Metric-Free Individual Fairness in Online Learning

Yahav Bechavod [1] **Christopher Jung** [2] **Zhiwei Steven Wu** [3]

Abstract

We study an online learning problem subject to the constraint of individual fairness, which requires that similar individuals are treated similarly. Unlike prior work on individual fairness, we do not assume the similarity measure among individuals is known, nor do we assume that such measure takes a certain parametric form. Instead, we leverage the existence of an *auditor* who detects fairness violations without enunciating the quantitative measure. In each round, the auditor examines the learner's decisions and attempts to identify a pair of individuals that are treated unfairly by the learner. We provide a general reduction framework that reduces online classification in our model to standard online classification, which allows us to leverage existing online learning algorithms to achieve sub-linear regret and number of fairness violations. Surprisingly, in the stochastic setting where the data are drawn independently from a distribution, we are also able to establish PAC-style fairness and accuracy generalization guarantees ((23)), despite only having access to a very restricted form of fairness feedback. Our fairness generalization bound qualitatively matches the uniform convergence bound of (23), while also providing a meaningful accuracy generalization guarantee. Our results resolve an open question by (9) by showing that online learning under an unknown individual fairness constraint is possible even without assuming a strong parametric form of the underlying similarity measure.[1]

1. Introduction

As machine learning increasingly permeates many critical aspects of society, including education, healthcare, criminal justice, and lending, there is by now a vast literature that studies how to make machine learning algorithms fair (see, e.g., (5); (22); (6)). Most of the work in this literature tackles the problem by taking the *statistical group fairness* approach that first fixes a small collection of high-level groups defined by protected attributes (e.g., race or gender) and then asks for approximate parity of some statistic of the predictor, such as positive classification rate or false positive rate, across these groups (see, e.g., (11; 4; 21; 1)). While notions of group fairness are easy to operationalize, they are aggregate in nature without fairness guarantees for finer subgroups or individuals (7; 12; 19).

In contrast, the *individual fairness* approach aims to address this limitation by asking for explicit fairness criteria at an individual level. In particular, the compelling notion of individual fairness proposed in the seminal work of (7) requires that similar people are treated similarly. The original formulation of individual fairness assumes that the algorithm designer has access to a task-specific fairness metric that captures how similar two individuals are in the context of the specific classification task at hand. In practice, however, such a fairness metric is rarely specified, and the lack of metrics has been a major obstacle for the wide adoption of individual fairness. There has been recent work on learning the fairness metric based on different forms of human feedback. For example, (13) provides an algorithm for learning the metric by presenting human arbiters with queries concerning the distance between individuals, and (9) provide an online learning algorithm that can eventually learn a Mahalanobis metric based on identified fairness violations. While these results are encouraging, they are still bound by several limitations. In particular, it might be difficult for humans to enunciate a precise quantitative similarity measure between individuals. Moreover, their similarity measure across individuals may not be consistent with any metric (e.g., it may not satisfy the triangle inequality) and is unlikely to be given by a simple parametric function (e.g., the Mahalanobis metric function).

To tackle these issues, this paper studies *metric-free* online learning algorithms for individual fairness that rely on a

[1]School of Computer Science and Engineering, Hebrew University [2]Computer and Information Science Department, University of Pennsylvania [3]Department of Computer Science, University of Minnesota. Correspondence to: Yahav Bechavod <yahav.bechavod@cs.huji.ac.il>, Christopher Jung <chrjung@seas.upenn.edu>, Zhiwei Steven Wu <zsw@umn.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020. Copyright 2020 by the author(s).

