Abstracting Fairness: Oracles, Metrics, and Interpretability

Cynthia Dwork
Harvard John A Paulson School of Engineering and Applied Sciences, Cambridge, MA, USA
Radcliffe Institute for Advanced Study, Cambridge, MA, USA
Microsoft Research, Mountain View, CA, USA
dwork@seas.harvard.edu

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

Guy N. Rothblum
Department of Computer Science and Applied Mathematics,
Weizmann Institute of Science, Rehovot, Israel
rothblum@alum.mit.edu

Pragya Sur
Harvard University, Center for Research on Computation and Society, Cambridge, MA, USA
pragya@seas.harvard.edu

Abstract

It is well understood that classification algorithms, for example, for deciding on loan applications, cannot be evaluated for fairness without taking context into account. We examine what can be learned from a *fairness oracle* equipped with an underlying understanding of "true" fairness. The oracle takes as input a (context, classifier) pair satisfying an arbitrary fairness definition, and accepts or rejects the pair according to whether the classifier satisfies the underlying fairness truth. Our principal conceptual result is an extraction procedure that learns the underlying truth; moreover, the procedure can learn an approximation to this truth given access to a *weak* form of the oracle. Since every "truly fair" classifier induces a coarse metric, in which those receiving the same decision are at distance zero from one another and those receiving different decisions are at distance one, this extraction process provides the basis for ensuring a rough form of *metric fairness*, also known as *individual fairness*.

Our principal technical result is a higher fidelity extractor under a mild technical constraint on the weak oracle's conception of fairness. Our framework permits the scenario in which many classifiers, with differing outcomes, may all be considered fair.

Our results have implications for *interpretablity* – a highly desired but poorly defined property of classification systems that endeavors to permit a human arbiter to reject classifiers deemed to be "unfair" or illegitimately derived.

2012 ACM Subject Classification Theory of computation → Machine learning theory

Keywords and phrases Algorithmic fairness, fairness definitions, causality-based fairness, interpretability, individual fairness, metric fairness

Digital Object Identifier 10.4230/LIPIcs.FORC.2020.8

Related Version A full version of the paper is available at `https://arxiv.org/pdf/2004.01840.pdf`.

Funding *Cynthia Dwork*: This research was supported in part by NSF grant CCF-1763665 and Microsoft Research.
Christina Ilvento: This work was supported by Microsoft Research and the Sloan Foundation.
Guy N. Rothblum: This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 819702), the Israel Science Foundation (grant number 5219/17), and Microsoft Research.

Pragya Sur: This research was supported by the Center for Research on Computation and Society, Harvard John A. Paulson School of Engineering and Applied Sciences, and in part by Microsoft Research.

Acknowledgements This research was conducted, in part, while the authors were at Microsoft Research, Silicon Valley.

1 Introduction

Definitions of fairness, for example, in the context of accept/reject classification algorithms, mostly fall into two main categories: group fairness definitions are requirements on various forms of statistical equality in the treatment of disjoint demographic groups; *individual* (or *metric*) fairness requires that individuals that are similar with respect to the classification task at hand should be treated similarly by the classifier. Although intuitively appealing, group fairness definitions suffer from internal inconsistency and incompatibility [4, 3, 21, 1]. (See also [27].)

On the other hand, individual fairness requires a task-specific similarity metric, which may be difficult to find.[1] *Counterfactual Fairness*, proposed by Kusner et al. in 2017 [22], is an approach to capturing individual fairness via *counterfactual reasoning* as put forth by Pearl [28]. Counterfactual fairness seeks to prevent discrimination based on protected attributes, such as race or sexual preference, by requiring that individuals' outcomes "would have been" the same in a counterfactual world in which these attributes have different values. To make such an assertion, the definition relies on a causal model that captures the ways in which these attributes influence other attributes relevant to classification. Thus, to evaluate whether a predictor for loan default is counterfactually fair for sexual orientation, one would construct a causal model reflecting the relationships between sexual orientation and the features weighed by the predictor, and then determine whether the predictions are inappropriately dependent on orientation. In this definition, the causal model replaces the metric as the specification of fairness. The classifier is evaluated for fairness in the context of the causal model just as in metric fairness the classifier is evaluated for fairness in the context of the metric.

As widely noted, and partially addressed in later work (Kilbertus et al., 2019 [18]), this approach suffers from the fact that different data generation models can give rise to the same distribution on outcomes. In particular, a blatantly unfair classifier can satisfy the definition when paired with a suitably contrived model, showing that the choice of model is itself a vector for unfairness.

