Fairness Through Computationally-Bounded Awareness

Michael P. Kim [1] **Omer Reingold** [1] **Guy N. Rothblum** [2]

Abstract

We study the problem of fair classification within the versatile framework of Dwork *et al.* [DHP$^+$12], which assumes the existence of a metric that measures similarity between pairs of individuals. Unlike earlier work, we do not assume that the entire metric is known to the learning algorithm; instead, the learner can query this *arbitrary* metric a bounded number of times. We propose a new notion of fairness called *metric multifairness* and show how to achieve this notion in our setting. Metric multifairness is parameterized by a similarity metric δ on pairs of individuals to classify and a rich collection $\mathcal{C}$ of (possibly overlapping) "comparison sets" over pairs of individuals. At a high level, metric multifairness guarantees that *similar subpopulations are treated similarly*, as long as these subpopulations are identified within the class $\mathcal{C}$.

1. Introduction

More and more, machine learning systems are being used to make predictions about people of significant personal consequence; for instance, *Is this person likely to repay a loan?* [Wad16] or *Is this person likely to recommit a crime?* [ALMK16]. As these classification systems have become more ubiquitous, concerns have also grown that classifiers obtained via machine learning might discriminate based on sensitive attributed like race, gender, or sexual orientation. Indeed, machine-learned classifiers run the risk of perpetuating or amplifying historical biases present in the training data. Examples of discrimination in classification have been well-illustrated [Wad16, ALMK16, CDPF$^+$17, KMR17, HPS16, BG18]; nevertheless, developing a systematic approach to fairness has been challenging. Often, it feels that the objectives of fair classification are at odds with obtaining high-utility predictions.

In an influential work, Dwork *et al.* [DHP$^+$12] proposed a framework to resolve the apparent conflict between utility and fairness, which they call "fairness through awareness." This framework takes the perspective that a fair classifier should *treat similar individuals similarly*. The work formalizes this abstract goal by assuming access to a task-specific similarity metric δ on pairs of individuals. The proposed notion of fairness requires that if the distance between two individuals is small, then the predictions of a fair classifier cannot be very different. More formally, for some small constant $\tau \geq 0$, we say a hypothesis $f : \mathcal{X} \to [-1, 1]$ satisfies (δ, τ)-*metric fairness*[1] if the following (approximate) Lipschitz condition holds for all pairs of individuals from the population $\mathcal{X}$.

$$\forall x, x' \in \mathcal{X} \times \mathcal{X} : \quad |f(x) - f(x')| \leq \delta(x, x') + \tau \qquad (1)$$

Subject to these intuitive similarity constraints, the classifier may be chosen to maximize utility. Note that, in general, the metric may be designed externally (say, by a regulatory agency) to address legal and ethical concerns, independent from the task of learning. In particular, in certain settings, the metric designers may have access to a different set of features than the learner. For instance, perhaps the metric designers have access to sensitive attributes, but for legal, social, or pragmatic reasons, the learner does not. In addition to its conceptual simplicity, the modularity of fairness through awareness makes it a very appealing framework. Currently, there are many (sometimes contradictory) notions of what it means for a classifier to be fair [KMR17, CDPF$^+$17, FPCG16, HPS16, HKRR17], and there is much debate on which definitions should be applied in a given context. Discrimination comes in many forms and classification is used in a variety of settings, so naturally, it is hard to imagine any universally-applicable definition of fairness. Basing fairness on a similarity metric offers a flexible approach for formalizing a variety of guarantees and protections from discrimination.

Still, a challenging aspect of this approach is the assumption that the similarity metric is known for all pairs of individuals.[2] Deciding on an appropriate metric is itself a

[1] Computer Science Department, Stanford University, Stanford, CA. Supported by NSF Grant CCF-1763299. [2] Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel. Supported by ISF grant No. 5219/17.. Correspondence to: Michael P. Kim <mpk@cs.stanford.edu>.

[1] Note the definition given in [DHP$^+$12] is slightly different; they propose a more general Lipschitz condition, but fix $\tau = 0$.

[2] Indeed, [DHP$^+$12] identifies this assumption as "one of the most challenging aspects" of the framework.

delicate matter and could require human input from sociologists, legal scholars, and specialists with domain expertise. For instance, in the loan repayment example, a simple, seemingly-objective metric might be a comparison of credit scores. A potential concern, however, is that these scores might themselves be biased (i.e. encode historical discriminations). In this case, a more nuanced metric requiring human input may be necessary. Further, if the metric depends on features that are latent to the learner (e.g. some missing sensitive feature) then the metric could appear *arbitrarily complex* to the learner. As such, in many realistic settings, the resulting metric will not be a simple function of the learner's feature vectors of individuals.

