle \text{LC}(\overline{\lambda}), h \rangle$

Auditor best responds: Use the oracle CSC($\mathcal{G}$) to compute $\lambda^t = \text{argmax}_{\lambda} \mathbb{E}_{h \sim \overline{D}}[U(h, \lambda)]$

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

5.2 Description of Data

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[4]. 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

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

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. [5]

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.

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 classifier learned by the algorithm has nontrivial accuracy, as well as strong subgroup fairness properties.
- Whether the algorithm and dataset 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 2.12 and 2.11), we report all results here on the full C&C dataset of 1994 points, treating it as the true distribution of interest.

5.3 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

[5] Ongoing experiments on other datasets where fairness is a concern will be reported on in a forthcoming experimental paper.

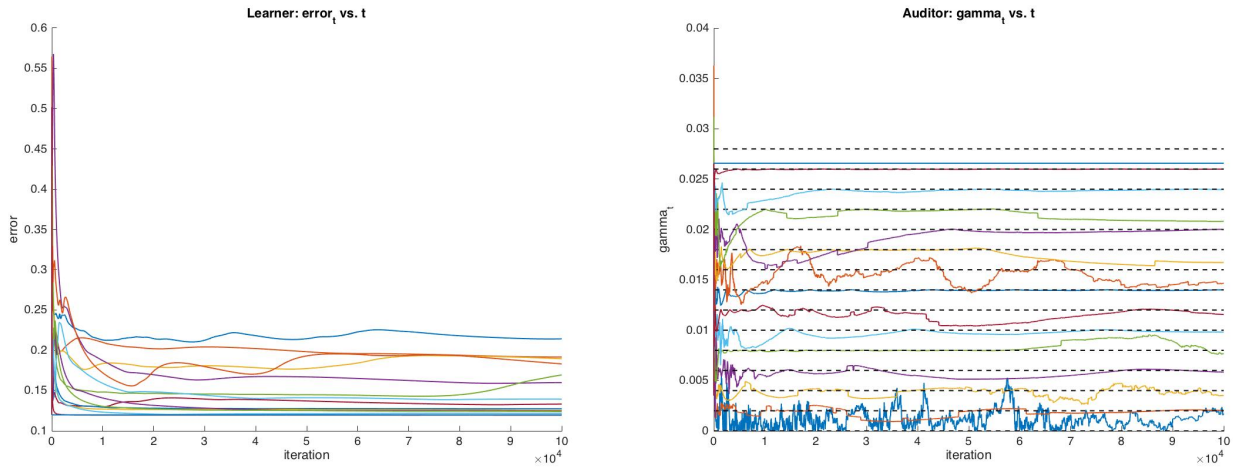
Learner does, except that Auditor only has access to the 18 sensitive features, rather than all 122.

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). While the theory does not provide an explicit bound or guide for choosing C, it needs to be large enough to permit the dual player to force the min-max value of the game. For our experiments we chose $C = 10$, which despite being a relatively small value seems to suffice for (approximate) convergence.

The other and more meaningful parameter of the algorithm is the bound γ in the Fair ERM optimization problem implemented by the game, which controls the amount of unfairness permitted. If on a given round the subgroup disparity found by the Auditor is greater than γ, the Learner must react by adding a fairness penalty for this subgroup to its objective function; if it is smaller than γ, the Learner can ignore it and continue to optimize its previous objective function. Ideally, and as we shall see, varying γ allows us to trace out a menu of trade-offs between accuracy and fairness.

5.4 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.