[1] The full version of this paper is available at: `https://arxiv.org/abs/2002.05474`.

weaker form of interactive human feedback and minimal assumptions on the similarity measure across individuals. Similar to the prior work of (9), we do not assume a prespecified metric, but instead assume access to an *auditor*, who observes the learner's decisions over a group of individuals that show up in each round and attempts to identify a fairness violation—a pair of individuals in the group that should have been treated more similarly by the learner. Since the auditor only needs to identify such unfairly treated pairs, there is no need for them to enunciate a quantitative measure – to specify the distance between the identified pairs. Moreover, we do not impose any parametric assumption on the underlying similarity measure, nor do we assume that it is actually a metric since we do not require that similarity measure to satisfy the triangle inequality. Under this model, we provide a general reduction framework that can take any online classification algorithm (without fairness constraint) as a black-box and obtain a learning algorithm that can simultaneously minimize cumulative classification loss and the number of fairness violations. Our results in particular remove many strong assumptions in (9), including their parametric assumptions on linear rewards and Mahalanobis distances, and thus answer several questions left open in their work. We include a more detailed overview of related work in the appendix.

1.1. Overview of Model and Results

We study an online classification problem: over rounds $t = 1, \dots, T$, a learner observes a small set of k individuals with their feature vectors $(x_\tau^t)_{\tau=1}^k$ in space $\mathcal{X}$. The learner tries to predict the label $y_k^t \in \{0, 1\}$ of each individual with a "soft" predictor π^t that predicts $\pi^t(x_\tau^t) \in [0, 1]$ on each x_τ^t and incurs classification loss $|\pi^t(x_\tau^t) - y_\tau^t|$. Then an auditor will investigate if the learner has violated the individual fairness constraint on any pair of individuals within this round, that is, if there exists $(\tau_1, \tau_2) \in [k]^2$ such that $|\pi^t(x_{\tau_1}^t) - \pi^t(x_{\tau_2}^t)| > d(x_{\tau_1}^t, x_{\tau_2}^t) + \alpha$, where $d\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ is an unknown distance function and α denotes the auditor's tolerance. If this violation has occurred on any number of pairs, the auditor will identify one of such pairs and incur a fairness loss of 1; otherwise, the fairness loss is 0. Then the learner will update the predictive policy based on the observed labels and the received fairness feedback. Under this model, our results include:

A Reduction from Fair Online Classification to Standard Online Classification. Our reduction-based algorithm can take any no-regret online (batch) classification learner as a black-box and achieve sub-linear cumulative fairness loss and sub-linear regret on mis-classification loss compared to the most accurate policy that is fair on every round. In particular, our framework can leverage the generic exponential weights method (8; 3; 2) and also oracle-efficient methods, including variants of Follow-the-Perturbed-Leader (FTPL) (e.g., (25; 24)), that further reduces online learning to standard supervised learning or optimization problems. We instantiate our framework using two online learning algorithms (exponential weights and CONTEXT-FTPL), both of which obtain a $\tilde{O}(\sqrt{T})$ on misclassification regret and cumulative fairness loss.

Fairness and Accuracy Generalization Guarantees. While our algorithmic results hold under adversarial arrivals of the individuals, in the stochastic arrivals setting we show that the uniform average policy over time is probably approximate correct and fair (PACF) (23)–that is, the policy is approximately fair on almost all random pairs drawn from the distribution and nearly matches the accuracy gurantee of the best fair policy. In particular, we show that the average policy π^{avg} with high probability satisfies

$$\Pr_{x,x'}[|\pi^{avg}(x) - \pi^{avg}(x')| > \alpha + 1/T^{1/4}] \le O(1/T^{1/4}),$$

which qualitatively achieves similar PACF uniform convergence sample complexity as (23).[2] However, we establish our generalization guarantee through fundamentally different techniques. While their work assumes a fully specified metric and i.i.d. data, the learner in our setting can only access the similarity measure through an auditor's limited fairness violations feedback. The main challenge we need to overcome is that the fairness feedback is inherently adaptive–that is, the auditor only provides feedback for the sequence of deployed policies, which are updated adaptively over rounds. In comparison, a fully known metric allows the learner to evaluate the fairness guarantee of all policies simultaneously. As a result, we cannot rely on their uniform convergence result to bound the fairness generalization error, but instead we leverage a probabilistic argument that relates the learner's regret to the distributional fairness guarantee.

