Probably Approximately Metric-Fair Learning

Guy N. Rothblum [*1] **Gal Yona** [*1]

Abstract

The seminal work of Dwork *et al.* [ITCS 2012] introduced a metric-based notion of individual fairness: given a task-specific similarity metric, their notion required that every pair of similar individuals should be treated similarly. In the context of machine learning, however, individual fairness does not generalize from a training set to the underlying population. We show that this can lead to computational intractability even for simple fair-learning tasks.

With this motivation in mind, we introduce and study a relaxed notion of *approximate metric-fairness*: for a random pair of individuals sampled from the population, with all but a small probability of error, if they are similar then they should be treated similarly. We formalize the goal of achieving approximate metric-fairness simultaneously with best-possible accuracy as Probably Approximately Correct and Fair (PACF) Learning. We show that approximate metric-fairness *does* generalize, and leverage these generalization guarantees to construct polynomial-time PACF learning algorithms for the classes of linear and logistic predictors.

1. Introduction

Machine learning is increasingly used to make consequential classification decisions about individuals. Examples range from predicting whether a user will enjoy a particular article, to estimating a felon's recidivism risk, to determining whether a patient is a good candidate for a medical treatment. Automated classification comes with great benefits, but it also raises substantial societal concerns (cf. (O'Neil, 2016) for a recent perspective). One prominent concern is that these algorithms might discriminate against individuals or groups in a way that violates laws or social and ethical norms. This might happen due to biases in the training data or due to biases introduced by the algorithm. To address these concerns, and to truly unleash the full potential of automated classification, there is a growing need for frameworks and tools to mitigate the risks of algorithmic discrimination. A growing literature attempts to tackle these challenges by exploring different fairness criteria.

Discrimination can take many guises. It can be difficult to spot and difficult to define. Imagine a protected minority population P (defined by race, gender identity, etc). A natural approach for protecting the members of P from discrimination is to make sure that they are not mistreated *on average*. For example, that on average members of P and individuals outside of P are classified in any particular way with roughly the same probability. This is a *"group-level"* fairness notion, sometimes referred to as *statistical parity*.

Pointing out several weakness of group-level notions of fairness, the seminal work of (Dwork et al., 2012) introduced a notion of *individual fairness*. Their notion relies on a *task-specific similarity metric* that specifies, for every two individuals, how similar they are with respect to the specific classification task at hand. Given such a metric, similar individuals should be treated similarly, i.e. assigned similar classification distributions (their focus was on probabilistic classifiers, as will be ours). In this work, we refer to their fairness notion as *perfect metric-fairness*.

Given a good metric, perfect metric-fairness provides powerful protections from discrimination. Furthermore, the metric provides a vehicle for specifying social norms, cultural awareness, and task-specific knowledge. While coming up with a good metric can be challenging, metrics arise naturally in prominent existing examples (such as credit scores and insurance risk scores), and in natural scenarios (a metric specified by an external regulator). Dwork *et al.* studied the goal of finding a (probabilistic) classifier that minimizes utility loss (or maximizes accuracy), subject to satisfying the perfect metric-fairness constraint. They showed how to phrase and solve this optimization problem for a given collection of individuals.

1.1. This Work

Building on these foundations, we study *metric-fair machine learning*. Consider a learner that is given a similarity metric and a training set of labeled examples, drawn from an underlying population distribution. The learner should output a *fair* classifier that (to the extent possible) accurately

[*]Equal contribution [1]Weizmann Institute of Science, Rehovot, Israel. Correspondence to: Guy N. Rothblum <rothblum@alum.mit.edu>, Gal Yona <gal.yona@weizmann.ac.il>.

classifies the underlying population.

This goal departs from the scenario studied in (Dwork et al., 2012), where the focus was on guaranteeing metric-fairness and utility for the dataset at hand. *Generalization* of the fairness guarantee is a key difference: we focus on guaranteeing fairness not just for the (training) data set at hand, but also for the underlying population from which it was drawn. We note that perfect metric-fairness does not, as a rule, generalize from a training set to the underlying population. This presents computational difficulties for constructing learning algorithms that are perfectly metric-fair for the underlying population. Indeed, we exhibit a simple learning task that, while easy to learn without fairness constraints, becomes computationally infeasible under the perfect metric-fairness constraint (given a particular metric).

We develop a relaxed *approximate metric-fairness* framework for machine learning, where fairness does generalize from the sample to the underlying population, and present polynomial-time fair learning algorithms in this framework.