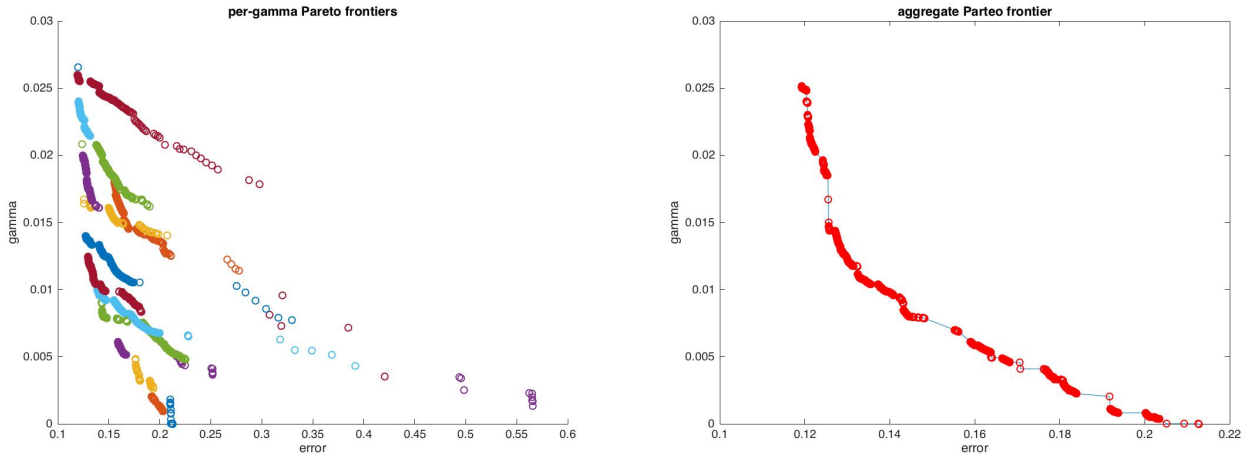
(a) (b)

Figure 1: Evolution of the error and unfairness of Learner's classifier across iterations, for varying choices of γ. (a) Error ε_t of Learner's model vs iteration t. (b) Unfairness γ_t of subgroup found by Auditor vs. iteration t, as measured by Definition 2.3. See text for details.

We begin by examining the evolution of the error and unfairness of Learner's model. In the left panel of Figure 1 we show the error of the model found by Learner vs. iteration for values of γ ranging from 0 to 0.029. Several comments are in order.

First, after an intial period in which there is a fair amount of oscillatory behavior, by 50,000 iterations most of the curves have largely flattened out, and by 100,000 iterations it appears most but not all have reached approximate convergence. Second, while the top-to-bottom ordering of these error curves is approximately aligned with decreasing γ — so larger γ generally results in lower error, as expected — there are many violations of this for small t, and even a few at large t. Third, and as we will examine more closely shortly, the converged values at large t do indeed exhibit a range of errors.

In the right panel of Figure 1, we show the corresponding unfairness γ_t of the subgroup found by the Auditor at each iteration t for the same runs and values of the parameter γ (indicated by horizontal dashed lines), with the same color-coding as for the left panel. Now the ordering is generally reversed — larger values of γ generally lead to higher γ_t curves, since the fairness constraint on the Learner is weaker. We again see a great deal of early oscillatory behavior, with most γ_t curves then eventually settling at or near their corresponding input γ value, as Learner and Auditor engage in a back-and-forth struggle for lower error for Learner and γ-subgroup fairness for Auditor.



(a) (b)

Figure 2: (a) Pareto-optimal error-unfairness values, color coded by varying values of the input parameter γ. (b) Aggregate Pareto frontier across all values of γ. Here the γ values cover the same range but are sampled more densely to get a smoother frontier. See text for details.

For any choice of the parameter γ, and each iteration t, the two panels of Figure 1 yield a pair of realized values $\langle \varepsilon_t, \gamma_t \rangle$ from the experiment, corresponding to a Learner model whose error is ε_t, and for which the worst subgroup the Auditor was able to find had unfairness γ_t. The set of all $\langle \varepsilon_t, \gamma_t \rangle$ pairs across all runs or γ values thus represents the different trade-offs between error and unfairness found by our algorithm on the data. Most of these pairs are of course Pareto-dominated by other pairs, so we are primarily interested in the undominated frontier.

In the left panel of Figure 2, for each value of γ we show the Pareto-optimal pairs, color-coded for the value of γ. Each value of γ yields a set or cloud of undominated pairs that are usually fairly close to each other, and as expected, as γ is increased, these clouds generally move leftwards and upwards (lower error and higher unfairness).