In most machine learning applications, where the universe of individuals is assumed to be very large, even writing down an appropriate metric could be completely infeasible. In these cases, rather than require the metric value to be specified for all pairs of individuals, we could instead ask a panel of experts to provide similarity scores for a *small sample* of pairs of individuals. While it is information-theoretically impossible to guarantee metric fairness from a sampling-based approach, we still might hope to provide a strong, provable notion of fairness that maintains the theoretical appeal and practical modularity of the fairness through awareness framework.

1.1. Fairness through computationally-bounded awareness

In this work, we propose a new theoretical framework for fair classification based on fairness through awareness – which we dub "fairness through computationally-bounded awareness" – that eliminates the considerable issue of requiring the metric to be known exactly. Our approach maintains the simplicity and flexibility of fairness through awareness, but provably only requires a small number of random samples from the underlying metric, even though we make no structural assumptions about the metric. In particular, our approach works even if the metric provably cannot be learned. Specifically, our notion will require that a fair classifier *treat similar subpopulations of individuals similarly*, in a sense we will make formal next. While our definition relaxes fairness through awareness, we argue that it still protects against important forms of discrimination that the original work aimed to combat; further, we show that stronger notions necessarily require a larger sample complexity from the metric. As in [DHP+12], we investigate how to learn a classifier that achieves optimal utility under similarity-based fairness constraints, assuming a weaker model of limited access to the metric. We give positive and negative results that show connections between achieving our fairness notion and learning.

Metric multifairness We define our relaxation of metric fairness with respect to a rich class of statistical tests on the pairs of individuals. Let a *comparison* be any subset of the pairs of $\mathcal{X} \times \mathcal{X}$. Our definition, which we call *metric multifairness*, is parameterized by a collection of comparisons $\mathcal{C} \subseteq 2^{\mathcal{X}\times\mathcal{X}}$ and requires that a hypothesis appear Lipschitz according to all of the statistical tests defined by the comparisons $S \in \mathcal{C}$. More formally, for some small constant $\tau > 0$, we require the following condition holds for all comparisons $S \in \mathcal{C}$ in the collection:

$$\mathop{\mathbf{E}}_{(x,x')\sim S}\left[\,|f(x) - f(x')|\,\right] \leq \mathop{\mathbf{E}}_{(x,x')\sim S}\left[\delta(x,x')\right] + \tau \quad (2)$$

where $(x, x') \sim S$ is a random draw from the data distribution $\mathcal{D}$, conditioned on $(x, x') \in S$.

To begin, note that metric multifairness is indeed a relaxation of metric fairness; if we take the collection $\mathcal{C} = \{\{(x,x')\} : x, x' \in \mathcal{X} \times \mathcal{X}\}$ to be the collection of all pairwise comparisons, then $(\mathcal{C}, \delta, \tau)$-metric multifairness is equivalent to (δ, τ)-metric fairness.

In order to achieve metric multifairness from a small sample from the metric, however, we need a lower bound on the density of each comparison in $\mathcal{C}$; in particular, we can't hope to enforce metric fairness from a small sample. For some $\gamma > 0$, we say that a collection of comparisons $\mathcal{C}$ is γ-*large* if for all $S \in \mathcal{C}$, $\Pr_{(x,x')\sim\mathcal{D}}[(x,x') \in S] \geq \gamma$. A natural next choice for $\mathcal{C}$ would be a collection of comparisons that represent the Cartesian products between traditionally-protected groups, defined by race, gender, etc. In this case, as long as the minority populations are not too small, then a random sample from the metric will give accurate empirical statistics, and we can enforce this statistical relaxation of metric fairness. While this approach is information-theoretically feasible, its protections are very weak.

To highlight this weakness, suppose we want to predict the probability individuals will repay a loan, and our metric is an adjusted credit score. Even after adjusting scores, two populations $P, Q \subseteq \mathcal{X}$ (say, defined by race) may have large average distance because *overall* P has better credit than Q; still, within P and Q, there may be significant *subpopulations* $P' \subseteq P$ and $Q' \subseteq Q$ that should be treated similarly (possibly representing the qualified members of each group). In this case, a coarse statistical relaxation of metric fairness will not require that a classifier treat P' and Q' similarly; instead, the classifier could treat everyone in P better than everyone in Q – including treating *unqualified* members of P better than *qualified* members of Q. Indeed, the weaknesses of broad-strokes statistical definitions served as motivation for the original work of [DHP+12]. We would like to choose a class $\mathcal{C}$ that strengthens the fairness guarantees of metric multifairness, but maintains its efficient sample complexity.