More generally, the maxim "All models are wrong but some models are useful" highlights the dangers of a fairness definition that evaluates a classifier in the context of a stated model: What are the semantics of having the (model, classifier) pair satisfy the definition when the model is wrong (which is always the case!)?

To complete the counterfactual fairness approach (among others), one might assume the existence of an expert that can judge whether or not a classifier is "truly fair" in a given context. For example, a domain expert may reject a (causal model, classifier) pair for home loan decisions that satisfies the technical definition of counterfactual fairness but in which the model has been contrived to use zip code instead of race in order to obfuscate racial bias. In this work we investigate what can be learned by interacting with such an expert.

[1] See, however, the recent proposal of [11].

We abstract the problem by instantiating "true fairness" (and the expert who knows what this is) via an oracle that holds a collection $\mathcal{T} \subseteq \{0,1\}^{|\mathcal{X}|}$ of vectors specifying the classification outcome for each individual in the universe $\mathcal{X}$ of possible individuals[2]. Each $t \in \mathcal{T}$ corresponds to a classifier that the oracle considers to be fair, at least in some context. As an example, one might imagine that the oracle has access to the true data generation model, and it evaluates classifiers in this single context.

Our principal conceptual result is an (inefficient) *extraction procedure* that learns the underlying truth (collection $\mathcal{T}$) held by the oracle under the assumption that the contexts of interest are of bounded size. Once the assumption is cleanly stated it is not surprising that $\mathcal{T}$ can be extracted by brute force, so this first contribution is the conceptual framing of the problem (Sections 2 and 3). This result makes no assumptions about the set of fair classifiers accepted by the oracle, nor about the particular context(s) that make the oracle accept a classifier. We extract the full set of classifiers for which there exists *some* context that makes the oracle accept.

Under the assumption that counterfactual fairness (or any other causality-based definition, such as path-specific effects [26]), combined with the true causal model (or an appropriate approximation), genuinely captures fairness, our results imply that one can extract, from an oracle with access to the true model, a coarse metric for individual fairness. This holds because every classifier induces a coarse metric in which those receiving positive decisions are at distance zero from one another, and similarly for those receiving negative decisions. (See Remark 1.)

We then turn to *weak* oracles, which solve a more relaxed promise problem. Each weak oracle $\tilde{\mathcal{O}}$ is a relaxation, based on a given notion of closeness of classifiers, of a strong oracle, $\mathcal{O}$. Weak oracles always accept the (context, classifier) pairs accepted by their strong counterparts, but only reject (context, classifier) pairs where the classifier is "far" from an accepted classifier for the given distance notion.

We consider two types of closeness in defining weak oracles: Hamming distance, where the reconstruction problem is straightforward, provided the members of $\mathcal{T}$ are sufficiently separated[3], and an asymmetric *transportation cost* $\mathcal{C}$ that does not satisfy the triangle inequality. Our transportation cost is closely related to individual fairness: $\mathcal{C}(t \to c)$ captures the number of pairs of individuals that are treated similarly in t but differently in c. In essence, the transportation cost notion requires *less* of the oracle: δ-weak oracles[4] with this notion of distance may not know *how* individuals should be treated for the task at hand, but may have a sense of who should be treated similarly to whom. This lack of decisiveness on the part of the oracle makes extraction much more difficult. Not only does it lead to a transportation cost that is not even a distance function, but it also limits what can possibly be extracted even if $\delta = 0$: under this notion, the distance between a classifier and its complement is 0! [5] In consequence, rather than aiming to extract the set of fair classifiers, we extract a set of fair partitions, where each partition specifies which individuals are similar to each other. The partition can also be viewed as a coarse metric.

[2] We focus on the case of binary, deterministic classifiers. Such a classifier can only satisfy individual fairness if for all individuals u, v, $d(u,v) \in \{0,1\}$, where d is the task-specific metric. In Remark 1 we discuss amplification of this technique to a richer class.

[3] Much as it is possible to learn a mixture of Gaussians provided the means are sufficiently far apart.

[4] Oracles only guaranteed to reject classifiers at distance greater than δ from all $t \in \mathcal{T}$.

[5] Our techniques apply to a symmetrized version of $\mathcal{C}(t \to c)$, defined by the fraction of pairs of individuals that disagree between t and c (Section 3). This case is, in fact, easier than the transportation cost.

