Metric Learning for Individual Fairness

Christina Ilvento
Harvard University, John A. Paulson School of Engineering and Applied Science,
Cambridge, MA, USA
cilvento@g.harvard.edu

Abstract

There has been much discussion concerning how "fairness" should be measured or enforced in classification. Individual Fairness [2], which requires that similar individuals be treated similarly, is a highly appealing definition as it gives strong treatment guarantees for individuals. Unfortunately, the need for a task-specific similarity metric has prevented its use in practice. In this work, we propose a solution to the problem of approximating a metric for Individual Fairness based on human judgments. Our model assumes access to a human fairness arbiter who is free of explicit biases and possesses sufficient domain knowledge to evaluate similarity. Our contributions include definitions for metric approximation relevant for Individual Fairness, constructions for approximations from a limited number of realistic queries to the arbiter on a sample of individuals, and learning procedures to construct hypotheses for metric approximations which generalize to unseen samples under certain assumptions of learnability of distance threshold functions.

2012 ACM Subject Classification Theory of computation → Machine learning theory

Keywords and phrases metric learning, individual fairness, fair machine learning

Digital Object Identifier 10.4230/LIPIcs.FORC.2020.2

Related Version `https://arxiv.org/abs/1906.00250`

Funding *Christina Ilvento*: This work was supported by the Smith Family Fellowship, Microsoft Research and the Sloan Foundation.

Acknowledgements The author is grateful for the comments of Cynthia Dwork.

1 Introduction

Determining what it means for an algorithm or classifier to be "fair" and how to enforce any such determination has become a subject of considerable interest as automated decision-making increasingly takes the place of direct human judgment. One attractive definition proposed is Individual Fairness [2], which states that similar individuals should be treated similarly, where similarity is encoded in a task-specific *metric*.

▶ **Definition 1** (Individual Fairness [2]). *Given a universe U, a metric $\mathcal{D} : U \times U \to [0,1]$ for a classification task with outcome set O, and a distance measure $d : \Delta(O) \times \Delta(O) \to [0,1]$, a randomized classifier $C : U \to \Delta(O)$ is Individually Fair if and only if for all $u, v \in U$, $\mathcal{D}(u,v) \geq d(C(u), C(v))$.*

Individual Fairness is appealing because each person is assured that her treatment is similar to that of any person similar to her.[1] However, the value of this assurance critically depends on the extent to which the similarity metric ($\mathcal{D}$) faithfully represents society's best

[1] By way of contrast, notions of fairness based on group level statistics can only provide individuals with the guarantee that if they are treated poorly, either someone in a different group is also treated poorly or someone in their group is treated well. Furthermore, many popular notions of statistical group fairness conflict with each other and cannot be satisfied simultaneously [1, 6].

1st Symposium on Foundations of Responsible Computing (FORC 2020).
Editor: Aaron Roth; Article No. 2; pp. 2:1–2:11
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

understanding of what constitutes similarity for a given task. Thus, the most significant barrier to implementing Individual Fairness in practice is the need to construct a similarity metric for each classification setting.

In this work we set out a path for constructing metrics for Individual Fairness based on judgments made by a qualified, fair-minded "human fairness arbiter." Our contributions include: (1) a framework for useful approximations to a metric for Individual Fairness; (2) a limited, realistic query model for determining the arbiter's judgments of who is similar to whom; (3) a method for constructing approximations to the true metric with limited queries to the arbiter by using distances from a (set of) representative individual(s); (4) a procedure for generalizing these approximations to unseen samples based on limited learnability assumptions. Throughout this work we make no assumption on the form of the metric or the features included in the learning procedure with the clearly stated exception of Assumption 1 concerning learnability of threshold functions. As our results are built upon a series of sequential steps including new terminology and machinery, we first present an extended introduction to highlight the key concepts, logic and results. These results are expanded and discussed in detail in the full version of the paper [4].

1.1 Model

In this work, we take the viewpoint that fairness is not well described by either accuracy or group statistics alone. Instead, we view fairness as a highly contextual property one can identify but not necessarily describe.[2] Our goal is to produce a metric which results in similarity judgments with which fair-minded people would agree, rather than satisfying any particular statistical properties.[3] The core of our model is the human fairness arbiter, a fair-minded individual who is free from explicit biases or arbitrary preferences, is motivated to engage ethically and honestly in the query protocol, and has sufficient knowledge and contextual understanding of who is similar to whom for a particular task. The arbiter is not expected to provide us a description or specification of the distance metric.

A critical part of learning metrics based on human judgments is determining the type of queries to ask in order to solicit consistent, fast responses. To that end, we assume that we cannot ask the arbiter to consider more than a few individuals at a time, e.g., it is not realistic to ask the arbiter to find the closest pair of elements in the universe.