2. Model and Preliminaries

We define the instance space to be $\mathcal{X}$ and its label space to be $\mathcal{Y}$. Throughout this paper, we will restrict our attention to binary labels, that is $\mathcal{Y} = \{0, 1\}$. We write $\mathcal{H} : \mathcal{X} \to \mathcal{Y}$ to denote the hypothesis class and assume that $\mathcal{H}$ contains a constant hypothesis – i.e. there exists h such that $h(x) = 0$ for all $x \in \mathcal{X}$. Also, we allow for convex combination of hypotheses for the purpose of randomizing the prediction and denote the simplex of hypotheses by $\Delta\mathcal{H}$; we call a randomized hypothesis a *policy*. Sometimes, we assume the existence of an underlying (but unknown) distribution $\mathcal{D}$ over $(\mathcal{X}, \mathcal{Y})$. For each prediction $\hat{y} \in \mathcal{Y}$ and its true label $y \in \mathcal{Y}$, there is an associated misclassification loss,

[2] (23) show that if a policy π is α-fair on all pairs in a i.i.d. dataset of size m, then π satisfies $\Pr_{x,x'}[|\pi(x) - \pi(x')| > \alpha + \epsilon] \le \epsilon$, as long as $m \ge \tilde{\Omega}(1/\epsilon^4)$.

$\ell(\hat{y}, y) = \mathbb{1}(\hat{y} \neq y)$. For simplicity, we overload the notation and write

$$\ell(\pi(x), y) = (1-\pi(x))\cdot y + \pi(x)\cdot(1-y) = \mathop{\mathbb{E}}_{h\sim\pi}[\ell(h(x), y)].$$

2.1. Individual Fairness and Auditor

We want our deployed policy π to behave fairly in some manner, and we use the individual fairness definition from (7) that asserts that "similar individuals should be treated similarly." We assume that there is some distance function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ over the instance space $\mathcal{X}$ which captures the distance between individuals in $\mathcal{X}$, although d doesn't have to satisfy the triangle inequality. The only requirement on d is that it is always non-negative and symmetric $d(x, x') = d(x', x)$.

Definition 2.1 ((α, β)-fairness)**.** *Assume $\alpha, \beta > 0$. A policy $\pi \in \Delta\mathcal{H}$ is said to be α-fair on pair (x, x'), if*

$$|\pi(x) - \pi(x')| \leq d(x, x') + \alpha.$$

We say policy π's α-fairness violation on pair (x, x') is

$$v_\alpha(\pi, (x, x')) = \max(0, |\pi(x) - \pi(x')| - d(x, x') - \alpha).$$

A policy is π is said to be (α, β)-fair on distribution $\mathcal{D}$, if

$$\Pr_{(x,x')\sim\mathcal{D}|_\mathcal{X}\times\mathcal{D}|_\mathcal{X}}[|\pi(x) - \pi(x')| > d(x, x') + \alpha] \leq \beta.$$

A policy π is said to be α-fair on set $S \subseteq \mathcal{X}$, if for all $(x, x') \in S^2$, it is α-fair.

Although individual fairness is intuitively sound, individual fairness notion requires the knowledge of the distance function d which is often hard to specify. Therefore, we rely on an auditor $\mathcal{J}$ that can detect instances of α-unfairness.

Definition 2.2 (Auditor $\mathcal{J}$)**.** *An auditor $\mathcal{J}_\alpha$ which can have its own internal state takes in a* reference set *$S \subseteq \mathcal{X}$ and a policy π. Then, it outputs ρ which is either $null$ or a pair of indices from the provided reference set to denote that there is some positive α-fairness violation for that pair. For some $S = (x_1, \ldots, x_n)$,*

$$\mathcal{J}_\alpha(S, \pi) = \begin{cases} \rho = (\rho_1, \rho_2) & \textit{there's } \alpha\textit{-fairness violation}(\rho_1, \rho_2) \\ null & \textit{otherwise} \end{cases}$$

If there exists multiple pairs with some α-violation, the auditor can choose one arbitrarily.