Our principal technical result is a high fidelity extractor in the transportation cost model, under a mild technical constraint on the weak oracle's conception of fairness. For $t \in \mathcal{T}$, define $t^{\text{flip}}(x) = 1 - t(x)$ for all $x \in \mathcal{X}$. The assumption is: for $t \in \mathcal{T}$ for which the weak oracle rejects t^{flip}, it also rejects all classifiers very close (in Hamming distance) to t^{flip}.

Interpretability

Our results have implications for *interpretablity* – a highly desired but poorly defined property of classification systems that endeavors to permit a human arbiter to reject classifiers deemed to be "unfair" or illegitimately derived. If "interpretability" permits a knowledgeable human to distinguish truly fair from truly unfair classifiers, then there is a procedure to extract from the human information a measure of similarity for pairs of individuals. Roughly speaking, we can get our hands on a metric, even when the closeness notion for classifiers is the Hamming distance on their vector representation, which is unrelated to metric fairness!

▶ Remark 1. In this work, the "metric" we extract from the oracle is crude: all distances are either 0 or 1. Metrics of this type can be amplified to yield a richer class of metrics by considering a collection of oracles with varying tolerance for unfairness. For example, given a metric $d : \mathcal{X} \times \mathcal{X} \to [0, 1]$, we can instantiate k approximations $\{d'_1, \ldots, d'_k\}$ of d such that $d'_i(u, v) := 1$ if $d(u, v) > \frac{k}{i}$ and 0 otherwise. Given access to an oracle for each threshold, we can apply the extraction procedure multiple times to (approximately) recover this set of $\{0, 1\}-$metrics. The recovered collection can then be combined to form an approximation of d, using the threshold combination procedure developed in [11]. See also [8, 12, 14] for demonstrations of the usefulness of coarse metrics.

Related Work

There is a vast literature on algorithmic fairness. The theory of algorithmic fairness was first studied by Dwork et al. in 2012 [4]. In addition to defining individual fairness, this work noted that sensitive attributes may be holographically embedded in the data, showed the benefits of utilizing, rather than trying to suppress, the sensitive information; showed the power of Individual Fairness when given a metric; examined the group fairness property of demographic parity and gave examples motivating its dismissal as a fairness solution concept, and provided a metric-based approach to Fair Affirmative Action. Earlier work suggested concrete approaches based on training on a modified dataset in which the proportion of positive labels is equal in disjoint demographic groups, in the hopes that a classifier trained on these new labels will imbibe the group fairness properties of the training data [29, 15]. A second approach added a regularization term to the classification training objective to quantify the degree of bias or discrimination [16, 2]. Subsequent work saw heavy investment in algorithms satisfying group-based criteria, even in the face of the negative results about the compatibility of natural group fairness objectives [27, 3, 21, 17, 5]. Individual fairness, predicated on access to a similarity metric, proceeded more slowly, although the literature contains several works extending the theory [5, 30, 8, 20]. Recent work [11] combines insights from HCI and computational learning theory to learn an approximation to a metric known to a human arbiter with surprisingly few queries. An intriguing "middle ground" enforces *calibration* (in the case of scoring functions [10]) *simultaneously* for large numbers of intersecting subpopulations (see [6] for a treatment of fair *rankings* in this setting). A variant of the multiple intersecting groups approach [17] enforces *Equalized Odds* [9] among all pairs of groups simultaneously. An economics justification for Equalized Odds is put forth in [13]. Equalized Odds and related candidate fairness criteria are criticized through the lens of graphical models [1].

Still other work employs deep learning to build *fair representations* of individuals that, speaking intuitively, retain much useful information for classification or even transfer learning, but "screen out" sensitive demographic information [31, 7, 25]. Finally, there is also a vast literature on *interpretability*. See [24] for a discussion of what this might mean (and hurdles to be overcome); the course notes of Lakkaraju [23] contain a wealth of examples and references for this literature.

Our work was inspired by the elegant proposal of Counterfactual Fairness by Kusner, Loftus, Russell, and Silva [22]. A related definition of fairness concentrates on path-specific effects [26] (see also [19]). Kilbertus et al. design tools to assess the sensitivity of fairness measures to unmeasured confounding for a popular class of noise models [18].

Organization

The rest of this paper is organized as follows. Section 2 introduces the definitions used in this work. Section 3 states the main contributions and motivates our use of oracles. Section 4 describes our algorithms, and Section 5 concludes with a discussion of necessary assumptions.