We anticipate that the practical use of our algorithm would, as we have done, explore many values of γ and then pick a model corresponding to a point on the aggregated Pareto frontier across all γ, which represents the collection of all undominated models and the overall error-unfairness trade-off. This aggregate frontier is shown in the right panel of Figure 2, and shows a relatively smooth menu of options, ranging from error about 0.21 and no unfairness at one extreme, to error about 0.12 and unfairness 0.025 at the other, and an appealing assortment of intermediate trade-offs. Of course, in a real application the selection of a particular point on the frontier should be made in a domain-specific manner by the stakeholders or policymakers in question.

5.5 Protecting Marginal Subgroups is not Sufficient

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 only 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. In this case our more elaborate techniques and guarantees would not be needed except for in theory.

To explore 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 discover 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, and the subgroup can thus 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. We

suspect this phenomenon is common on many datasets, and that our methods and algorithms are needed to address it.

Acknowledgements We thank Alekh Agarwal, Richard Berk, Miro Dudík, Akshay Krishnamurthy, 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.

Yoav Freund and Robert E. Schapire. Game theory, on-line prediction and boosting. In *Proceedings of the Ninth Annual Conference on Computational Learning Theory, COLT 1996, Desenzano del Garda, Italy, June 28-July 1, 1996.*, pages 325–332, 1996.

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.

Úrsula Hébert-Johnson, Michael P Kim, Omer Reingold, and Guy N Rothblum. Calibration for the (computationally-identifiable) masses. *arXiv preprint arXiv:1711.08513*, 2017.

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 and Santosh Vempala. Efficient algorithms for online decision problems. *J. Comput. Syst. Sci.*, 71(3):291–307, 2005.

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.

Bianca Zadrozny, John Langford, and Naoki Abe. Cost-sensitive learning by cost-proportionate example weighting. In *Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM 2003), 19-22 December 2003, Melbourne, Florida, USA*, page 435, 2003.

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 Chernoff-Hoeffding Bound

We use the following concentration inequality.

Theorem A.1 (Real-vaued Additive Chernoff-Hoeffding Bound)**.** *Let $X_1, X_2, \ldots, X_m$ be i.i.d. random variables with $\mathbb{E}[X_i] = \mu$ and $a \le X_i \le b$ for all i. Then for every $\alpha > 0$,*

$$\Pr\left[\left|\frac{\sum_i X_i}{m} - \mu\right| \ge \alpha\right] \le 2\exp\left(\frac{-2\alpha^2 m}{(b-a)^2}\right)$$

B Generalization Bounds

Proof of Theorems 2.11 and 2.12. We give a proof of Theorem 2.11. The proof of Theorem 2.12 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 B.1.

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

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 $\pi_{\mathcal{F}_{\mathcal{H},\mathcal{G}}}(S) \le \pi_{\mathcal{F}}(S)\pi_{\mathcal{G}}(S)$. Now by the Sauer-Shelah Lemma (see e.g. Kearns and Vazirani [1994]), $\pi_{\mathcal{F}}(S) = O(m^{\text{VCDIM}(\mathcal{H})}), \pi_{\mathcal{G}}(S) = O(m^{\text{VCDIM}(\mathcal{G})})$. Thus $\pi_{\mathcal{F}_{\mathcal{H},\mathcal{G}}}(S) = 2^m \le O(m^{\text{VCDIM}(\mathcal{H})+\text{VCDIM}(\mathcal{G})})$, which implies that $m = \tilde{O}(\text{VCDIM}(\mathcal{H}) + \text{VCDIM}(\mathcal{G}))$, 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 $f_{h,g} \in \mathcal{F}_{\mathcal{H},\mathcal{G}}$:

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

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

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

This completes our proof. □

C Missing Proofs in Section 4

Theorem 4.5. *Let $(\hat{D}, \hat{\lambda})$ be a ν-approximate minmax solution to the Λ-bounded Lagrangian problem in the sense that*

$$\mathcal{L}(\hat{D}, \hat{\lambda}) \le \min_{D \in \Delta_{\mathcal{H}(S)}} \mathcal{L}(D, \hat{\lambda}) + \nu \quad \textit{and,} \quad \mathcal{L}(\hat{D}, \hat{\lambda}) \ge \max_{\lambda \in \Lambda} \mathcal{L}(\hat{D}, \lambda) - \nu.$$

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

$$\alpha_{FP}(g,\mathcal{P})\,\beta_{FP}(g,\hat{D},\mathcal{P}) \le \gamma + \frac{1+2\nu}{C}.$$

Proof of Theorem 4.5. Let D^* be the optimal feasible solution for our constrained optimization problem. Since D^* is feasible, we know that $\mathcal{L}(D^*,\hat{\lambda}) \le err(D^*,\mathcal{P})$.