Remark 2.3. *Our assumptions on the auditor are much more relaxed than those of (9), which require that the auditor outputs whether the policy is 0-fair (i.e. with no slack) on all pairs S^2 exactly. Furthermore, the auditor in (9) can only handle Mahalanobis distances. In our setting, because of the internal state of the auditor, the auditor does not have to be a fixed function but rather can be adaptively changing in each round. Finally, we never rely on the fact the distance function d stays the same throughout rounds, meaning all our results extend to the case where the distance function governing the fairness constraints is changing every round.*

2.2. Online Batch Classification

We now describe our online batch classification setting. In each round $t = 1, \ldots, T$, the learner deploys some model $\pi^t \in \Delta\mathcal{H}$. Upon seeing the deployed policy π^t, the environment chooses a batch of k individuals, $(x^t_\tau, y^t_\tau)^k_{\tau=1}$ and possibly, a pair of individuals from that round on which π^t will be responsible for any α-fairness violation. For simplicity, we write $\bar{x}^t = (x^t_\tau)^k_{\tau=1}$ and $\bar{y}^t = (y^t_\tau)^k_{\tau=1}$. The strategy $z^t_{\text{FAIR-BATCH}} \in \mathcal{Z}_{\text{FAIR-BATCH}}$ that the environment chooses can be described by

$$z^t_{\text{FAIR-BATCH}} = (\bar{x}^t, \bar{y}^t) \times \rho^t,$$

where $\rho^t \in [k]^2 \cup \{null\}$. Often, we will omit the subscript and simply write z^t. If $\rho^t = (\rho^t_1, \rho^t_2)$, then π^t will be responsible for the α-fairness violation on the pair $(x^t_{\rho^t_1}, x^t_{\rho^t_2})$. There are two types of losses that we are interested in: misclassification and fairness loss.

Definition 2.4 (Misclassification Loss)**.** *The (batch) misclassification loss Err*[3] *is*

$$\textit{Err}(\pi, z^t) = \sum_{\tau=1}^{k} \ell(\pi(x^t_\tau), y^t_\tau).$$

Definition 2.5 (Fairness Loss)**.** *The α-fairness loss $\textit{Unfair}_\alpha$ is*

$$\textit{Unfair}_\alpha(\pi, z^t) = \begin{cases} \mathbb{1}\left(\pi(x^t_{\rho^t_1}) - \pi(x^t_{\rho^t_2}) - d(x^t_{\rho^t_1}, x^t_{\rho^t_2}) - \alpha > 0\right) & \textit{if } \rho^t = (\rho^t_1, \rho^t_2) \\ 0 & \textit{otherwise} \end{cases}$$

We want the total misclassification and fairness loss over T rounds to be as small as any $\pi^* \in Q$ for some competitor set Q, which we describe now. As said above, each round's reference set, a set of pairs for which the deployed policy will possibly be responsible in terms of α-fairness, will be defined in terms of the instances that arrive within that round $\bar{x}^t$. The baseline Q_α that we compete against will be all policies that are α-fair on $\bar{x}^t$ for all $t \in [T]$:

$$Q_\alpha = \{\pi \in \Delta\mathcal{H} : \pi \text{ is } \alpha\text{-fair on } \bar{x}^t \text{ for all } t \in [T]\}$$

Note that because $\mathcal{H}$ contains a constant hypothesis which must be 0-fair on all instances, Q_α cannot be empty. The

[3] We will overload the notation for this loss; regardless of what $\mathcal{Z}$ is, we'll assume $\text{Err}(\pi, z^t)$ is well-defined as long as z^t includes $(\bar{x}^t, \bar{y}^t)$.