We ask the arbiter to answer two types of queries in this work: relative distance queries, (e.g., is a closer to b or c), and real-valued distance queries.

▶ **Definition 2** (Real-valued distance query). $\mathsf{O}_{\mathsf{REAL}}(u, v) := \mathcal{D}(u, v)$.

▶ **Definition 3** (Triplet query). $\mathsf{O}_{\mathsf{TRIPLET}}(a, b, c) := \{1 \textit{ if } \mathcal{D}(a, b) < \mathcal{D}(a, c), 0 \textit{ if } \mathcal{D}(a, c) \leq \mathcal{D}(a, b)\}$.

Producing a consistent set of real-valued distances is not a natural judgment most people are accustomed to making, so we assume that real-valued queries are very "expensive" for the arbiter to answer. Furthermore, maintaining internal consistency may *increase* the query cost as the number of queries increases. Relative distance queries have been used successfully for human evaluation in image processing and computer vision, e.g. [8, 9], and we anticipate they will be significantly easier for the arbiter to evaluate. Demonstrating how to replace difficult queries with easy queries is a significant part of our contribution.

[2] [3] takes a similar approach in which a judge "knows it when she sees it," but is not required to articulate why a decision is unfair.

[3] We discuss different types of agreement, and the extent to which we fully achieve this goal, in Section 8 of the full paper.

We make several simplifying assumptions about the nature of the human fairness arbiter in the main results of this work. (1) There is either one arbiter or all arbiters agree on all decisions. (2) The arbiter does not change her opinion over the query period. (3) The arbiter's responses are consistent, i.e., if she answers that a is closer to b than it is to c, her responses to real-valued queries will also reflect this relative judgment.[4] For the majority of this work, we focus on the query model specified above, which requires the arbiter to answer with arbitrary precision. We also present a relaxed model which allows the arbiter to answer real-valued queries with bounded noise and does not require arbitrarily small distinctions in relative distances queries. The main results presented are replicated in the relaxed model. As the results are similar, we focus on the more simple exact model in the main presentation of our results.[5]

1.2 Contributions

Approximating the metric by contracting. Our first key observation is that Individual Fairness only requires that we do not *overestimate* distances. This motivates our definition of a *submetric*, which is a contraction of the original metric and can be substituted for the original metric and still maintain Individual Fairness.

Constructing submetrics based on distances from representative elements. Taking the difference in distance to a single reference or "representative" point is one of the simplest ways to produce an underestimate of the distance between two elements. Submetrics based on distances from representative elements form the basis of all of our constructions, and although this may seem simplistic, it has a significant advantage when it comes to deciding which queries to ask the arbiter: *ordering.* An ordering of elements by increasing distance from the representative can be constructed with relative distance (easy) queries used as a comparator. Once this ordering is established, real-valued distances at a given granularity can be layered on top in a *sublinear* number of real-valued (hard) queries.

Choosing representatives. A single representative may not be sufficient to capture all relevant distance information, but combining the information from multiple representative elements can produce a more complete picture of the distances between all pairs of individuals. But which representatives should we choose to maximize distance preservation? We discuss a general, randomized approach and show that given certain properties of the metric, i.e. how tightly packed individuals are, a random set of representatives of reasonable size will have good distance preservation properties.

Generalizing submetrics to unseen samples. Once we have established how to construct a submetric for a fixed sample of elements, our next step is to generalize to unseen samples. Our results are based on an assumption that threshold functions, i.e. binary indicators of whether an element is closer to a representative than a given threshold, are efficiently learnable. We show how to combine threshold functions to simulate rounding distances to a representative and then exhibit appropriate parameters to construct an efficient combined learning procedure.

[4] Please see Section 8 in the full paper for additional details.

[5] Extended discussion of the exact query model and a more general definition of relative queries is included in Section 3 of the full paper. The relaxed query model is discussed in detail in Section 7 of the full paper.

Relaxing arbiter requirements. Finally, we present a relaxation of the arbiter query model in which the arbiter (1) may respond to real-valued queries with arbitrary bounded noise and (2) is not required to make arbitrarily precise distinctions between distances and may instead declare relative comparisons to be "too close to call." This model more closely matches the reality of human arbiters, and our results extend with improvements in query complexity at the cost of increased error magnitude.

