Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness

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

Abstract

The most prevalent notions of fairness in machine learning fix a small collection of pre-defined groups (such as race or gender), and then ask for approximate parity of some statistic of the classifier (such as false positive rate) across these groups. Constraints of this form are susceptible to *fairness gerrymandering*, in which a classifier is fair on each individual group, but badly violates the fairness constraint on structured *subgroups*, such as certain combinations of protected attribute values. We thus consider fairness across exponentially or infinitely many subgroups, defined by a structured class of functions over the protected attributes. We show the problem of auditing subgroup fairness for both equality of false positive rates and statistical parity is computationally equivalent to the problem of weak agnostic learning — which means it is hard in the worst case, even for simple structured subclasses. We then derive an algorithm that provably converges in a polynomial number of steps to the best subgroup-fair distribution over classifiers (given access to an oracle which can solve the agnostic learning problem), and show that we can effectively both audit and learn fair classifiers on a real dataset. The full version of our paper is available on arXiv (Kearns et al., 2017).

1. Introduction

Approaches to fairness in machine learning can be broadly 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, such as equality of false positive or negative rates. These statistical notions of fairness are the most common in the literature (see e.g. (Kamiran & Calders, 2012; Hajian & 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 (Joseph et al., 2016; Dwork et al., 2012). 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 ask 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. Similar examples for classification are easily constructed.*

To avoid such problems, we would like to be able to satisfy a

[*] Equal contribution [1] University of Pennsylvania, Philadelphia, PA, USA [2] Mcrosoft Research, New York City, NY, USA. Correspondence to: Michael Kearns <mkearns@cis.upenn.edu>, Seth Neel <sethneel@wharton.upenn.edu>, Aaron Roth <aaroth@cis.upenn.edu>, Zhiwei Steven Wu <steven7woo@gmail.com>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

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.

We consider the problem of *auditing* binary classifiers for stastical parity and equality of false positive/negative rates, 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. It is, however, sensible to ask for fairness for large *structured* subsets of individuals: so long as these subsets have (say) bounded VC dimension, the *statistical* problem of learning and auditing fair classifiers is easy, so long as the dataset is sufficiently large. Our investigation focuses on the computational challenges, both in theory and in practice.

1.1. Our Results

Briefly, our contributions are:

- Formalization of the problems of auditing and learning classifiers for fairness with respect to rich classes of subgroups $\mathcal{G}$.
- Results proving (under certain assumptions) the computational equivalence of auditing $\mathcal{G}$ and (weak) agnostic learning of $\mathcal{G}$. While these results imply theoretical intractability of auditing for some natural classes $\mathcal{G}$, they also suggest that practical machine learning heuristics can be applied to the auditing problem.
- A provably convergent, polynomial-time algorithm for learning classifiers that are fair with respect to $\mathcal{G}$ under the assumption of oracles for agnostic learning. The algorithm is formulated as a two-player zero-sum game between a Learner (the primal player) and an Auditor (the dual player). The Learner uses *Follow the Perturbed Leader* (Kalai & Vempala, 2005) and the Auditor uses best response dynamics, leading to efficient no-regret convergence to approximate equilibrium.
- An implementation of a variant of the algorithm in which both players use Fictitious Play and heuristic oracles, and a demonstration of its effectiveness on a real dataset in which subgroup fairness is a concern.

Related Work Independent of our work, (Hébert-Johnson et al., 2017) also consider a related and complementary notion of fairness that they call “multicalibration”. They give an algorithmic result that is broadly similar to the one we give for learning subgroup fair classifiers, but our techniques differ from theirs significantly. Technically, the most closely related piece of work (and from which we take inspiration for our algorithm in Section 3) 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. Though it is beyond the scope of this short paper, experimental results show that in practice equalizing rates across polynomially many subgroups determined by individual features does not prevent imbalance in natural subgroups that are protected under our model (e.g. groups determined by simple functions of protected features). 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, leading to our use and analysis of Follow the Perturbed Leader for the Learner and best response for the Auditor.