We will first focus on the case where $\hat{D}$ is not a feasible solution, that is

$$\max_{(g,\bullet)\in\mathcal{G}(S)\times\{\pm\}} \Phi_\bullet(\hat{D},g) > 0$$

Let $(\hat{g},\hat{\bullet}) \in \operatorname{argmax}_{(g,\bullet)} \Phi_\bullet(\hat{D},g)$ and let $\lambda' \in \Lambda$ be a vector with $(\lambda')^{\hat{\bullet}}_{\hat{g}} = C$ and all other coordinates zero. By Lemma 4.8, we know that $\lambda' \in \operatorname{argmax}_{\lambda\in\Lambda} \mathcal{L}(\hat{D},\lambda)$. By the definition of a ν-approximate minmax solution, we know that $\mathcal{L}(\hat{D},\hat{\lambda}) \ge \mathcal{L}(\hat{D},\lambda') - \nu$. This implies that

$$\mathcal{L}(\hat{D},\hat{\lambda}) \ge err(\hat{D},\mathcal{P}) + C\,\Phi_{\hat{\bullet}}(\hat{D},\hat{g}) - \nu \tag{15}$$

Note that $\mathcal{L}(D^*,\hat{\lambda}) \le err(D^*,\mathcal{P})$, and so

$$\mathcal{L}(\hat{D},\hat{\lambda}) \le \min_{D\in\Delta_{\mathcal{H}(S)}} \mathcal{L}(D,\hat{\lambda}) + \nu \le \mathcal{L}(D^*,\hat{\lambda}) + \nu \tag{16}$$

Combining Equations (15) and (16), we get

$$err(\hat{D},\mathcal{P}) + C\,\Phi_{\hat{\bullet}}(\hat{D},\hat{g}) \le \mathcal{L}(\hat{D},\hat{\lambda}) + \nu \le \mathcal{L}(D^*,\hat{\lambda}) + 2\nu \le err(D^*,\mathcal{P}) + 2\nu$$

Note that $C\,\Phi_{\hat{\bullet}}(\hat{D},\hat{g}) \ge 0$, so we must have $err(\hat{D},\mathcal{P}) \le err(D^*,\mathcal{P}) + 2\nu = \mathrm{OPT} + 2\nu$. Furthermore, since $err(\hat{D},\mathcal{P}), err(D^*,\mathcal{P}) \in [0,1]$, we know

$$C\,\Phi_{\hat{\bullet}}(\hat{D},\hat{g}) \le 1 + 2\nu,$$

which implies that maximum constraint violation satisfies $\Phi_{\hat{\bullet}}(\hat{D},\hat{g}) \le (1+2\nu)/C$. By applying Claim 4.4, we get

$$\alpha_{FP}(g,\mathcal{P})\,\beta_{FP}(g,\hat{D},\mathcal{P}) \le \gamma + \frac{1+2\nu}{C}.$$

Now let us consider the case in which $\hat{D}$ is a feasible solution for the optimization problem. Then it follows that there is no constraint violation by $\hat{D}$ and $\max_\lambda \mathcal{L}(\hat{D},\lambda) = err(\hat{D},\mathcal{P})$, and so

$$err(\hat{D},\mathcal{P}) = \max_\lambda \mathcal{L}(\hat{D},\lambda) \le \mathcal{L}(\hat{D},\hat{\lambda}) + \nu \le \min_D \mathcal{L}(D,\hat{\lambda}) + 2\nu \le \mathcal{L}(D^*,\hat{\lambda}) + 2\nu \le err(D^*,\mathcal{P}) + 2\nu$$