1.2. Related Work

There is a growing body of work attempting to study the question of algorithmic discrimination. This literature is characterized by an abundance of definitions, each capturing different discrimination concerns and notions of fairness. One high-level distinction can be drawn between *group* and *individual* notions of fairness.

Group-fairness notions assume the existence of a protected attribute (e.g gender, race), which induces a partition of the instance space into some small number of groups. A fair classifier is one that achieves parity of some statistical measure across these groups. Some prominent measures include classification rates (statistical parity, see e.g (Feldman et al., 2015)), calibration, and false positive or negative rates (Kleinberg et al., 2016; Chouldechova, 2017; Hardt et al., 2016). It has been established that some of these notions are inherently incompatible with each other, in all but trivial cases (Kleinberg et al., 2016; Chouldechova, 2017).

Individual fairness (Dwork et al., 2012) posits that "similar individuals should be treated similarly". This powerful guarantee is formalized via a Lipschitz condition (with respect to an existing task-specific similarity metric) on the classifier mapping individuals to distributions over outcomes. Recent works (Joseph et al., 2016; Joseph et al.) study different individual-level fairness guarantees in the contexts of reinforcement and online learning.

Our notion of approximate metric-fairness can be interpreted as staking a middle-ground between individual- and group-fairness. In this sense, it is similar to recent works that protect large collections of sufficiently-large groups (Hébert-Johnson et al., 2017; Kearns et al., 2017; Kim et al., 2018). A distinction from these works is in protecting *every* sufficiently-large group, rather than a large collection of groups that is fixed a priori. (Kim et al., 2018) consider a (computational) relaxation of individual fairness, focusing on settings where the metric itself is not fully known.

Finally, several works have studied fair regression (Kamishima et al., 2012; Calders et al., 2013; Zafar et al., 2017; Berk et al., 2017). The main differences in our work are a focus on metric-based fairness, a strong rigorous fairness guarantee, and proofs of competitive accuracy (both stated with respect to the underlying distribution).

2. Approximate Metric-Fairness

Taking inspiration from Valiant's celebrated PAC learning model (Valiant, 1984), we allow a small fairness error and a small probability of a complete fairness failure. We require that with all but α probability over a choice of two individuals from the underlying distribution, if the two individuals are similar then they get similar classification distributions, up to an additive slack γ in the similarity measure. We refer to this condition as *approximate metric-fairness (MF)*. We think of $\alpha, \gamma \in [0, 1)$ as small constants, and note that setting $\alpha = \gamma = 0$ recovers the definition of *perfect* metric-fairness (thus, setting α, γ to be a small constants larger than 0 is indeed a relaxation).

Definition 2.1 *A predictor h is (α, γ) approximately metric-fair (MF) with respect to a similarity metric d and a data distribution $\mathcal{D}$ if:*

$$\Pr_{x,x' \sim \mathcal{D}}[|h(x) - h(x')| > d(x, x') + \gamma] \leq \alpha \qquad (1)$$

The *MF loss* of the classifier h on the pair (x, x') is 1 if the (internal) inequality in Equation (1) holds, and 0 otherwise (hence, we refer to this as a $0/1$ fairness loss). A *classifier* is (α, γ)-approximately MF if with all but α probability over two individuals (x, x') sampled from $\mathcal{D}$, the MF loss is 0.

Similarly to PAC learning, we also allow a small δ probability of failure. This probability is taken over the choice of the training set and over the learner's coins. For example, δ bounds the probability that the randomly sampled training set is not representative of the underlying population. We think of δ as very small or even negligible. A learning algorithm is *probably approximately metric-fair* if with all but δ probability over the sample (and the learner's coins), it outputs a classifier that is (α, γ)-approximately MF.

Given a well-designed metric, approximate metric-fairness (for sufficiently small α, γ) guarantees that almost every individual gets fair treatment compared to almost every other individual. *Every* group P of fractional size significantly larger than α is protected in the sense that, on average, members of P are treated similarly to similar individuals

outside of P. We note, however, that this guarantee does not protect single individuals or small groups.

3. Accurate and Fair Learning

Our goal is obtaining learning algorithms that are probably approximately metric-fair, and that simultaneously guarantee non-trivial accuracy. Recall that fairness, on its own, can always be obtained by outputting a constant classifier that ignores its input and treats all individuals identically. It is the combination of the fairness and the accuracy objectives that makes for an interesting task. We follow (Dwork et al., 2012) in focusing on finding a classifier that maximizes accuracy, subject to the approximate metric-fairness constraint. This is a natural formulation, as we think of fairness as a hard requirement (imposed, for example, by a regulator), and thus fairness cannot be traded off for better accuracy.