Computationally-bounded awareness. While we can define metric multifairness with respect to any collection $\mathcal{C}$, typically, we will think of $\mathcal{C}$ as a rich class of overlapping subsets; equivalently, we can think of the collection $\mathcal{C}$ as an expressive class of boolean functions, where for $S \in \mathcal{C}$, $c_S(x, x') = 1$ if and only if $(x, x') \in S$. In particular, $\mathcal{C}$ should be much more expressive than simply defining comparisons across traditionally-protected groups. The motivation for choosing such an expressive class $\mathcal{C}$ is exemplified in the following proposition.

Proposition 1. *Suppose there is some $S \in \mathcal{C}$, such that $\mathbf{E}_{(x,x')\sim S}[\delta(x, x')] \le \varepsilon$. Then if f is $(\mathcal{C}, \delta, \tau)$-metric multifair, then f satisfies $(\delta, (\varepsilon + \tau)/p)$-metric fairness for at least a $(1 - p)$-fraction of the pairs in S.*

That is, if there is some subset $S \in \mathcal{C}$ that identifies a set of pairs whose metric distance is small, then any metric multifair hypothesis must also satisfy the stronger metric fairness notion on many pairs from S. This effect will compound if many different (possibly overlapping) comparisons are identified that have small average distance. We emphasize that these small-distance comparisons are not known before sampling from the metric; indeed, this would imply the metric was (approximately) known *a priori*. Still, if the class $\mathcal{C}$ is rich enough to correlate well with various comparisons that reveal significant information about the metric, then any metric multifair hypothesis will satisfy *individual-level* fairness on a significant fraction of the population!

While increasing the expressiveness of $\mathcal{C}$ increases the strength of the fairness guarantee, in order to learn from a small sample, we cannot choose $\mathcal{C}$ to be arbitrarily complex. Thus, in choosing $\mathcal{C}$ we must balance the strength of the fairness guarantee with the information bottleneck in accessing δ through random samples. Our resolution to these competing needs is complexity-theoretic: while we can't hope to ensure fair treatment across *all* subpopulations, we can hope ensure fair treatment across *efficiently-identifiable* subpopulations. For instance, if we take $\mathcal{C}$ to be a family defined according to some class of computations of bounded dimension – think, the set of conjunctions of a constant number of boolean features or short decision trees – then we can hope to accurately estimate and enforce the metric multifairness conditions. Taking such a bounded $\mathcal{C}$ ensures that a hypothesis will be fair on all comparisons identifiable within this computational bound. This is the sense in which metric multifairness provides fairness through *computationally-bounded* awareness.

2. Our Contributions

High-level setting. We will take $\mathcal{C}$ to be our collection of comparisons, τ to be the additive error in the multifairness definition, $\gamma \le \Pr[(x, x') \in S]$ be the minimum density of $S \in \mathcal{C}$, and ξ to be the failure probability.

Our results focus on two settings. In the first setting, we receive a set of N individuals, each with an associated "ideal" utility-maximizing prediction; our algorithmic task is to find adjusted predictions for each individual that satisfy metric multifairness. Of course, any set of constant predictions is always metric multifair but will have no predictive power; as such, subject to the fairness constraints, our objective is to optimize utility (according to some given loss function). Beyond the multifairness constraints, we place no modeling restrictions on the predictions we output. Our first result shows that there is an algorithm for outputting metric multifair predictions for a given set of individuals.

Theorem 2 (Informal)**.** *There is an algorithm that, given N labeled individuals, outputs a set of near-optimal metric multifair predictions with probability at least $1 - \xi$ over the draw of n metric samples provided $n \ge \tilde{\Omega}\left(\frac{\log(|\mathcal{C}|/\xi)}{\gamma\tau^2}\right)$, where γN^2 is the minimum cardinality of any comparison $S \in \mathcal{C}$. The algorithm runs in time $O(|\mathcal{C}| \cdot N^2 \cdot \mathrm{poly}(1/\gamma, 1/\tau))$.*

The predictions we learn achieve loss comparable to the best metric multifair hypotheses. The exact guarantee on the utility is a bit technical, but as one point of comparison, the utility achieved nearly matches or improves on the utility of the best hypothesis satisfying metric fairness of [DHP+12].