Therefore, the stated bounds hold for both cases. □

Lemma 4.8. *Fix any* $\overline{D} \in \Delta_{\mathcal{H}(S)}$ *such that that* $\max_{g\in\mathcal{G}(S)}\{\Phi_+(\overline{D},g),\Phi_-(\overline{D},g)\} > 0$. *Let* $\lambda' \in \Lambda$ *be vector with one non-zero coordinate* $(\lambda')^{\bullet'}_{g'} = C$, *where*

$$(g',\bullet') = \operatorname*{argmax}_{(g,\bullet)\in\mathcal{G}(S)\times\{\pm\}} \{\Phi_\bullet(\overline{D},g)\}$$

Then $\mathcal{L}(\overline{D},\lambda') \ge \max_{\lambda\in\Lambda} \mathcal{L}(\overline{D},\lambda)$.

Proof of Lemma 4.8. Observe:

$$\begin{aligned}\operatorname*{argmax}_{\lambda\in\Lambda} \mathcal{L}(\overline{D},\lambda) &= \operatorname*{argmax}_{\lambda\in\Lambda} \mathop{\mathbb{E}}_{h\sim\overline{D}}[err(h,\mathcal{P})] + \sum_{g\in\mathcal{G}(S)} \left(\lambda^+_g\,\Phi_+(\overline{D},g) + \lambda^-_g\,\Phi_-(\overline{D},g)\right)\\ &= \operatorname*{argmax}_{\lambda\in\Lambda} \sum_{g\in\mathcal{G}} \left(\lambda^+_g\,\Phi_+(\overline{D},g) + \lambda^-_g\,\Phi_-(\overline{D},g)\right)\end{aligned}$$

Note that this is a linear optimization problem over the non-negative orthant of a scaling of the ℓ_1 ball, and so has a solution at a vertex, which corresponds to a single group $g \in \mathcal{G}(S)$. Thus, there is always a best response λ' that puts all the weight C on the coordinate $(\lambda')^{\bullet}_g$ that maximizes $\Phi_{\bullet}(\overline{D}, g)$. □

Lemma 4.10. *Let T be the time horizon for the no-regret dynamics. Let $D^1, \ldots, D^T$ be the sequence of distributions maintained by the Learner's FTPL algorithm with $\eta = \frac{n}{(1+C)}\sqrt{\frac{1}{\sqrt{n}T}}$, and $\lambda^1, \ldots, \lambda^T$ be the sequence of plays by the Auditor. Then*

$$\sum_{t=1}^{T} \mathop{\mathbb{E}}_{h\sim D^t}\left[U(h, \lambda^t)\right] - \min_{h\in\mathcal{H}(S)} \sum_{t=1}^{T} U(h, \lambda^t) \leq 2n^{1/4}(1+C)\sqrt{T}$$

Proof of Lemma 4.10. To instantiate the regret bound in Theorem 2.9, we just need to provide a bound on the maximum absoluate value over the coordinates of the loss vector (the quantity M in Theorem 2.9). For any $\lambda \in \Lambda$, the absolute value of the i-th coordinate of $\mathrm{LC}(\lambda)$ is bounded by:

$$\begin{aligned}
&\frac{1}{n} + \left|\frac{1}{n}\sum_{g\in\mathcal{G}(S)} (\lambda_g^+ - \lambda_g^-)\left(\Pr[g(x) = 1 \mid y = 0] - 1\right) \mathbf{1}[g(x_i) = 1]\right| \\
\leq &\frac{1}{n} + \frac{1}{n}\left(\sum_{g\in\mathcal{G}(S)} \left|\lambda_g^+ - \lambda_g^-\right|\right) \max_{g\in\mathcal{G}(S)} \left(\Pr[g(x) = 1 \mid y = 0]\mathbf{1}g(x_i) = 1\right) \\
\leq &\frac{1}{n} + \frac{1}{n}\left(\sum_{g\in\mathcal{G}(S)} \left|\lambda_g^+\right| + \left|\lambda_g^-\right|\right) \leq \frac{1+C}{n}
\end{aligned}$$

Also note that the dimension of the optimization is the size of the dataset n. This means if we set $\eta = \frac{n}{(1+C)}\sqrt{\frac{1}{\sqrt{n}T}}$, the regret of the learner will then be bounded by $2n^{1/4}(1+C)\sqrt{T}$. □