Algorithm 1 Online Fair Batch Classification FAIR-BATCH

```
$t = 1, \ldots, T$ Learner deploys $\pi^t$
Environment chooses $(\bar{x}^t, \bar{y}^t)$
Environment chooses the pair $\rho^t$
$z^t = (\bar{x}^t, \bar{y}^t) \times \rho^t$
Learner incurs misclassfication loss $\mathrm{Err}(\pi^t, z^t)$
Learner incurs fairness loss $\mathrm{Unfair}(\pi^t, z^t)$
```

Algorithm 2 Online Batch Classification BATCH

```
$t = 1, \ldots, T$ Learner deploys $\pi^t$
Environment chooses $z^t = (\bar{x}^t, \bar{y}^t)$
Learner incurs misclassification loss $\mathrm{Err}(\pi^t, z^t)$
```

Figure 1. Comparison between Online Fair Batch Classification and Online Batch Classification: each is summarized by the interaction between the learner and the environment: $(\Delta\mathcal{H} \times \mathcal{Z}_{\text{FAIR-BATCH}})^T$ and $(\Delta\mathcal{H} \times \mathcal{Z}_{\text{BATCH}})^T$ where $\mathcal{Z}_{\text{FAIR-BATCH}} = \mathcal{X}^k \times \mathcal{Y}^k \times ([k]^2 \cup \{null\})$ and $\mathcal{Z}_{\text{BATCH}} = \mathcal{X}^k \times \mathcal{Y}^k$.

difference in total loss between our algorithm and a fixed π^* is called 'regret', which we formally define below.

Definition 2.6 (Algorithm $\mathcal{A}$)**.** *An algorithm $\mathcal{A} : (\Delta\mathcal{H} \times \mathcal{Z})^* \to \Delta\mathcal{H}$ takes in its past history $(\pi^\tau, z^\tau)_{\tau=1}^{t-1}$ and deploys a policy $\pi^t \in \Delta\mathcal{H}$ at every round $t \in [T]$.*

Definition 2.7 (Regret). *For some $Q \subseteq \Delta\mathcal{H}$, the regret of algorithm $\mathcal{A}$ with respect to some loss $L : \Delta\mathcal{H} \times \mathcal{Z} \to \mathbb{R}$ is denoted as $\mathbf{Regret}^L(\mathcal{A}, Q, T)$, if for any $(z_t)_{t=1}^T$,*

$$\sum_{t=1}^T L\left(\pi^t, z^t\right) - \inf_{\pi^* \in Q} \sum_{t=1}^T L\left(\pi^*, z^t\right) \leq \mathbf{Regret}^L(\mathcal{A}, Q, T),$$

where $\pi^t = \mathcal{A}((\pi^j, z^j)_{j=1}^{t-1})$. When it is not clear from the context, we will use subscript to denote the setting – e.g. $\mathbf{Regret}^L_{\text{FAIR-BATCH}}$.

We wish to develop an algorithm such that both the misclassfication and fairness loss regret is sublinear, which is often called no-regret. Note that because $\pi^* \in Q_\alpha$ is α-fair on $\bar{x}^t$ for all $t \in [T]$, we have $\mathrm{Unfair}_\alpha(\pi^*, z^t) = 0$ for all $t \in [T]$. Hence, achieving $\mathbf{Regret}^{\text{Unfair}_\alpha}_{\text{FAIR-BATCH}}(\mathcal{A}, Q, T) = o(T)$ is equivalent to ensuring that the total number of rounds with any α-fairness violation is sublinear. Therefore, our goal is equivalent to developing an algorithm $\mathcal{A}$ so that for any $(z^t)_{t=1}^T$,

$$\mathbf{Regret}^{\text{Err}}_{\text{FAIR-BATCH}}(\mathcal{A}, Q, T) = o(T)$$
$$\sum_{t=1}^T \mathrm{Unfair}_\alpha(\pi^t, z^t) = o(T).$$

To achieve the result above, we will reduce our setting to a setting with no fairness constraint, which we call *online batch classification* problem. Similar to the online fair batch classification setting, in each round t, the learner deploys a policy π^t, but the environment chooses only a batch of instances $(x^t_\tau, y^t_\tau)_{\tau=1}^k$. In online batch classification, we denote the strategy that the environment can take with $\mathcal{Z}_{\text{BATCH}} = \mathcal{X}^k \times \mathcal{Y}^k$. We compare the two settings in figure 1.