2. Model and Preliminaries

We model each individual in a population as 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. 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 (which may be either the result of learning, or fixed code written by a human), 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 $\mathcal{P}_{\text{audit}}(D)$ 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 the fairness constraint of equality of false positive rates (also known as *equality of opportunity*).[1] Auditing for equality of false negative rates is symmetric and so we do not explicitly consider it. The 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 (such as black women over age 55).

Definition 2.1 (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*

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

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

where $\mathrm{FP}(D) = \Pr_{D,\mathcal{P}}[D(X) = 1 \mid y = 0]$ *and* $\mathrm{FP}(D, g) = \Pr_{D,\mathcal{P}}[D(X) = 1 \mid g(x) = 1, y = 0]$ *denote the overall false positive rate of D and the false positive rate of D on group g respectively. We say D satisfies γ-***False Positive (FP) Fairness** *with respect to $\mathcal{P}$ and $\mathcal{G}$ if for every $g \in \mathcal{G}$*

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

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

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 a γ-*unfair certificate* for $(D, \mathcal{P})$. An *auditing algorithm* for either notion of fairness is given sample access to $\mathcal{P}_{\text{audit}}(D)$ for some classifier D. It will either deem D to be fair with respect to $\mathcal{P}$, or produce a certificate of unfairness, under polynomial time. As we show in the full version, auditing of a given class is closely related to weak agnostic learning (Kearns et al., 1994; Kalai et al., 2008) of that same class. (See the full version for the formal definitions.) This means that from a worst-case 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 obviously 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 subsequent development of our algorithmic results and their experimental evaluation.

3. Learning With Subgroup Fairness

We provide 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 $\mathcal{P}$ 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 error subject to the fairness constraint over $\mathcal{G}$. We will design an iterative algorithm that, when given access to a agnostic learning oracle, computes an optimal randomized classifier in polynomial time.

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

$$\min_{p \in \Delta_{\mathcal{H}}} \mathop{\mathbb{E}}_{h \sim D} [err(h, \mathcal{P})] \quad (1)$$

$$\text{such that } \forall g \in \mathcal{G} \quad \alpha_{FP}(g, \mathcal{P})\, \beta_{FP}(g, D, \mathcal{P}) \le \gamma. \quad (2)$$

where $err(h, \mathcal{P}) = \Pr_{\mathcal{P}}[h(x, x') \neq y]$, and the quantities α_{FP} and β_{FP} are as in Definition 2.1. We will write OPT to denote the objective value at the optimum for the Fair ERM problem, which is the minimum error achieved by a γ-fair distribution over the class $\mathcal{H}$.

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

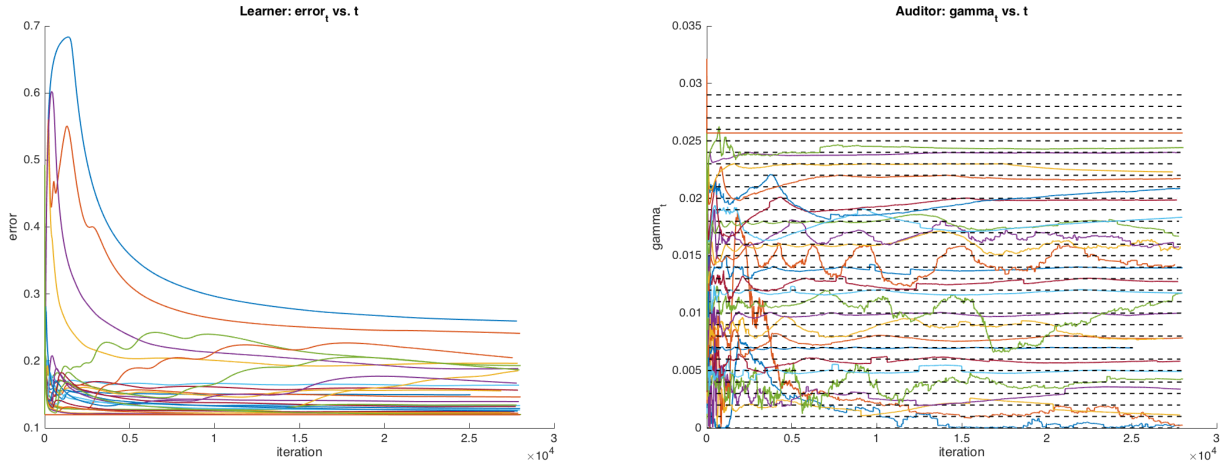
Theorem 3.1. *Fix any $\nu, \delta \in (0, 1)$. Then there exists an algorithm that, given as input n data points with empirical distribution $\mathcal{P}$, parameters ν, δ, and access to agnostic learning oracles over $\mathcal{H}$ and $\mathcal{G}$, runs in polynomial time, and with probability at least $1 - \delta$, outputs a randomized classifier $\hat{D}$ such that $err(\hat{D}, \mathcal{P}) \le \mathrm{OPT} + \nu$, and for all $g \in \mathcal{G}$, obeys the fairness constraint*