Problem Statement. A *learning problem* is defined by an instance domain $\mathcal{X}$ and a class $\mathcal{H}$ of predictors (probabilistic classifiers) $h : \mathcal{X} \to [0, 1]$, where we interpret $h(x)$ as the probability that the label is 1. A *fair* learning problem also includes a similarity metric $d : \mathcal{X}^2 \to [0, 1]$. The learning algorithm gets as input the metric d and a sample of labeled examples, drawn i.i.d. from a distribution $\mathcal{D}$ over labeled examples from $(\mathcal{X} \times \pm 1)$, and its goal is to output a predictor that is both fair and as accurate as possible. For a learned (real-valued) predictor h, we use $err_D(h)$ to denote the expected ℓ_1 error of h (the absolute loss) on a random sample from $\mathcal{D}$.[1]

Accuracy guarantee. As discussed above, the goal in metric-fair and accurate learning is optimizing the predictor's accuracy subject to the fairness constraint. Ideally, we aim to approach (as the sample size grows) the error rate of the most accurate classifier that satisfies the fairness constraints. A more relaxed benchmark is guaranteeing (α, γ)-approximate metric-fairness, while approaching the accuracy of the best classifier that is (α', γ')-approximately metric-fair, for $\alpha' \in [0, \alpha]$ and $\gamma' \in [0, \gamma]$. Our efficient learning algorithms will achieve this more relaxed accuracy goal (see below). We note that even relaxed competitiveness means that the classifier is (at the very least) competitive with the best *perfectly* metric-fair classifier.

These goals are captured in the following definition of *probability approximately correct and fair (PACF) learning*. Crucially, both fairness and accuracy goals are stated with respect to the (unknown) underlying distribution.

Definition 3.1 (PACF Learning) *A learning algorithm $\mathcal{A}$ PACF-learns a hypothesis class $\mathcal{H}$ if for every metric d and population distribution $\mathcal{D}$, every required fairness parameters $\alpha, \gamma \in [0, 1)$, every failure probability $\delta \in (0, 1)$, and every error parameters $\epsilon, \epsilon_\alpha, \epsilon_\gamma \in (0, 1)$, there exists a sample complexity $m = \text{poly}\left(\frac{\log|\mathcal{X}|\cdot\log(1/\delta)}{\alpha\cdot\gamma\cdot\epsilon\cdot\epsilon_\alpha\cdot\epsilon_\gamma}\right)$ and constants $\alpha', \gamma' \in [0, 1)$ (specified below), such that with all but δ probability over an i.i.d. sample of size m and $\mathcal{A}$'s coin tosses, the output predictor h satisfies the following two conditions:*

1. **Fairness**: *h is (α, γ)-approximately metric-fair w.r.t. the metric d and the distribution $\mathcal{D}$.*
2. **Accuracy**: *Let $\mathcal{H}'_F$ be the subclass of hypotheses in $\mathcal{H}$ that are $(\alpha' - \epsilon_\alpha, \gamma' - \epsilon_\gamma)$-approximately MF, then:*

$$err_D(h) \leq \min_{h' \in \mathcal{H}'_F} err_D(h') + \epsilon$$

We say that $\mathcal{A}$ is efficient *if it runs in time* $\text{poly}(m)$. *If accuracy holds for $\alpha' = \alpha$ and $\gamma' = \gamma$, then we say that $\mathcal{A}$ is a* strong *PACF learning algorithm. Otherwise, we say that $\mathcal{A}$ is a* relaxed *PACF learning algorithm.*

Note that the accuracy guarantee is *agnostic*: we make no assumptions about the way the training labels are generated. Agnostic learning is particularly well suited to our setting: since we make no assumptions about the metric d, even if the labels are generated by $h \in \mathcal{H}$, it might be the case that d does not allow for accurate predictions, in which case a fair learner cannot compete with h's accuracy.

4. Generalization

Generalization is a key issue in learning theory. We develop strong generalization bounds for approximate MF, showing that guaranteeing *empirical* approximate MF on a training set also guarantees approximate MF on the underlying distribution (w.h.p. over the choice of sample S). These bounds open the door to polynomial-time algorithms that can focus on guaranteeing fairness (and accuracy) on the sample and effectively rules out the possibility of creating a "false facade" of fairness (i.e, a classifier that appears fair on a random sample, but is not fair w.r.t new individuals).