2 Definitions

We consider a universe $\mathcal{X}$ of individuals, each represented by a vector of p attributes. We will assume each vector of attributes represents a unique individual. Determining whether or not the representation of the individuals is sufficient to permit fair classification is a fascinating topic beyond the reach of this paper; here we assume an affirmative answer. Since our work may be viewed as negative results, this assumption only strengthens the contribution.

A *classifier* maps individuals to $\{0,1\}$, $C : \mathcal{X} \to \{0,1\}$. It is often convenient to think of classifiers as vectors $c \in \{0,1\}^{|\mathcal{X}|}$, with $c_i \in \{0,1\}$ being the classification of the ith individual in some canonical ordering. We completely identify an individual and its index i, so we will often write $i \in \mathcal{X}$ to denote the ith individual in this ordering.

It is sometimes convenient to think of a classifier as partitioning $\mathcal{X}$ into two groups according to their classification outcomes. Unless otherwise specified, we use lower case letters to denote classifiers and the corresponding upper case letter to denote the partition. For a classifier c, we let $c^0 = \{i \in \mathcal{X} | c_i = 0\}$ and $c^1 = \{i \in \mathcal{X} | c_i = 1\}$. We sometimes refer to c^0 as the *Left Hand Side* of the partition c, denoted $\mathcal{LHS}(c)$, and c^1 as the *Right Hand Side*, denoted $\mathcal{RHS}(c)$. The *flip* of a partition is a swap of its left and right sides; in vector form, $c^{\text{flip}} = 1 - c$, *i.e.*, $\forall i \in \mathcal{X}, c_i^{\text{flip}} = 1 - c_i$. Constant classifiers have the property that for some $v \in \{0,1\}$, $c_i = v,\ \forall i \in \mathcal{X}$.

It is also sometimes convenient to think of individuals in $\mathcal{X}$ as vertices, and to think of the classifier as a two-coloring of the complete graph on $\mathcal{X}$ (see Figure 1). Monochromatic edges indicate pairs of individuals who are treated the same by the classifier.

Contexts and Valid Pairs

Many fairness notions require that classifiers be considered in some form of *context*. For example, in the case of counterfactual fairness the context is given by a causal model [6]. We therefore abstract the notion of a fairness definition $\mathcal{W}$ as a set of (context, classifier) pairs.

[6] See [1] for a general discussion of the need for context.

▶ **Definition 2** (validity). *If $(\mathcal{A}_\mathcal{W}, c) \in \mathcal{W}$ then $(\mathcal{A}_\mathcal{W}, c)$ is said to be a valid pair* under $\mathcal{W}$. *In our work $\mathcal{W}$ is typically fixed, in which case we may simply refer to* valid pairs.

Boundedness

We assume there is a procedure for enumerating all contexts, whose running time is a fixed function of $|\mathcal{X}|$. For example, we might consider the case in which the context is given by a causal graph constrained to have a number of vertices linear in p (the number of attributes) and the functions computed at each vertex can be described by circuits of size polynomial in $|\mathcal{X}|$. We note that without this assumption it is not even clear how to represent a context $\mathcal{A}_\mathcal{W}$ for the purposes of determining whether or not some $(\mathcal{A}_\mathcal{W}, c) \in \mathcal{W}$.

Oracles

We view the fairness definition $\mathcal{W}$ as a filter, and hypothesize the existence of an *oracle* to rule on the acceptability of valid pairs. A useful intuition, for example, with counterfactual fairness in mind, is that the oracle knows the true data generation model ${\mathcal{A}_\mathcal{W}}^*$, and is willing to accept exactly valid pairs $({\mathcal{A}_\mathcal{W}}^*, c) \in \mathcal{W}$; alternatively, the oracle may be willing to accept valid pairs $(\mathcal{A}_\mathcal{W}, c)$ whenever $\mathcal{A}_\mathcal{W}$ enjoys certain properties. However, we make no explicit assumptions: Formally, the oracle is specified by a subset of $\mathcal{W}$. It takes as input a valid pair $(\mathcal{A}_\mathcal{W}, t) \in \mathcal{W}$ and either accepts ($\mathcal{O}(\mathcal{A}_\mathcal{W}, t) = 1$) or rejects ($\mathcal{O}(\mathcal{A}_\mathcal{W}, t) = 0$).