Theorem 2 shows that metric multifairness is attainable with respect to any metric δ and any collection of comparisons $\mathcal{C}$ from a small number of metric samples. Unfortunately, the running time depends linearly on $|\mathcal{C}|$; as we think of $\mathcal{C}$ as a large and expressive collection of comparisons, $|\mathcal{C}|$ may be prohibitively expensive. Restricting our attention to structured classes of $\mathcal{C}$, we improve the dependence on $|\mathcal{C}|$. Specifically, viewing the collection $\mathcal{C}$ as a boolean concept class, we show that if $\mathcal{C}$ admits an efficient agnostic learner, then we can learn a metric multifair hypothesis efficiently.

Theorem 3 (Informal)**.** *Suppose there is an algorithm $\mathcal{A}$ for agnostic learning $\mathcal{C}$ that achieves accuracy ε with probability $1 - \xi$ in time $\mathrm{poly}(1/\varepsilon, \log(1/\xi))$ from $\mathrm{poly}(1/\varepsilon, \log(1/\xi))$ labeled samples. Then, there is an algorithm that, given N labeled individuals, with probability at least $1 - \xi$ over n metric samples outputs a set of near-optimal metric multifair predictions that runs in time $O(N^2 \cdot \mathrm{poly}(1/\gamma, 1/\tau, \log(1/\xi)))$ and requires $n = \tilde{O}\left(\frac{\log(|\mathcal{C}|/\xi)}{\gamma\tau^2} + \mathrm{poly}(1/\gamma, 1/\tau, \log(1/\xi))\right)$ metric samples.*

While agnostic learning is typically considered a hard computational problem, in practice, machine learning techniques tend to perform very well. For a related notion of fairness, [KGZ18] show experimentally that using heuristic methods for learning decision trees and linear functions can improve the fairness of models significantly.

In the second setting, we turn our attention to learning a high-utility metric multifair hypothesis from a small sample of training examples. In order to guarantee the learned hypothesis will generalize to the unseen examples, we focus on learning over linear families. Specifically, we restrict our attention to learning hypotheses of the form $f_w(x) = \langle w, x \rangle$, for $w \in \mathcal{F} = [-B, B]^d$ given some bound $B > 0$. We show that a variant of stochastic gradient descent due to Nesterov learns a metric multifair hypothesis with time and sample complexities that depend polynomially on the dimension.

Theorem 4. *There is an algorithm that, given access to $T = \mathrm{poly}(d, B, 1/\gamma, 1/\tau, \log(1/\xi))$ random training examples and $n = \tilde{O}\left(\frac{\log(|\mathcal{C}|/\xi)}{\gamma\tau^2}\right)$ metric samples, with probability at least $1 - \xi$ over the random training examples and the metric samples, outputs a near-optimal metric multifair linear hypothesis $w \in \mathcal{F}$. The algorithm runs in time $O(|\mathcal{C}| \cdot T)$.*

As in Theorem 3, we can remove the linear dependence on $|\mathcal{C}|$ if we can agnostically learn $\mathcal{C}$.

Finally, we show that, in various respects, our positive results for learning a metric multifair hypothesis cannot be improved significantly. We give a reduction from a generic PAC learning task to learning a metric multifair hypothesis.

Theorem 5 (Informal)**.** *There is an efficient reduction from PAC learning a concept class $\mathcal{C}$ to learning a near-optimal metric multifair hypothesis for a related class $\mathcal{C}'$.*

This reduction has two immediate corollaries. First, we show that our algorithmic sample complexity is unconditionally tight up to constant factors. The second corollary is that under cryptographic assumptions (as in [Val84, GGM84, BR17]), the worst-case time required to learn a metric multifair hypothesis depends polynomially on $|\mathcal{C}|$. Taken together, these hardness results demonstrate that our algorithm's sample complexity cannot be improved and its running time is tight up to $\mathrm{poly}(|\mathcal{C}|)$ factors.

3. Broader Context

Many exciting recent works have investigated fairness in machine learning. In particular, there is much debate on the very definitions of what it means for a classifier to be fair [KMR17, Cho17, PRW+17, HPS16, CDPF+17, HKRR17]. Beyond the work of Dwork *et al.* [DHP+12], our work bears most similarity to a few recent works on subpopulation fairness [HKRR17, KNRW17, KGZ18]. As in this work, these papers investigate notions of fairness that aim to strengthen the guarantees of statistical notions, while maintaining their practicality. These works also draw connections between achieving notions of fairness and efficient agnostic learning. [KNRW17] and [KGZ18] show that using heuristic methods to agnostically "audit" models for fairness seems to work well in practice.