Towards proving generalization, we define the empirical fairness loss on a sample S (a training set). Fixing a fairness parameter γ, a predictor h and a pair of individuals x, x' in the training set, consider the MF loss on the "edge" between x and x' (recall that the MF loss is 1 if the "internal" inequality of Equation (1) holds, and 0 otherwise). Observe that the losses on the $\binom{|S|}{2}$ edges are not independent random variables (over the choice of S), because each individual $x \in S$ affects many edges. Thus, rather than count the empirical MF loss over all edges, we restrict ourselves to a "matching" $M(S)$ in the complete graph whose vertices are S: a collection of edges, where each individual is involved in exactly one edge. The empirical MF loss of h on S is defined as the average MF loss over edges in $M(S)$.

[1] All results also translate to ℓ_2 error (the squared loss).

Theorem 4.1 *Let $\mathcal{H}$ be a hypothesis class with Rademacher complexity $R_m(\mathcal{H}) = (r/\sqrt{m})$. For every $\delta \in (0,1)$ and every $\epsilon_\alpha, \epsilon_\gamma \in (0,1)$, there exists a sample complexity $m = O\left(\frac{r^2 \cdot \ln(1/\delta)}{\epsilon_\alpha^2 \cdot \epsilon_\gamma^2}\right)$, such that with probability at least $1-\delta$ over an i.i.d sample $S \sim \mathcal{D}^m$, simultaneously for every $h \in \mathcal{H}$: if h is (α, γ)-approximately metric-fair on the sample S, then h is also $(\alpha + \epsilon_\alpha, \gamma + \epsilon_\gamma)$-approximately metric-fair on the underlying distribution $\mathcal{D}$.*

4.1. Information-Theoretic Sample Complexity

The fairness-generalization result of Theorem 4.1 implies that, from a sample-complexity perspective, any hypothesis class is strongly PACF learnable, with sample complexity comparable to that of standard PAC learning.

Theorem 4.2 *Let $\mathcal{H}$ be a hypothesis class with Rademacher complexity $R_m(\mathcal{H}) = (r/\sqrt{m})$. Then $\mathcal{H}$ is information-theoretically strongly PACF learnable with sample complexity $m = O\left(\frac{r^2 \ln(1/\delta)}{(\epsilon')^2}\right)$, for $\epsilon' = \min\{\epsilon, \epsilon_\alpha, \epsilon_\gamma\}$.*

5. Efficient Fair Learning

One of our primary contributions is the construction of polynomial-time relaxed-PACF learning algorithms for expressive hypothesis classes. We focus on linear classification tasks, where the labels are determined by a separating hyperplane. Learning linear classifiers is a central tool in machine learning. By embedding a learning problem into a higher-dimensional space, linear classifiers (over the expanded space) can capture surprisingly strong classes, such as polynomial threshold functions (see, for example, the discussion in (Hellerstein & Servedio, 2007)). The "kernel trick" (see, e.g, (Shalev-Shwartz & Ben-David, 2014)) can allow for efficient solutions even over very high (or infinite) dimensional embeddings. Many of the known (distribution-free) PAC learning algorithms can be derived by learning linear threshold functions (Hellerstein & Servedio, 2007).

5.1. Linear Regression

Linear regression, the task of learning linear predictors, is an important and well-studied problem in the machine learning literature. In terms of accuracy, this is an appealing class when we expect a linear relationship between the probability of the label being 1 and the distance from a hyperplane. Taking the domain $\mathcal{X}$ to be the unit ball, we define the class of linear predictors as:

$$H_{lin} \stackrel{\text{def}}{=} \{\mathbf{x} \mapsto (1 + \langle \mathbf{w}, \mathbf{x} \rangle)/2 : \ \|\mathbf{w}\| \leq 1\}.$$

We show a relaxed PACF learning algorithm for H_{lin}:

Theorem 5.1 *H_{lin} is relaxed PACF learnable with sample and time complexities of poly$(\frac{1}{\epsilon_\gamma}, \frac{1}{\epsilon_\alpha}, \frac{1}{\epsilon}, \log\frac{1}{\delta})$. For every $\gamma' \in [0,1)$ and $\alpha' = (\alpha \cdot \gamma - \gamma')$, the accuracy of the learned predictor approaches (or beats) the most accurate (α', γ')-approximately MF predictor.*

5.2. Logistic Regression

Logistic regression is another appealing class. Here, the prediction need not a be a linear function of the distance from a hyperplane. Rather, we allow the use of a sigmoid function $\phi_\ell : [-1,1] \to [0,1]$ defined as $\phi_\ell(z) = \frac{1}{1+\exp(-4\ell \cdot z)}$ (which is continuous and ℓ-Lipschitz). The sigmoidal transfer function gives the predictor the power to exhibit sharper transitions from low predictions to high predictions around a certain distance threshold. The class of logistic predictors is formed by composing a linear function with a sigmoidal transfer function:

$$H_{\phi,L} \stackrel{\text{def}}{=} \{\mathbf{x} \mapsto \phi_\ell(\langle w, \mathbf{x} \rangle) : \ \|\mathbf{w}\| \leq 1, \ \ell \in [0, L]\} \quad (2)$$

Our primary technical contribution is a polynomial-time relaxed PACF learner for $H_{\phi,L}$ where L is constant.