▶ **Definition 3** (Strong Oracle). *A strong oracle is completely specified by the valid pairs that it accepts.*

It is convenient to name the collection of classifiers associated with acceptance by the strong oracle, that is, to define $\mathcal{T} = \{t \in \{0,1\}^{|X|} \mid \exists \mathcal{A}_\mathcal{W} :\ \mathcal{O}(\mathcal{A}_\mathcal{W}, t) = 1\}$.

Weak Oracles

Every weak oracle is a relaxation of a strong oracle. Weak oracles differ from their corresponding strong oracles by relaxation of the conditions for acceptance: weak oracles will accept whatever the associated strong oracles accept, but may also accept valid pairs.

▶ **Definition 4** (δ-Weak Oracle for Hamming distance). *Fix an arbitrary strong oracle $\mathcal{O}$ with associated classifiers $\mathcal{T}$. For $\delta > 0$ we say that $\mathcal{O}_\mathcal{H}$ is a* δ-weak oracle relaxation of $\mathcal{O}$, *based on the Hamming distance, if*

1. *$\mathcal{O}_\mathcal{H}$ accepts all valid pairs accepted by $\mathcal{O}$;*
2. *$\mathcal{O}_\mathcal{H}$ rejects valid pairs whose classifiers are far (in Hamming distance) from all classifiers in $\mathcal{T}$: Let $(\mathcal{A}_\mathcal{W}, c) \in \mathcal{W}$. If $\forall t \in \mathcal{T}, d_\mathrm{H}(c,t) > \delta$, then $\tilde{\mathcal{O}}(\mathcal{A}_\mathcal{W}, c) = 0$. Here, for $u, v \in \{0,1\}^{|\mathcal{X}|}$, $d_\mathrm{H}(u,v) := \{i \in \mathcal{X} \mid u_i \neq v_i\}$.*

On the remaining valid pairs, $\mathcal{O}_\mathcal{H}$ may behave arbitrarily.

The definition of a weak oracle *based on transportation cost* requires one additional concept.

▶ **Definition 5** (δ-faithfulness). *For $c, t \in \{0,1\}^{|\mathcal{X}|}$, we say that c is* δ-faithful *to t if*

$$\frac{1}{\binom{n}{2}} \sum_{i \neq j} \mathbf{1}\{t_i = t_j, c_i \neq c_j\} \leq \delta. \tag{1}$$

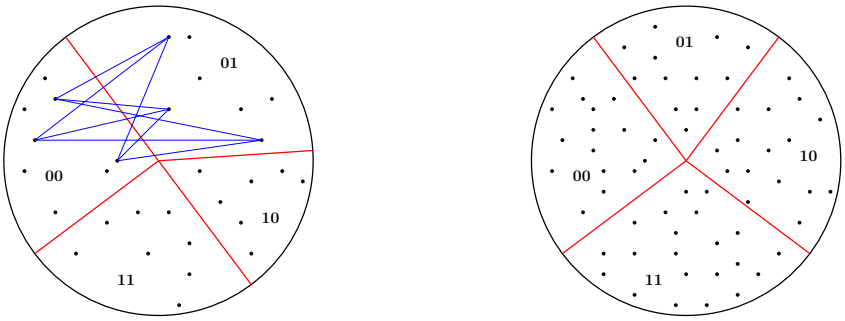

Figure 1 (Left) The (complete) labeled graph $G^{u \to v}$ for an ordered pair of classifiers (u, v). There is a vertex for each $i \in \mathcal{X}$. Vertices are labeled with two-bit strings indicating their classifications under u (first bit) and v (second bit). So vertices i in the quadrant $G_{00}^{u \to v}$ have $u_i = v_i = 0$, vertices in $G_{01}^{u \to v}$ have $u_i = 0$ and $v_i = 1$, and so on. The transportation cost $\mathcal{C}(u \to v)$ is captured by the number $|G_{00}^{u \to v}| \cdot |G_{01}^{u \to v}|$ of edges between the 00 and 01 quadrants (a few drawn in blue on left) plus the number of edges between the 11 and 10 quadrants (none drawn). For example, if $i \in G_{00}^{u \to v}$ and $j \in G_{01}^{u \to v}$ this says that the edge (i, j) was monochromatic (both vertices colored zero) in u but is polychromatic in v (because $v_i = 0 \neq v_j = 1$). (Right) Illustration of Assumption 14(2): all quadrants are substantial.