Lemma 4.11. *Fix any $\xi, \delta \in (0,1)$ and any distribution D over $\mathcal{H}(S)$. Let $h^1, \ldots, h^m$ be m i.i.d. draws from p, and $\hat{D}$ be the empirical distribution over the realized sample. Then with probability at least $1-\delta$ over the random draws of h^j's, the following holds,*

$$\max_{\lambda\in\Lambda} \left| \mathop{\mathbb{E}}_{h\sim\hat{D}}[U(h,\lambda)] - \mathop{\mathbb{E}}_{h\sim D}[U(h,\lambda)] \right| \leq \xi,$$

as long as $m \geq c_0 \frac{C^2(\ln(1/\delta) + d_2 \ln(n))}{\xi^2}$ for some absolute constant c_0 and $d_2 = \mathrm{VCDIM}(\mathcal{G})$.

Proof of Lemma 4.11. Recall that for any distribution D' over $\mathcal{H}(S)$ the expected payoff function is defined as

$$\begin{aligned}
\mathop{\mathbb{E}}_{h\sim\hat{D}}[U(h,\lambda)] - \mathop{\mathbb{E}}_{h\sim D}[U(h,\lambda)] = &\mathop{\mathbb{E}}_{h\sim\hat{D}}[err(h,\mathcal{P})] + \mathop{\mathbb{E}}_{h\sim\hat{D}}\left[\sum_{g\in\mathcal{G}(S)} \left(\lambda_g^+\Phi_+(h,g) + \lambda_g^-\Phi_-(h,g)\right)\right] \\
&- \mathop{\mathbb{E}}_{h\sim D}[err(h,\mathcal{P})] + \mathop{\mathbb{E}}_{h\sim D}\left[\sum_{g\in\mathcal{G}(S)} \left(\lambda_g^+\Phi_+(h,g) + \lambda_g^-\Phi_-(h,g)\right)\right]
\end{aligned}$$

By the triangle inequality, it suffices to show that with probability $(1-\delta)$, $A = |\mathbb{E}_{h\sim D}[err(h,\mathcal{P})] - \mathbb{E}_{h\sim\hat{D}}[err(h,\mathcal{P})]| \le \xi/2$ and for all $\lambda \in \Lambda$ and $g \in \mathcal{G}(S)$,

$$B = \left| \mathop{\mathbb{E}}_{h\sim\hat{D}} \left[\sum_{g\in\mathcal{G}(S)} \left(\lambda_g^+ \Phi_+(h,g) + \lambda_g^- \Phi_-(h,g) \right) \right] - \mathop{\mathbb{E}}_{h\sim D} \left[\sum_{g\in\mathcal{G}(S)} \left(\lambda_g^+ \Phi_+(h,g) + \lambda_g^- \Phi_-(h,g) \right) \right] \right| \le \xi/2$$

The first part follows directly from a simple application of the Chernoff-Hoeffding bound (Theorem A.1): with probability $(1-\delta/2)$, $A \le \xi/2$, as long as $m \ge 2\ln(4/\delta)/\xi^2$.

To bound the second part, we first note that by Hölder's inequality, we have

$$B \le \|\lambda\|_1 \max_{(g,\bullet)\in\mathcal{G}(S)\times\{\pm\}} |\Phi_\bullet(D,g) - \Phi_\bullet(\hat{D},g)|$$

Since for all $\lambda \in \Lambda$ we have $\|\lambda\|_1 \le C$, it suffices to show that with probability $1-\delta/2$, $|\Phi_\bullet(D,g) - \Phi_\bullet(\hat{D},g)| \le \xi/(2C)$ holds for all $\bullet \in \{-,+\}$ and $g \in \mathcal{G}(S)$. Note that

$$\begin{aligned} |\Phi_\bullet(D,g) - \Phi_\bullet(\hat{D},g)| = \Bigg| &\left(\mathop{\mathbb{E}}_{h\sim D}[\mathrm{FP}(h)] - \mathop{\mathbb{E}}_{h\sim\hat{D}}[\mathrm{FP}(h)] \right) \Pr[y=0, g(x)=1] \\ &+ \left(\mathop{\mathbb{E}}_{h\sim D}[\Pr[h(X)=1, y=0, g(x)=1]] - \mathop{\mathbb{E}}_{h\sim\hat{D}}[\Pr[h(X)=1, y=0, g(x)=1]] \right) \Bigg| \end{aligned}$$