Metric multifairness does not directly generalize either [HKRR17] or [KNRW17], but we argue that it provides a more flexible alternative to these approaches for subpopulation fairness. In particular, these works aim to achieve specific notions of fairness – either calibration or equalized error rates – across a rich class of subpopulations. As has been well-documented [KMR17, Cho17, PRW+17], calibration and equalized error rates, in general, cannot be simultaneously satisfied. Often, researchers frame this incompatibility as a choice: either you satisfy calibration or you satisfy equalized error rates; nevertheless, there are many applications where some interpolation between accuracy (à la calibration) and corrective treatment (à la equalized error rates) seems appropriate. Metric-based fairness offers a way to balance these conflicting fairness desiderata. In particular, one could design a similarity metric that preserves accuracy in predictions and separately a metric that performs corrective treatment, and then enforce metric multifairness on an appropriate combination of the metrics. Different combinations of the two metrics would place different weights on the degree of calibration and corrective discrimination in the resulting predictor.

Finally, two recent works also investigate extensions to the fairness through awareness framework of [DHP+12]. Gillen *et al.* [GJKR18] study metric-fairness in online decision-making with an unknown fairness metric. Their work makes a strong learnability assumption about the underlying metric, whereas our focus is on fair classification when the metric is unknown and unrestricted. Rothblum and Yona [RY18] study fair machine learning under a different relaxation of metric fairness, which they call *approximate* metric fairness. They assume that the metric is fully specified and known to the learning algorithm. Their notion of approximate metric fairness aims to protect *all* (large enough) groups, and thus, is more strict than metric multifairness.

Conclusion Achieving fairness in classification systems is a nuanced goal. While most researchers agree that the community needs to take steps towards ensuring the fairness of these systems, the very definition of what it means to be fair is currently under fierce debate. Often, the arguments for one definition over another boil down to a fundamental difference in the type of task. As such, we believe metric multifairness – which provides a practically-motivated framework for task-specific fairness – represents an important contribution to the quiver of fairness and anti-discrimination tools. We defer a more in-depth technical coverage to the arXiv version [KRR].

References

[ALMK16] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*, 2016.

[BG18] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91, 2018.

[BR17] Andrej Bogdanov and Alon Rosen. Pseudorandom functions: Three decades later. In *Tutorials on the Foundations of Cryptography*, pages 79–158. Springer, 2017.

[CDPF$^+$17] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. *KDD*, 2017.

[Cho17] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 2017.

[DHP$^+$12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science (ITCS)*, pages 214–226, 2012.

[FPCG16] Avi Feller, Emma Pierson, Sam Corbett-Davies, and Sharad Goel. A computer program used for bail and sentencing decisions was labeled biased against blacks. it's actually not that clear. *The Washington Post*, 2016.

[GGM84] Oded Goldreich, Shafi Goldwasser, and Silvio Micali. How to construct random functions. In *Foundations of Computer Science, 1984. 25th Annual Symposium on*, pages 464–479. IEEE, 1984.

[GJKR18] Stephen Gillen, Christopher Jung, Michael J. Kearns, and Aaron Roth. Online learning with an unknown fairness metric. *arXiv preprint arXiv:1802.06936*, 2018.

[HKRR17] Úrsula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Calibration for the (computationally-identifiable) masses. *arXiv Preprint*, 1711.08513, 2017.

[HPS16] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016.

[KGZ18] Michael P. Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box postprocessing for fairness in classification. *arXiv Preprint*, 1805.12317, 2018.

[KMR17] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *ITCS*, 2017.

[KNRW17] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *arXiv preprint arXiv:1711.05144v3*, 2017.

[KRR] Michael P Kim, Omer Reingold, and Guy N Rothblum. Fairness through computationally-bounded awareness. *arXiv Preprint*.

[PRW$^+$17] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. On fairness and calibration. *NIPS*, 2017.

[RY18] Guy N Rothblum and Gal Yona. Probably approximately metric-fair learning. *arXiv Preprint*, 1803.03242, 2018.

[Val84] Leslie G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[Wad16] Kaveh Waddell. How algorithms can bring down minorities' credit scores. *The Atlantic*, 2016.