Note that faithfulness is not symmetric. Typically, we will consider faithfulness when c is a candidate classifier and t is an element of the set $\mathcal{T}$ associated with an oracle. δ-faithfulness suggests a natural transportation cost capturing the answer to the question, "Starting from t, how many monochromatic edges in t do we need to "break" when we transition to c?" We let $\mathcal{C}(t \to c)$ denote this transportation cost (Figure 1). This transportation cost is asymmetric and does not satisfy the triangle inequality.

▶ **Definition 6** (δ-neighborhood). *The* δ-neighborhood *of a classifier* $t \in \{0,1\}^{|\mathcal{X}|}$*, denoted* $\Gamma_\delta(t)$*, is the set of all* $c \in \{0,1\}^{|\mathcal{X}|}$ *such that* c *is* δ*-faithful to* t*.*

▶ **Definition 7** (δ-Weak Oracle for transportation cost). *Fix an arbitrary strong oracle* $\mathcal{O}$ *with associated classifiers* $\mathcal{T}$*. For* $\delta > 0$ *we say that* $\tilde{\mathcal{O}}$ *is a* δ-weak oracle relaxation of $\mathcal{O}$*, based on the transportation cost* $\mathcal{C}$*, if*

1. $\tilde{\mathcal{O}}$ *accepts all valid pairs accepted by* $\mathcal{O}$*;* $\forall(\mathcal{A}_\mathcal{W}, c)$ *such that* $\mathcal{O}(\mathcal{A}_\mathcal{W}, c) = 1$*,* $\tilde{\mathcal{O}}(\mathcal{A}_\mathcal{W}, c) = 1$*;*
2. $\tilde{\mathcal{O}}$ *rejects valid pairs whose classifiers are far (in transportation cost) from all classifiers in* $\mathcal{T}$*: Let* $(\mathcal{A}_\mathcal{W}, c) \in \mathcal{W}$*. If* $\forall t \in \mathcal{T}, c \notin \Gamma_\delta(t)$*, then* $\tilde{\mathcal{O}}(\mathcal{A}_\mathcal{W}, c) = 0$*.*

There are no further constraints on oracles other than being deterministic. Note that there may be many weak oracle relaxations of a given strong oracle $\mathcal{O}$.

3 Main Contributions

Our principle contributions are extraction procedures that recover the underlying truth held by oracles. Recall that a strong oracle is associated with a set $\mathcal{T}$ of classifiers. Formally, an *extraction procedure* is a program that, using only access to an oracle $\mathcal{O}$, outputs the list $\mathcal{T}$ associated with $\mathcal{O}$. Intuitively, one may imagine that this set arises from some ground truth provided by the concept $\mathcal{W}$. To illustrate, let $\mathcal{W}$ be the notion of counterfactual fairness and $\mathcal{M}$ the true causal model explaining actual functional relationships between all the relevant variables for the task. An oracle may believe that all classifiers that satisfy the counterfactual fairness definition with respect to this "true" causal model are indeed truly fair. Then, $\mathcal{T}$ equals the set of all counterfactually fair classifiers with respect to $\mathcal{M}$. The data analysts

have no knowledge of $\mathcal{M}$ whatsoever, but hope to learn about fair classifiers by interacting with the oracle. We provide algorithms that achieve this goal – starting with the simpler case of the strong oracle, subsequently moving on to weak oracles.

Recall that every strong oracle $\mathcal{O}$ has an associated set $\mathcal{T}$ of classifiers, such that $\forall t \in \mathcal{T}$, $\exists \mathcal{A}_{\mathcal{W}}$ with $(\mathcal{A}_{\mathcal{W}}, t) \in \mathcal{W}$ and $\mathcal{O}(\mathcal{A}_{\mathcal{W}}, t) = 1$; $\mathcal{O}$ rejects all valid pairs $(\mathcal{A}_{\mathcal{W}}, c)$ with $c \notin \mathcal{T}$. To begin with, we establish the following.