1.3 Preliminary terminology and definitions

We refer to the universe of individuals as U, a distribution over the universe of individuals as $\mathcal{U}$, and the size of the universe as $|U| = N$. We write $\mathcal{U}^*$ for the uniform distribution over U. We assume $\mathcal{D} : U \times U \to [0, 1]$ for simplicity. Individual Fairness does not require that distances between individuals be maintained exactly, only that they not be exceeded. This observation motivates our definition of a *submetric* which is a contraction of the true metric, i.e., it does not *overestimate* any distance beyond a small additive error term.[6]

▶ **Definition 4** ($\alpha-$submetric). *Given a metric $\mathcal{D}$, $\mathcal{D}' : U \times U \to [0, 1]$ is an α-submetric of $\mathcal{D}$ if for all $u, v \in U$, $\mathcal{D}'(u, v) \leq \mathcal{D}(u, v) + \alpha$.*

Any classifier which satisfies the distance constraints of the submetric $\mathcal{D}'$ will also satisfy those of $\mathcal{D}$, modulo small additive error.[7] Given an α-submetric it is possible to eliminate the additive error by taking $\max\{0, \mathcal{D}'(x, y) - \alpha\}$. On the other hand, we want to avoid contracting distances to the point of triviality. We say that a submetric is $(\beta, c)-$nontrivial if a β fraction of distances between pairs preserve at least a $c-$fraction of their original distance.[8]

▶ **Definition 5** ($(\beta, c)-$nontrivial). *Given a metric $\mathcal{D}$, a submetric $\mathcal{D}'$ of $\mathcal{D}$ is (β, c)-nontrivial for the distribution $\mathcal{U}$ if $\Pr_{u,v\sim\mathcal{U}\times\mathcal{U}}\left[\frac{\mathcal{D}'(u,v)}{\mathcal{D}(u,v)} \geq c\right] \geq \beta$.*

1.4 Constructing submetrics from arbiter judgments

A core component of this work is constructing submetrics based on distance information (either exact or underestimated) from a single representative element. We define the *representative submetric* $\mathcal{D}_r$ in the following Lemma. (The proof of follows from triangle inequality.)

▶ **Lemma 6.** *Given a representative r, $\mathcal{D}_r(x, y) := |\mathcal{D}(r, x) - \mathcal{D}(r, y)|$ is a 0-submetric of $\mathcal{D}$.*

Given a sample of N individuals, $\mathcal{D}_r$ can be constructed from $O(N)$ queries to $\mathsf{O}_{\mathsf{REAL}}$. Although $O(N)$ may seem good compared with the $O(N^2)$ queries required to reconstruct the whole metric, it can be improved to $O(\log(N))$ by supplementing with relative distance queries. Our general strategy will be to show that (1) an *ordering* of elements by distance from a representative can be constructed using $\mathsf{O}_{\mathsf{TRIPLET}}$ as a comparator, and (2) given this ordering, the real-valued distances between each element and the representative can be closely approximated by labeling the ordering with distances at granularity α, which requires a sublinear number of real-valued queries. Algorithm 1 outlines this process.[9]

[6] This relaxation is very similar to the notion of (d, τ) metric fairness of [5] and approximate metric fairness of [10].

[7] As originally noted in [2], the distance measure need not be a true metric, i.e. it does not strictly need to obey triangle inequality or distinguish unequal elements.

[8] Nontriviality is defined over a product of identical distributions of elements in the universe. There is no general obstacle to extending our results to more complicated scenarios, but definitions of density (presented in the full version in Section 6) would need to be adjusted.

[9] See Section 4 in the full version for the detailed specifications and analysis.

Algorithm 1 (Pseudocode).

```
Inputs: the representative $r$, a set of elements $U$, error parameter $\alpha$, interfaces $\mathsf{O}_{\mathsf{TRIPLET}}$ and $\mathsf{O}_{\mathsf{REAL}}$.
Output: an $\alpha-$submetric $\mathcal{D}'_r$.
Initialize the submetric $\mathcal{D}'_r(x,y) \leftarrow 0$ for all $x, y \in U \times U$.
Order the elements of $U$ by distance from $r$ using $\mathsf{O}_{\mathsf{TRIPLET}}$ as a comparator.
Designate the entire ordered list as the first continuous range.
while there are still ranges left to be labeled do
  Select a range left to be labeled.
  Query $\mathsf{O}_{\mathsf{REAL}}(r, \text{first})$ and $\mathsf{O}_{\mathsf{REAL}}(r, \text{last})$ for the first and last elements in the range.
    if the difference in distances is $> \alpha$ then
      Split into two continuous ranges, each with half of the elements in the current range.
    else set $\mathcal{D}'_r(r,x)$ to $\mathsf{O}_{\mathsf{REAL}}(r, \text{first})$ for each element $x$ in the range.
Set $\mathcal{D}'_r(x,y) = |\mathcal{D}'_r(r,x) - \mathcal{D}'_r(r,y)|$ for all $x, y$ in the ordering.
return $\mathcal{D}'_r$.
```

Theorem 7 states that Algorithm 1 produces an $\alpha-$submetric, which follows from observing that rounding $\mathcal{D}(r,x)$ and $\mathcal{D}(r,y)$ down by at most α results in an increase (or decrease) of at most α in $|\mathcal{D}(r,x) - \mathcal{D}(r,y)|$. The bound of $O(N\log(N))$ relative distance queries follows from a straightforward analysis of sorting. The bound of $O(\max\{\frac{1}{\alpha}, \log(N)\})$ real-valued queries is included in Section 4 in the full paper. Briefly, the analysis considers the maximum number of continuous ranges that, when split, result in one range with difference greater than α and one with less. In the worst case, this results in logarithmic dependency on N or $\frac{1}{\alpha}$.