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

4. 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. 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. Fictitious Play is only known to converge to equilibrium in the limit (Robinson, 1951), rather than in a polynomial number of rounds – nevertheless it performs well on real data, despite the fact that it has weaker theoreti-

[1] In the full version, we also consider the constraint of equality of classification rates (also known as *statistical parity*).



(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 2.1. See text for details.

cal guarantees compared to the algorithm we presented in the last section.

4.1. 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[2]. Each record in this dataset describes the aggregate demographic properties of a different U.S. community. The total number of records is 1994, and the number of features is 122. The variable we predict is the binary indicator of whether the rate of violent crime in the community is above the 70th percentile of that value, indicating it is a high-crime community. 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.

For our experiments, both Learner and Auditor use the model class of hyperplanes, over 122 and 18 features respectively. We implement the cost sensitive classification oracle via a two stage regression procedure which classifies a data point according to whether the cost for predicting 0 or 1 is higher, as estimated by separate linear regressions. We set the parameter $C = 10$, 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, this relatively small value seems to suffice for (approximate) convergence. The other and more meaningful parameter of the algorithm is the bound γ, which controls the amount of unfairness permitted. Ideally, and as we shall see, varying γ allows us to trace out a menu of trade-offs between accuracy and fairness.

4.2. Results

We examine 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. 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. 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. It is clear that the exhibited relationship between γ and error is largely sensible, and that we trace out a useful pareto curve of error vs. γ (though space precludes showing it). The question of which precise γ is appropriate is most certainly task-specific, and is left as a policy question for the practitioner. To aid in their quest we have given them a powerful practical tool for auditing and learning classifiers with respect to a rich family of fairness guarantees.

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

References

Agarwal, Alekh, Beygelzimer, Alina, Dudík, Miroslav, and Langford, John. A reductions approach to fair classification. *Fairness, Accountability, and Transparency in Machine Learning (FATML)*, 2017. URL http://fatml.mysociety.org/media/documents/reductions_approach_to_fair_classification.pdf.

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

Dwork, Cynthia, Hardt, Moritz, Pitassi, Toniann, Reingold, Omer, and Zemel, Richard. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 214–226. ACM, 2012.

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

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

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

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

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

Kalai, Adam Tauman and Vempala, Santosh. Efficient algorithms for online decision problems. *J. Comput. Syst. Sci.*, 71(3):291–307, 2005. doi: 10.1016/j.jcss.2004.10.016. URL https://doi.org/10.1016/j.jcss.2004.10.016.

Kalai, Adam Tauman, Mansour, Yishay, and Verbin, Elad. 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*, pp. 629–638, 2008. doi: 10.1145/1374376.1374466. URL http://doi.acm.org/10.1145/1374376.1374466.

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

Kearns, Michael, Neel, Seth, Roth, Aaron, and Wu, Zhiwei Steven. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *arXiv preprint arXiv:1711.05144*, 2017.

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

Kleinberg, Jon, Mullainathan, Sendhil, and Raghavan, Manish. 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.

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

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