Theorem 5.2 *For every constant $L > 0$, $H_{\phi,L}$ is relaxed PACF learnable with sample and time complexities of poly$(\frac{1}{\epsilon_\gamma}, \frac{1}{\epsilon_\alpha}, \frac{1}{\epsilon}, \log\frac{1}{\delta})$. For every $\gamma' \in [0,1)$ and $\alpha' = (\alpha \cdot \gamma - \gamma')$, the learned predictor's accuracy approaches the best (α', γ')-approximately MF predictor.*

6. Hardness of Perfect Metric-Fairness

As discussed above, perfect metric-fairness *does not generalize* from a training set to the underlying population. For example, consider a small subset of the population that isn't represented in the training set. A classifier that discriminates against this small subset might be perfectly metric-fair *on the training set*. The failure of generalization poses serious challenges to constructing learning algorithms. Indeed, we show that perfect MF can make simple learning tasks computationally intractable (with respect to a particular metric).

Under mild cryptographic assumptions (specifically, that one-way functions exist), we exhibit a learning problem and a similarity metric where: (i) there exists a *perfectly fair and perfectly accurate* simple (linear) predictor, but (ii) any polynomial-time perfectly metric-fair learner can only find a trivial predictor, whose error approaches 1/2. In contrast, (iii) there *does* exist a polynomial-time (relaxed) PACF learning algorithm for this task. This is an important motivation for our study of *approximate* MF.

References

Berk, Richard, Heidari, Hoda, Jabbari, Shahin, Joseph, Matthew, Kearns, Michael, Morgenstern, Jamie, Neel, Seth, and Roth, Aaron. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*, 2017.

Calders, Toon, Karim, Asim, Kamiran, Faisal, Ali, Wasif, and Zhang, Xiangliang. Controlling attribute effect in linear regression. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, pp. 71–80. IEEE, 2013.

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

Dwork, Cynthia, Hardt, Moritz, Pitassi, Toniann, Reingold, Omer, and Zemel, Richard S. Fairness through awareness. In *Innovations in Theoretical Computer Science (ITCS)*, pp. 214–226, 2012.

Feldman, Michael, Friedler, Sorelle A, Moeller, John, Scheidegger, Carlos, and Venkatasubramanian, Suresh. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 259–268. ACM, 2015.

Hardt, Moritz, Price, Eric, and Srebro, Nathan. Equality of opportunity in supervised learning. *CoRR*, abs/1610.02413, 2016. URL http://arxiv.org/abs/1610.02413.

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

Hellerstein, Lisa and Servedio, Rocco A. On PAC learning algorithms for rich boolean function classes. *Theor. Comput. Sci.*, 384(1):66–76, 2007. doi: 10.1016/j.tcs.2007.05.018. URL https://doi.org/10.1016/j.tcs.2007.05.018.

Joseph, Matthew, Kearns, Michael, Morgenstern, Jamie, Neel, Seth, and Roth, Aaron. Better fair algorithms for contextual bandits.

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.

Kamishima, Toshihiro, Akaho, Shotaro, Asoh, Hideki, and Sakuma, Jun. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 35–50. Springer, 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.

Kim, Michael P., Reingold, Omer, and Rothblum, Guy N. Fairness through computationally-bounded awareness. *CoRR*, abs/1803.03239, 2018. URL http://arxiv.org/abs/1803.03239.

Kleinberg, Jon, Mullainathan, Sendhil, and Raghavan, Manish. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

O'Neil, Cathy. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown, New York, 2016.

Shalev-Shwartz, Shai and Ben-David, Shai. Understanding machine learning: From theory to algorithms. chapter 16, pp. 215–226. Cambridge university press, 2014.

Valiant, Leslie G. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, 1984. doi: 10.1145/1968.1972. URL http://doi.acm.org/10.1145/1968.1972.

Zafar, Muhammad Bilal, Valera, Isabel, Gomez Rodriguez, Manuel, and Gummadi, Krishna P. Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259*, 2017.