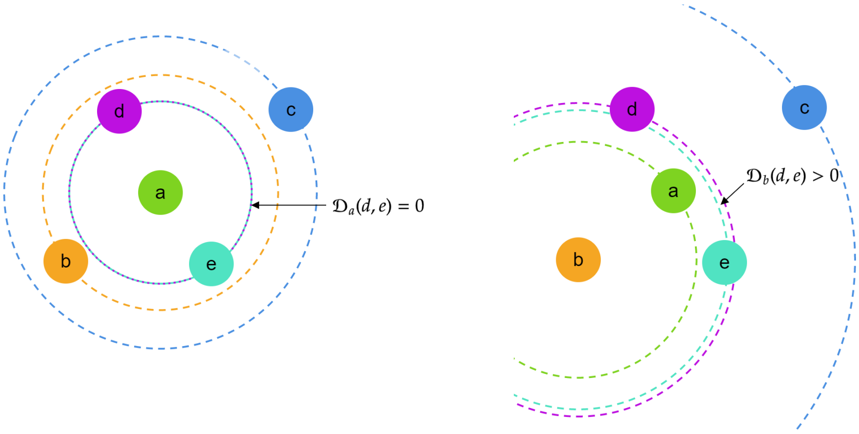
▶ **Theorem 7.** *Algorithm 1 produces an $\alpha-$submetric of $\mathcal{D}$ which preserves $\mathcal{D}(r,u)$ for each $u \in U$ (with additive error $\leq \alpha$) from $O(\max\{\frac{1}{\alpha}, \log(N)\})$ queries to $\mathsf{O}_{\mathsf{REAL}}$ and $O(N\log(N))$ queries to $\mathsf{O}_{\mathsf{TRIPLET}}$.*

The submetric produced by Algorithm 1 preserves distances between r and other elements well, as $\mathcal{D}'_r(r,x)$ is rounded down by at most α, but we cannot make guarantees on distance preservation between arbitrary pairs without further information. For example, with only the information that u and v are equally distant from r, it is impossible to distinguish whether the distance between u and v is zero or equal to twice their distance from r. (See Figure 1). Submetrics constructed based on different representatives preserve different information about the underlying metric, so we can construct more expressive submetrics by aggregating information from multiple representatives. Taking $\mathsf{maxmerge}(\{\mathcal{D}_i\}, x, y) := \max_i \mathcal{D}_i(x,y)$, it's straightforward to show that if all $\mathcal{D}_i$ are submetrics of $\mathcal{D}$, then the $\mathsf{maxmerge}$ of the set is also a submetric of $\mathcal{D}$, and that the merge preserves the "best" distance known for each pair.[10]

1.5 Choosing good representative elements

Although the $\mathsf{maxmerge}$ of submetrics based on multiple representatives is an improvement over a single representative, we still cannot make any guarantees about distances between pairs which do not include a representative. There are two approaches one might take to give non-triviality guarantees for arbitrary pairs: (1) develop specialized strategies for combining

[10] Formal analysis of $\mathsf{maxmerge}$ and the proof of Lemma 6 appear in Section 3 of the full paper. The proof of Theorem 7 as well as a precise description of Algorithm 1 appear in Section 4 of the full paper.



(a) a chosen as representative.

(b) b chosen as representative.

Figure 1 The impact of representative choice on distance preservation. The distance between each element and the chosen representative is the radius of the shell containing the element. The difference in radii of each pair of shells indicates the distance between the pair of elements under $\mathcal{D}_a$ or $\mathcal{D}_b$. If a is chosen as a representative, notice that d and e are indistinguishable using distance from a alone. Choosing representative b preserve distances better than a, but still does not distinguish d and e very well.

representative submetrics which depend on the structure of the metric, e.g., Euclidean distance, or (2) characterize generic randomized representative selection strategies. In this extended introduction, we focus on the randomized strategies for full generality.

Distance preservation via γ-nets. The crux of the argument for nontriviality with random representatives is (1) a random set of representatives is likely to be "close to" a significant portion of the distribution $\mathcal{U}$, and (2) we can bound the magnitude of underestimates based on the distance from a representative. Below, we formally define a $\gamma-$net to capture the notion of being "close to" or "covering" a set of elements.