▶ **Theorem 8.** *For any fairness notion $\mathcal{W}$ satisfying the boundedness condition, and for any strong oracle $\mathcal{O}$ accepting a subset of $\mathcal{W}$, there exists an extraction procedure interacting with $\mathcal{O}$ whose running time is bounded by a function of $|\mathcal{X}|$. The output of the extraction procedure is the set $\mathcal{T}$ associated with $\mathcal{O}$.*

Under the assumption of bounded length contexts, Theorem 8 can be achieved simply via exhaustive search, since our extraction procedures are allowed to be inefficient. The primary contribution of this result is thus conceptual – that it is feasible to extract the ground truth from $\mathcal{O}$ under our framing of the problem. In the spirit of prior examples, if $\mathcal{W}$ is counterfactual fairness and $\mathcal{T}$ the set of all counterfactually fair classifiers with respect to the true causal model (which we do not have any access to, but let's say the oracle has complete knowledge about), then in principle one can learn all of these classifiers. Note that each of these is equivalent to a partitioning of the universe, and can therefore be viewed as a metric (albeit a simple one). Intuitively, one can interpret the oracle as a highly knowledgeable human expert with a deep understanding of the true underlying relationships between the variables relevant for the task, but who is unable to enunciate them – however, the expert is able to tell whether a classifier is fair or not by "looking" at it. Our result demonstrates that, given access to such an expert, a systematic strategy can successfully learn all the fair classifiers. Thus, in settings where learning the true causal model is extremely hard (if not impossible), and hence, reliably implementing counterfactual fairness (or any other causality-based notion) may be out of scope, our results suggest that developing efficient query models to interact with human experts suffices for fair classification, since these directly learn metric information from the expert (recall Remark 1).

While it is helpful to think of an all-knowing expert, who can accurately identify fair classifiers and task-appropriate contexts, our framework can be applied to *any* expert. We can also extract from an imperfect expert, *e.g.* one who can only reason about simple contexts, and accepts a subset of the fair classifiers (or even accepts some unfair ones!). The better the expert, the better the classifiers (or metrics) we extract will be.

Weak Oracles

A weak oracle accepts every valid pair accepted by a strong oracle; in addition, it rejects valid pairs whose classifiers are far from all classifiers in $\mathcal{T}$. However, a weak oracle may behave arbitrarily on the remaining pairs. Nonetheless, we are able to extract even when we do not know how the oracle will behave on these remaining pairs, and this is a strength of our framework. Since the oracle only provides fuzzy information, in the sense that it may behave at will on several pairs, we can only hope to recover $\mathcal{T}$ up to some error. Different notions of distance that determine what should be judged "far" lead to different instantiations of the weak oracle. The conversation around the right notion of distance lies beyond the scope of this paper. Here, we consider two notions, Hamming distance and transportation cost, and provide algorithms in each case that approximately recover $\mathcal{T}$.

Extraction from a Hamming distance based weak oracle

To begin with, we consider a natural distance measure – the Hamming distance. Recall that any weak oracle is a relaxation of a strong oracle $\mathcal{O}$ with an associated set $\mathcal{T}$ of classifiers. A weak oracle $\tilde{\mathcal{O}}$ based on the Hamming distance accepts any valid pair accepted by $\mathcal{O}$, and rejects a valid pair $(\mathcal{A}_{\mathcal{W}}, c)$ for which $d_{\mathrm{H}}(c,t) > \delta$ for every $t \in \mathcal{T}$, and otherwise behaves arbitrarily. Recovery of an approximation to the elements in $\mathcal{T}$ is then trivial (via exhaustive search) as long as any two elements $t, u \in \mathcal{T}$ satisfy $d_{\mathrm{H}}(t,u) > 4\delta$. (Details omitted.)

Extraction from a transportation cost based weak oracle

Our primary technical contribution is an extraction algorithm that approximately recovers elements of $\mathcal{T}$ from a weak oracle based on the transportation cost $\mathcal{C}(t \to c)$ (Definition 7). Theorem 9, stated next, says that the Sharp Extraction Algorithm (Algorithm 3, Section 4) produces a list of classifiers, each of which corresponds to a unique member of $\mathcal{T}$. Recall that for any classifier c, $c^0 = \{i \in \mathcal{X} | c_i = 0\}$ and $c^1 = \{i \in \mathcal{X} | c_i = 1\}$.