3. Results

In this section, we state the main results of our paper.

First, we show that one can reduce our problem of achieving no regret with respect to misclassification and fairness loss simultaneously to online batch classification problem.

Theorem 3.1. *For any sequence of $(z^t)_{t=1}^T \in \mathcal{Z}^T_{\text{FAIR-BATCH}}$ and $Q \subseteq \Delta\mathcal{H}$,*

$$\sum_{t=1}^T \mathcal{L}_{C,\alpha}(\pi^t, z^t) - \sum_{t=1}^T \mathcal{L}_{C,\alpha}(\pi^*, z^t) \leq \mathbf{Regret}^{Err}_{\text{BATCH}}(\mathcal{A}, Q, T),$$

where $\pi^t = \mathcal{A}_{\text{BATCH}}\left((\pi^j, \bar{x}'^j, \bar{y}'^j)_{j=1}^{t-1}\right)$. In other words,

$$\mathbf{Regret}^{C,\alpha,\mathcal{J}_{\alpha+\epsilon}}(\mathcal{A}, Q_\alpha, T) \leq \mathbf{Regret}^{Err}_{\text{BATCH}}(\mathcal{A}, Q, T).$$

Then, we show that by leveraging the Follow-the-Perturbed-Leader approach from (25), there exists an oracle-efficient algorithm to achieve no regret for both misclassification and fairness loss simultaneously.

Theorem 3.2. *If the separator set S for $\mathcal{H}$ is of size s, then* CONTEXT-FTPL *achieves the following misclassification and fairness regret in the online fair batch classification setting.*

$$\mathbf{Regret}^{Err}_{\text{FAIR-BATCH}}(\mathcal{A}, Q_\alpha, T) \leq O\left(\left(\frac{sk}{\epsilon}\right)^{\frac{3}{4}} \sqrt{T \log(|\mathcal{H}|)}\right)$$

$$\sum_{t=1}^T \mathit{Unfair}_{\alpha+\epsilon}(\pi^t, z^t) \leq O\left(\left(\frac{sk}{\epsilon}\right)^{\frac{3}{4}} \sqrt{T \log(|\mathcal{H}|)}\right)$$

Finally, we further show that in the stochastic setting, that is $\{\{(x^t_\tau, y^t_\tau)\}_{\tau=1}^k\}_{t=1}^T \sim_{i.i.d.} \mathcal{D}^{Tk}$, we get generalization for both misclassification and fairness loss.

Theorem 3.3 (Accuracy Generalization)**.** *With probabilty $1 - \delta$, the misclassification loss of π^{avg} is upper bounded by*

$$\begin{aligned}
&\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(\pi^{avg}(x), y)] \\
&\leq \inf_{\pi \in Q_\alpha} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(\pi(x), y)] \\
&+ \frac{1}{kT}\mathbf{Regret}^{C,\alpha,\mathcal{J}_{\alpha+\epsilon}}(\mathcal{A}, Q_\alpha, T) + \sqrt{\frac{8\ln\left(\frac{4}{\delta}\right)}{T}}
\end{aligned}$$

Theorem 3.4 (Fairness Generalization). *Assume that for all t, π^t is (α, β^t)-fair $(0 \leq \beta^t \leq 1)$. With probability $1 - \delta$, for any integer $q \leq T$, π^{avg} is $(\alpha' + \frac{q}{T}, \beta^*)$-fair where*

$$\beta^* = \frac{1}{q}\left(\mathbf{Regret}^{C,\alpha,\mathcal{J}_{\alpha+\epsilon}}(\mathcal{A}, Q_\alpha, T) + \sqrt{2T\ln\left(\frac{2}{\delta}\right)}\right).$$