We can rewrite the absolute value of first term:

$$\begin{aligned} &\left| \left(\mathop{\mathbb{E}}_{h\sim D}[\mathrm{FP}(h)] - \mathop{\mathbb{E}}_{h\sim\hat{D}}[\mathrm{FP}(h)] \right) \Pr[y=0, g(x)=1] \right| \\ =&\left| \left(\mathop{\mathbb{E}}_{h\sim D}[\Pr[h(X)=1 \mid y=0]] - \mathop{\mathbb{E}}_{h\sim\hat{D}}[\Pr[h(X)=1 \mid y=0]] \right) \Pr[g(x)=1 \mid y=0] \right| \\ \le&\left| \left(\mathop{\mathbb{E}}_{h\sim D}[\Pr[h(X)=1, y=0]] - \mathop{\mathbb{E}}_{h\sim\hat{D}}[\Pr[h(X)=1, y=0]] \right) \right| \end{aligned}$$

where the last inequality follows from $\Pr[g(x)=1 \mid y=0] \le 1$.

Note that $\mathbb{E}_{h\sim\hat{D}}[\Pr[h(X)=1, y=0, g(x)=1]] = \frac{1}{m}\sum_{j=1}^m \Pr[h^j(X)=1, y=0, g(x)=1]$, which is an average of m i.i.d. random variables with expectation $\mathbb{E}_{h\sim D}[\Pr[h(X)=1, y=0, g(x)=1]]$. By the Chernoff-Hoeffding bound (Theorem A.1), we have

$$\Pr\left[\left| \mathop{\mathbb{E}}_{h\sim D}[\Pr[h(X)=1, y=0]] - \mathop{\mathbb{E}}_{h\sim\hat{D}}[\Pr[h(X)=1, y=0]] \right| > \frac{\xi}{4C} \right] \le 2\exp\left(-\frac{\xi^2 m}{8C^2} \right) \tag{17}$$

In the following, we will let $\delta_0 = 2\exp\left(-\frac{\xi^2 m}{8C^2}\right)$. Similarly, we also have for each $g \in \mathcal{G}(S)$,

$$\Pr\left[\left| \mathop{\mathbb{E}}_{h\sim D}[\Pr[h(X)=1, y=0, g(x)=1]] - \mathop{\mathbb{E}}_{h\sim\hat{D}}[\Pr[h(X)=1, y=0, g(x)=1]] \right| > \frac{\xi}{4C} \right] \le \delta_0 \tag{18}$$

By taking the union bound over (17) and (18) over all choices of $g \in \mathcal{G}(S)$, we have with probability at least $(1-\delta_0(1+|\mathcal{G}(S)|))$,

$$\left| \mathop{\mathbb{E}}_{h\sim D}[\Pr[h(X)=1, y=0]] - \mathop{\mathbb{E}}_{h\sim\hat{D}}[\Pr[h(X)=1, y=0]] \right| \le \frac{\xi}{4C} \tag{19}$$

and,

$$\left| \mathop{\mathbb{E}}_{h\sim D}[\Pr[h(X)=1, y=0, g(x)=1]] - \mathop{\mathbb{E}}_{h\sim\hat{D}}[\Pr[h(X)=1, y=0, g(x)=1]] \right| \le \frac{\xi}{4C} \quad \text{for all } g \in \mathcal{G}(S). \tag{20}$$