▶ **Definition 8.** *A set $R \subseteq U$ is said to form a $\gamma-$net for a subset $V \subseteq U$ under $\mathcal{D}$ if for all balls of radius γ (determined by $\mathcal{D}$) containing at least one element $v \in V$, the ball also contains $r \in R$.*

Intuitively, the distance between r and x will be nearly identical to the distance between a close neighbor of r and x, so we can conclude that if a set of representatives forms a $\gamma-$net for a subset of U, then pairs with at least one element in the net will have their original distance preserved up to a 2γ contraction. (Proofs of Lemmas 9 and 10 follow from triangle inequality.)

▶ **Lemma 9.** *For all $u, v \in U \backslash \{r\}$, $\mathcal{D}(u,v) - \mathcal{D}_r(u,v) \leq \min\{2\mathcal{D}(r,u), 2\mathcal{D}(r,v)\}$, where $\mathcal{D}_r(u,v) := |\mathcal{D}(r,u) - \mathcal{D}(r,v)|$.*

▶ **Lemma 10.** *If a set of representatives $R \subseteq U$ forms a $\gamma-$net for $V \subseteq U$, then for every pair $x, y \in V \times U$ there exists $r \in R$ such that $\mathcal{D}(x,y) - \mathcal{D}_r(x,y) \leq 2\gamma$, where $\mathcal{D}_r(x,y) := |\mathcal{D}(r,x) - \mathcal{D}(r,y)|$.*

Of course, forming a $\gamma-$net for an *arbitrary* γ isn't enough to give a good nontriviality guarantee. To understand how representatives which form a $\gamma-$net will preserve distances, we define *density* and *diffusion* below to characterize the relevant properties of the metric and

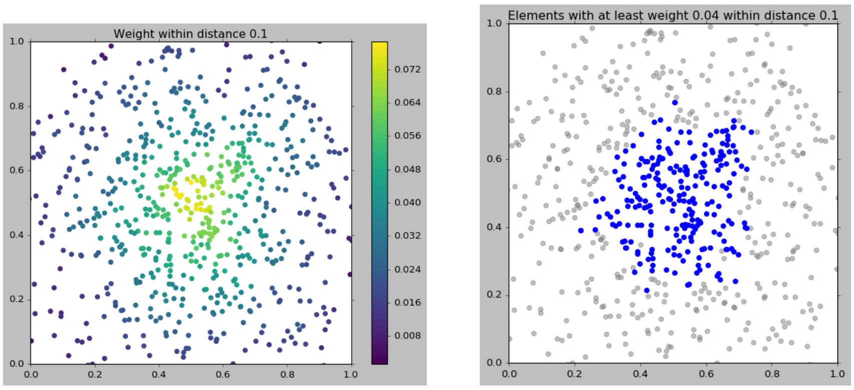


Figure 2 A visualization of the weight of elements, b, within distance $\gamma = .1$ of each element under $\mathcal{U}^*$ for an example universe of points in $[0,1]^2$ where $\mathcal{D}$ is taken as Euclidean distance. The color assigned to each point on the left indicates the weight of elements in the universe which are within distance $\gamma = 0.1$ from the element under the uniform distribution $\mathcal{U}^*$. On the right, the points with at least weight $b = 0.04$ of $\mathcal{U}^*$ within distance $\gamma = 0.1$ are highlighted in blue. This example is $(\gamma = 0.1, a = .31, b = 0.04)-$dense for $\mathcal{U}^*$. That is, 31% of elements in the universe are within distance 0.1 of 4% of the rest of the universe.

distribution. The notion of $(\gamma, a, b)-$dense is intended to capture the weight (a) of elements that have a significant weight (b) on their close neighbors (distance γ) under $\mathcal{U}$ as a way to characterize how likely it is that a randomly chosen representative will be γ-close to a significant fraction of elements.

▶ **Definition 11** ($(\gamma, a, b)-$dense)**.** *Given a distribution $\mathcal{U}$ over U, a metric $\mathcal{D}$ is $(\gamma, a, b)-$dense for $\mathcal{U}$ if there exists a subset $A \subseteq U$ with weight a under $\mathcal{U}$ such that for all $u \in A$ $\Pr_{v\sim\mathcal{U}}[\mathcal{D}(u,v) \leq \gamma] \geq b$.*

$(p, \zeta)-$diffuse, defined below, captures what fraction of distances can tolerate a contraction proportional to ζ without becoming trivial.

▶ **Definition 12** ($(p, \zeta)-$diffuse)**.** *Given a distribution $\mathcal{U}$, a metric $\mathcal{D}$ is $(p, \zeta)-$diffuse if the fraction of distances between pairs of elements in $\mathcal{U} \times \mathcal{U}$ greater than ζ is p, i.e. $\Pr_{u,v\sim\mathcal{U}\times\mathcal{U}}[\mathcal{D}(u,v) \geq \zeta] \geq p$.*

A metric can be described by many combinations of density and diffusion parameters, as illustrated in Figure 2. These parameters are highly related, and we generally consider the combination of $(\gamma, a, b)-$dense and $(p, \frac{2\gamma}{1-c})-$diffuse. Although $\frac{2\gamma}{1-c}$ initially seems an arbitrary quantity, it indicates that a $p-$fraction of pairs will have distances preserved by a factor of c if the maximum contraction for those pairs is no more than 2γ. Thus the values of γ and c, which in turn dictate p, a, and b, (assuming $\zeta = \frac{2\gamma}{1-c}$) can loosely be seen as a tradeoff between how many pairs will have distance preservation guarantees and how significant the guarantees will be.