Acknowledgements

We thank Sampath Kannan, Akshay Krishnamurthy, Katrina Ligett, and Aaron Roth for helpful conversations at an early stage of this work. Part of this work was done while YB, CJ, and ZSW were visiting the Simons Institute for the Theory of Computing. YB is supported in part by Israel Science Foundation (ISF) grant #1044/16, the United States Air Force and DARPA under contracts FA8750-16-C-0022 and FA8750-19-2-0222, and the Federmann Cyber Security Center in conjunction with the Israel national cyber directorate. CJ is supported in part by NSF grant AF-1763307. ZSW is supported in part by the NSF FAI Award #1939606, an Amazon Research Award, a Google Faculty Research Award, a J.P. Morgan Faculty Award, a Facebook Research Award, and a Mozilla Research Grant. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Air Force and DARPA.

References

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna M. Wallach. A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 60–69, 2018.

[2] Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(1):121–164, 2012.

[3] Nicolò Cesa-Bianchi, Yoav Freund, David Haussler, David P. Helmbold, Robert E. Schapire, and Manfred K. Warmuth. How to use expert advice. *J. ACM*, 44(3):427–485, May 1997.

[4] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, Special Issue on Social and Technical Trade-Offs, 2017.

[5] Alexandra Chouldechova and Aaron Roth. The frontiers of fairness in machine learning. *CoRR*, abs/1810.08810, 2018.

[6] Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv, 2018.

[7] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*, pages 214–226, 2012.

[8] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1):119–139, 1997.

[9] Stephen Gillen, Christopher Jung, Michael J. Kearns, and Aaron Roth. Online learning with an unknown fairness metric. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pages 2605–2614, 2018.

[10] Swati Gupta and Vijay Kamble. Individual fairness in hindsight. In *Proceedings of the 2019 ACM Conference on Economics and Computation, EC 2019, Phoenix, AZ, USA, June 24-28, 2019*, pages 805–806, 2019.

[11] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Neural Information Processing Systems (NIPS)*, 2016.

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

[13] Christina Ilvento. Metric learning for individual fairness. *CoRR*, abs/1906.00250, 2019.

[14] Shahin Jabbari, Matthew Joseph, Michael J. Kearns, Jamie Morgenstern, and Aaron Roth. Fairness in reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 1617–1626, 2017.

[15] Matthew Joseph, Michael J. Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. Meritocratic fairness for infinite and contextual bandits. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2018, New Orleans, LA, USA, February 02-03, 2018*, pages 158–163, 2018.

[16] Matthew Joseph, Michael J. Kearns, Jamie H. Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 325–333, 2016.

[17] Christopher Jung, Michael J. Kearns, Seth Neel, Aaron Roth, Logan Stapleton, and Zhiwei Steven Wu. Eliciting and enforcing subjective individual fairness. *CoRR*, abs/1905.10660, 2019.

[18] Sampath Kannan, Michael J. Kearns, Jamie Morgenstern, Mallesh M. Pai, Aaron Roth, Rakesh V. Vohra, and Zhiwei Steven Wu. Fairness incentives for myopic agents. In *Proceedings of the 2017 ACM Conference on Economics and Computation, EC '17, Cambridge, MA, USA, June 26-30, 2017*, pages 369–386, 2017.

[19] Michael J. Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 2569–2577, 2018.

[20] Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Fairness through computationally-bounded awareness. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 4847–4857, 2018.

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

[22] John Podesta, Penny Pritzker, Ernest J. Moniz, John Holdren, and Jefrey Zients. Big data: Seizing opportunities and preserving values. 2014.

[23] Guy N. Rothblum and Gal Yona. Probably approximately metric-fair learning. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 5666–5674, 2018.

[24] Arun Sai Suggala and Praneeth Netrapalli. Online non-convex learning: Following the perturbed leader is optimal. *CoRR*, abs/1903.08110, 2019.

[25] Vasilis Syrgkanis, Akshay Krishnamurthy, and Robert E. Schapire. Efficient algorithms for adversarial contextual learning. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 2159–2168, 2016.