Note that by Sauer's lemma (Lemma 4.3), $|\mathcal{G}(S)| \le O\left(n^{d_2}\right)$. Thus, there exists an absolute constant c_0 such that $m \ge c_0 \frac{C^2(\ln(1/\delta)+d_2\ln(n))}{\xi^2}$ implies that failure probability above $\delta_0(1+|\mathcal{G}(S)|) \le \delta/2$. We will assume m satisifies such a bound, and so the events of (19) and (20) hold with probaility at least $(1-\delta/2)$. Then by the triangle inequality we have for all $(g,\bullet) \in \mathcal{G}(S)\times\{\pm\}$, $|\Phi_\bullet(D,g) - \Phi_\bullet(\hat{D},g)| \le \xi/(2C)$, which implies that $B \le \xi/2$. This completes the proof. □

Claim C.1. *Suppose there are two distributions D and $\hat{D}$ over $\mathcal{H}(S)$ such that*

$$\max_{\lambda\in\Lambda}\left|\mathop{\mathbb{E}}_{h\sim\hat{D}}[U(h,\lambda)] - \mathop{\mathbb{E}}_{h\sim D}[U(h,\lambda)]\right| \le \xi.$$

Let

$$\hat{\lambda} \in \operatorname*{argmax}_{\lambda'\in\Lambda} \mathop{\mathbb{E}}_{h\sim\hat{D}}[U(h,\lambda')]$$

Then

$$\max_{\lambda} \mathop{\mathbb{E}}_{h\sim D}[U(h,\lambda)] - \xi \le \mathop{\mathbb{E}}_{h\sim D}\left[U(h,\hat{\lambda})\right],$$

Lemma 4.12. *Let T be the time horizon for the no-regret dynamics. Let $D^1,\dots,D^T$ be the sequence of distributions maintained by the Learner's FTPL algorithm. For each D^t, let $\hat{D}^t$ be the empirical distribution over m i.i.d. draws from D^t. Let $\lambda^1,\dots,\lambda^T$ be the Auditor's best responses against $\hat{D}^1,\dots,\hat{D}^T$. Then with probability $1-\delta$,*

$$\max_{\lambda\in\Lambda}\sum_{t=1}^{T}\mathop{\mathbb{E}}_{h\sim D^t}[U(h,\lambda)] - \sum_{t=1}^{T}\mathop{\mathbb{E}}_{h\sim D^t}\left[U(h,\lambda^t)\right] \le T\sqrt{\frac{c_0C^2(\ln(T/\delta)+d_2\ln(n))}{m}}$$

for some absolute constant c_0 and $d_2 = \mathrm{VCDIM}(\mathcal{G})$.

Proof. Let γ_A^t be defined as

$$\gamma_A^t = \max_{\lambda\in\Lambda}\left|\mathop{\mathbb{E}}_{h\sim\hat{D}^t}[U(h,\lambda)] - \mathop{\mathbb{E}}_{h\sim D^t}[U(h,\lambda)]\right|$$

By instantiating Lemma 4.11 and applying union bound across all T steps, we know with probability at least $1-\delta$, the following holds for all $t\in[T]$:

$$\gamma_A^t \le \sqrt{\frac{c_0C^2(\ln(T/\delta)+d_2\ln(n))}{m}}$$

where c_0 is the absolute constant in Lemma 4.11 and $d_2 = \mathrm{VCDIM}(\mathcal{G})$.

Note that by Claim C.1, the Auditor is performing a γ_A^t-approximate best response at each round t. Then we can bound the Auditor's regret as follows:

$$\begin{aligned}\gamma_A = \frac{1}{T}\left[\max_{\lambda\in\Lambda}\sum_{t=1}^{T}\mathop{\mathbb{E}}_{h\sim D^t}[U(h,\lambda)] - \sum_{t=1}^{T}\mathop{\mathbb{E}}_{h\sim D^t}\left[U(h,\lambda^t)\right]\right] &\le \frac{1}{T}\sum_{t=1}^{T}\left(\max_{\lambda\in\Lambda}\mathop{\mathbb{E}}_{h\sim D^t}[U(h,\lambda)] - \mathop{\mathbb{E}}_{h\sim D^t}\left[U(h,\lambda^t)\right]\right)\\ &\le \max_T \gamma_A^t\end{aligned}$$

It follows that with probability $1-\delta$, we have

$$\gamma_A \le \sqrt{\frac{c_0C^2(\ln(T/\delta)+d_2\ln(n))}{m}}$$

which completes the proof. □