Nontriviality properties of $\gamma-$nets. Next, we relate the magnitude of γ to the non-triviality properties of the maxmerge of a set of representative submetrics. Lemma 13 states that a submetric based on a set of representatives which form a $\gamma-$net for a subset of U will have nontriviality properties related to the diffusion properties of $\mathcal{D}$ and the weight of the subset in $\mathcal{U}$.

▶ **Lemma 13.** *If a set of representatives* $R \subseteq U$ *form a* $\gamma-$*net for weight* w *of* $\mathcal{U}$ *and* $\mathcal{D}$ *is* $(p, \frac{2\gamma}{1-c})-$*diffuse on* $\mathcal{U}$*, then the submetric* $\mathcal{D}_R(x, y) := \mathsf{maxmerge}(\{\mathcal{D}_r | r \in R\}, x, y)$ *is* $(p', c)-$*nontrivial for* $\mathcal{U}$*, where* $p' \geq p - (1 - w)^2$*.*

The proof follows from a worst-case analysis of the fraction of pairs with at least one element in the net with distance large enough that a 2γ contraction leaves at least a c-fraction of the original distance. The nontriviality guarantees of Lemma 13 are conservative, and we stress that our goal is to show the possibility of positive results, rather than achieving optimal performance or guarantees.

Representative set size. We now consider how likely it is that a set of random representatives drawn from $\mathcal{U}$ will form a $\gamma-$net for a significant fraction of $\mathcal{U}$. Lemma 14 characterizes the necessary representative set size based on the density and diffusion properties of the metric. The proof follows from characterizing the probability of "hitting" a sufficient weight of the distribution with a sample of a given size, and arguing that no element in our subset of interest can be more than 3γ far from any of the "hitting" elements.

▶ **Lemma 14.** *Given access to unlimited queries to the arbiter, if a metric* $\mathcal{D}$ *is* $(\gamma, a, b)-$*dense and* $(p, \frac{6\gamma}{1-c})-$*diffuse on* $\mathcal{U}$*, then a random set of representatives* R *of size at least* $\frac{1}{b}\ln(\frac{1}{b\delta})$ *will produce a* $(p - (1 - a)^2, c)$*-nontrivial submetric for* $\mathcal{U}$ *with probability at least* $1 - \delta$*.*

Random sampling is not the only method to construct a $\gamma-$net, and our strategy is motivated by simplicity as much as generality. In practice it may be more efficient to use the distance information from previously selected representatives to inform the selection of the next representative. For example, omitting or down-weighting any candidates that are already very close to existing representatives, or using a greedy strategy.[11]

1.6 Generalizing arbiter judgments

Now that we have shown how to construct a nontrivial submetric with ongoing access to the arbiter, we consider the problem of generalizing the arbiter's responses to unseen samples. Our goal is to construct efficient learners for submetrics as in Valiant's Probably Approximately Correct (PAC) model of learning [7]. However, we do not want to be too prescriptive about the submetric concept class, particularly about the representation of elements. Instead, we will make an assumption about the learnability of *threshold functions* and construct learning procedures for submetrics using threshold functions as building blocks without any additional direct access to labeled or unlabeled samples. More formally, our goal is to produce an efficient submetric learner, defined below.

▶ **Definition 15** (Efficient submetric learner)**.** *A learning procedure is an efficient* $\alpha-$*submetric learner if for all* $\varepsilon, \delta \in (0, 1]$*, given access to labeled examples, with probability at least* $1 - \delta$ *over the randomness of the sampling and the learning procedure produces a hypothesis* h_r *such that* $\Pr_{x,y\sim\mathcal{U}\times\mathcal{U}}[h_r(x, y) - \mathcal{D}(x, y) \geq \alpha] \leq \varepsilon$ *in time* $O(poly(\frac{1}{\varepsilon}, \frac{1}{\delta}))$*.*

To show how to construct an efficient submetric learner, we first formalize our assumption of learnability of threshold functions. Next, we show how to combine the threshold function hypotheses for each representative to simulate rounding the distance between the representative and each element down to the nearest threshold. Finally, we specify the appropriate parameters for each component to achieve the desired bounds.

[11] Section 6 of the full paper contains proofs for Lemmas 9-14 and extended discussion of specialized strategies for representative selection, in particular strategies taking advantage of known metric structure.

Learnability of threshold functions. Assumption 1 (below) states that for every representative, there exists a set of thresholds and a learner for each threshold which, with high probability, produces an accurate hypothesis for the threshold function which generalizes to unseen samples.[12] ("With high probability" always refers to the probability over the randomness of sampling and the learner.) We first formally define a threshold function, which is a binary indicator of whether a particular element $u \in U$ is within distance $t \in [0, 1]$ of a representative r as $T_t^r(u) := \{1 \text{ if } \mathcal{D}(r, u) \leq t, 0 \text{ otherwise}\}$.

▶ **Assumption 1.** (Informal) *Given a metric $\mathcal{D}$ and a representative r, there exists a set of thresholds $\mathcal{T}$ such that $t \in [0, 1]$ for all $t \in \mathcal{T}$, $0 \in \mathcal{T}$, and $|\mathcal{T}| = O(1)$, and for every $t \in \mathcal{T}$ there exists an efficient learner $L_t^r(\varepsilon_t, \delta_t)$ which for all $\varepsilon_t, \delta_t \in (0, 1]$, with probability at least $1 - \delta_t$, produces a hypothesis h_t^r such that $\Pr_{x \sim \mathcal{U}}[h_t^r(x) \neq T_t^r(x)] \leq \varepsilon_t$ in time $O(poly(\frac{1}{\varepsilon_t}, \frac{1}{\delta_t}))$ with access to labeled samples of $T_t^r(u \sim \mathcal{U})$ for any distribution $\mathcal{U}$.*

Constructing submetric learners from threshold learners. Given Assumption 1, our next step is to determine how to combine the threshold learners into a learner for the representative submetric. (Notice that training data for the threshold function learners can be produced by post-processing the outputs of Algorithm 1.) Our strategy is similar to the rounding strategy used in Algorithm 1, using the threshold functions to identify the largest threshold which underestimates the distance between the representative and the element under consideration. The LinearVote mechanism takes in a set of hypotheses for the thresholds and outputs the threshold with which the most hypotheses agree. When all hypotheses output the correct value of their corresponding threshold function, LinearVote is equivalent to rounding $\mathcal{D}(r, x)$ down to the nearest threshold.

▶ **Definition 16** (LinearVote). *Given an ordered set of thresholds, $\mathcal{T} = \{t_1, t_2, \ldots, t_{|T|}\}$, and a set of hypotheses $H_{\mathcal{T}}^r = \{h_{t_1}^r, h_{t_2}^r, \ldots, h_{t_{|T|}}^r\}$, one corresponding to each threshold function,* $\mathsf{LinearVote}(\mathcal{T}, H_{\mathcal{T}}^r, x) := \arg\max_{t_i} \sum_{t_j < t_i} (1 - h_{t_j}^r(x)) + \sum_{t_j \geq t_i} h_{t_j}^r(x)$.

Algorithm 2 Pseudocode.

```
Inputs: error and failure probability parameters $\varepsilon, \delta$, density parameter $b$, a set of threshold function learners, the threshold set $\mathcal{T}$, and interfaces to the arbiter.
Sample a set of representatives $R \sim \mathcal{U}$ of size $\frac{1}{b} \ln(\frac{2}{b\delta})$ to produce $\gamma$−net with $\Pr \geq 1 - \frac{\delta}{2}$.
Generate labeled training data for the threshold learners via Algorithm 1.
Run each threshold learner $L_{t_i}^r$ with error parameters $\varepsilon_t \leftarrow \frac{\varepsilon}{2|R||\mathcal{T}|}$ and $\delta_t \leftarrow \frac{\delta}{2|R||\mathcal{T}|}$ to produce threshold function hypotheses $h_{t_i}^r$ for $r \in R$ and $t_i \in \mathcal{T}$.
For each representative, produce a hypothesis for distance from the representative by taking $h_r(x, y) := |\mathsf{LinearVote}(\mathcal{T}, \{h_{t_i}^r | t_i \in \mathcal{T}\}, x) - \mathsf{LinearVote}(\mathcal{T}, \{h_{t_i}^r | t_i \in \mathcal{T}\}, y)|$.
Combine the hypotheses for each representative into $h_R(x, y) := \mathsf{maxmerge}(\{h_r | r \in R\}, x, y)$.
return $h_R$.
```

Algorithm 2 combines all of our constructions thus far to create an efficient submetric learner: it chooses a set of representatives, learns threshold functions for each threshold for each representative, and combines the resulting hypotheses using LinearVote and maxmerge

[12] The formal statement of Assumption 1 is included in Section 5.1 of the full paper.

to produce a single submetric hypothesis.[13] Theorem 17 builds on the result of Lemma 14 and concludes that the parametrization of Algorithm 2 results in an efficient submetric learner.

▶ **Theorem 17.** *[Informal] Given a distance metric $\mathcal{D}$, and a distribution $\mathcal{U}$ over the universe, if there exist a set of thresholds $\mathcal{T}$ with maximum gap $\alpha_{\mathcal{T}}$ and efficient learners $\{L^{\tau}_{t_i \in \mathcal{T}}\}$ as in Assumption 1, and $\mathcal{D}$ is $(\gamma, a, b)-$dense and $(p, \frac{6\gamma+\alpha_{\mathcal{T}}}{1-c})-$diffuse on $\mathcal{U}$, then there exists an efficient $\alpha_{\mathcal{T}}$-submetric learner which produces a hypothesis h_R such that h_R is $(p-(1-a)^2-\varepsilon, c)-$nontrivial for $\mathcal{U}$.*

The proof of Theorem 17 follows from an analysis of the error parameter propagation. We briefly give some intuition for the analysis and implications of the theorem. First, the magnitude $\alpha_{\mathcal{T}}$ error follows from the same single direction rounding argument as for Algorithm 1. The error probability follows from noticing that at least one threshold function must be in error for one of the elements to result in an error in LinearVote. The failure probability "budget" is split evenly between failure to choose a good set of representatives (Line 1) as specified in Lemma 14, and failure of the underlying learning procedures (Line 3) derived by union bound. Compared with Lemma 14, the diffusion and nontriviality parameters are adjusted to take into account the additional rounding error magnitude of $\alpha_{\mathcal{T}}$ introduced by LinearVote and the combined hypothesis error probability ε. In practice, we expect that the set of thresholds which are learnable are unlikely to occur at regular intervals. Post-processing is a valuable tool to reduce the magnitude of $\alpha_{\mathcal{T}}$ (by re-mapping the threshold values in step 4 to reduce the maximum gap), but comes at the cost of reduced nontriviality guarantees.

The desired query complexity to the arbiter follows from basic analysis of the parameters. However, the query complexity bound can be improved significantly by observing that no independence of errors between threshold functions is assumed, allowing a single call to Algorithm 1 for each representative (rather than $|\mathcal{T}|$ calls). The dependence on $|R|$ can also be improved to logarithmic by sorting a single merged list of (representative, element) pairs, but we defer detailed discussion to Sections 5 and 6 of the full paper.

1.7 Relaxing the query model

Our results extend to a relaxed model in which arbiters are not expected to make arbitrarily small distinctions between distances or individuals and may answer real-valued queries with bounded noise. The relaxed model assumes that there are two fixed constants, α_L, the minimum precision with which the arbiter can distinguish elements or distances, and α_H, a bound on the magnitude of the (potentially biased) noise in the arbiter's real-valued responses. For any comparisons with difference smaller than α_L, the arbiter declares the elements indistinguishable or the difference "too close to call." The model allows for a "gray area" between α_L and α_H in which the arbiter may either respond with the true answer or "too close to call." For any differences larger than α_H, the arbiter responds with the true answer.

For the most part, our results translate to the relaxed model with minimal modification to the logic of the proofs to handle two-sided error in real-valued queries. Interestingly, the real-value query complexity improves to constant, as the worst-case behavior in Algorithm 1 is avoided as the arbiter "knows" not to worry about inconsequentially small distances. However, this does result in additional error magnitude, so the improved query complexity

[13] See Sections 5 and 6 in the full version.

does not come for free. Furthermore, unlike the exact model we won't necessarily be able to label a sample with perfect accuracy for every threshold function learner due to the bi-directional error. To handle this labeling problem, we modify the distribution of samples presented to each learner, eliminating samples whose labels are ambiguous, again resulting in increased error. Formal results in the relaxed model are discussed in Section 7 of the full paper.

References

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

2 Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM, 2012.

3 Stephen Gillen, Christopher Jung, Michael J. Kearns, and Aaron Roth. Online learning with an unknown fairness metric. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *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.

4 Christina Ilvento. Metric learning for individual fairness. *CoRR*, abs/1906.00250, 2019. arXiv:1906.00250.

5 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.

6 Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference, ITCS 2017, January 9-11, 2017, Berkeley, CA, USA*, pages 43:1–43:23, 2017.

7 Leslie G Valiant. A theory of the learnable. In *Proceedings of the sixteenth annual ACM symposium on Theory of computing*, pages 436–445. ACM, 1984.

8 Laurens Van Der Maaten and Kilian Weinberger. Stochastic triplet embedding. In *Machine Learning for Signal Processing (MLSP), 2012 IEEE International Workshop on*, pages 1–6. IEEE, 2012.

9 Michael J Wilber, Iljung S Kwak, and Serge J Belongie. Cost-effective hits for relative similarity comparisons. In *Second AAAI Conference on Human Computation and Crowdsourcing*, 2014.

10 Gal Yona and Guy N. Rothblum. 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.