▶ **Theorem 9.** *Suppose $\mathcal{O}$ is a strong oracle with an associated set $\mathcal{T}$ of classifiers, and $\tilde{\mathcal{O}}$ is a δ-weak relaxation of $\mathcal{O}$ under the transportation cost $\mathcal{C}(t \to c)$ (Definition 7). Then under Assumption 14, the list of classifiers $(P_1, \ldots, P_m, Q_1, \ldots, Q_m)$ obtained from Sharp Extraction (Algorithm 3, Section 4) satisfies the following: Fix any index j. There exists $t \in \mathcal{T}$ such that*

$$P_j{}^0 \subset t^0 \quad (\text{or } \ t^1) \qquad \text{and} \qquad Q_j{}^1 \subset t^1 \quad (\text{resp. } t^0), \tag{2}$$

simultaneously,

$$|P_j{}^1 \cap t^0| \quad (\text{resp. } \ t^1) \leq \left(\frac{\tau_j - \sqrt{\tau_j^2 - 2\delta}}{2} \right) n, \quad \text{and}$$

$$|Q_j{}^0 \cap t^1| \quad (\text{resp. } \ t^0) \leq \left(\frac{\tilde{\tau}_j - \sqrt{\tilde{\tau}_j^2 - 2\delta}}{2} \right) n. \tag{3}$$

Above, $\tau_j = |t^0|/n$ and $\tilde{\tau}_j = 1 - \tau_j$. For classifiers P_j, Q_j and P_k, Q_k with different indices, the corresponding elements of $\mathcal{T}$ that satisfy the aforementioned property are also different.

Sharp Extraction precisely pins down the elements of $\mathcal{T}$ up to a small error margin as specified by (3). The smaller the value of δ, the lower the overall error. In general, for every $t \in \mathcal{T}$, Sharp Extraction recovers nearly as many members of t^0 and t^1 as possible (without recovering the exact classification outcomes). To see this, observe that the fraction of pairs of individuals that P_j erroneously splits in two groups, when they belong to the same group in the underlying element of $\mathcal{T}$, can be bounded by

$$\left(\frac{\tau_j - \sqrt{\tau_j^2 - 2\delta}}{2} \right) \left(\frac{\tau_j + \sqrt{\tau_j^2 - 2\delta}}{2} \right) + \left(\frac{\tilde{\tau}_j - \sqrt{\tilde{\tau}_j^2 - 2\delta}}{2} \right) \left(\frac{\tilde{\tau}_j - \sqrt{\tilde{\tau}_j^2 + 2\delta}}{2} \right) = \delta.$$

Since $\tilde{\mathcal{O}}$ is a δ-weak relaxation of $\mathcal{O}$, intuitively, one cannot hope to accurately cluster additional pairs of individuals from this weak oracle model, suggesting that Sharp Extraction achieves the best we may hope for in such a setting.

Extraction from a weak oracle based on a symmetrized transportation cost

Extraction algorithms can also be developed for weak oracles based on the symmetrized version of the transportation cost

$$\mathcal{C}^s(u \leftrightarrow v) := \frac{1}{\binom{n}{2}} \sum_{i \neq j} \left[\mathbf{1}\{u_i = u_j, v_i \neq v_j\} + \mathbf{1}\{u_i \neq u_j, v_i = v_j\} \right]. \tag{4}$$

A weak oracle $\tilde{\mathcal{O}}$ based on this notion accepts any valid pair accepted by its associated strong oracle $\mathcal{O}$, and rejects any valid pair $(\mathcal{A}_{\mathcal{W}}, c)$ for which $\mathcal{C}^s(t \leftrightarrow c) > \delta$ for every $t \in \mathcal{T}$. This case is, in fact, easier to handle than the asymmetric version. Thus, algorithms that work for weak oracles based on the asymmetric transportation cost can be simplified to suit the needs of a weak oracle based on the symmetrized version (4).

Conclusion

Classification algorithms cannot be evaluated for fairness without taking context into account. Several works in the fairness literature posit the existence of fair and wise human judges, and the ability of humans to recognize unfairness when they see it seems to be a linchpin of interpretability. We have explored what can be learned from a fairness oracle that evaluates (context, classifier) pairs satisfying a definition of fairness, accepting or rejecting according to a hypothesized fairness "truth". The oracle abstraction captures any human judge, or algorithm, or benchmark test; the extraction procedures described here do not need to "understand" the oracle's decisions. Even so, the procedures produce rudimentary metrics for the classification task at hand. The procedure can be amplified to improve the expressive power of the metric. These existence proofs are evidence for the conjecture that a metric is *always* at the heart of fairness.

Metrics can be combined with arbitrary loss functions to obtain individually fair classifiers satisfying a wide range of objectives [4]. Metrics learned on a sample of the population can sometimes be generalized to unseen examples [11]. An *efficient* metric extraction procedure would mean that it is essentially no harder to find a metric than to build good causal models and accompanying classifiers. This is an exciting direction for future research.

4 Extraction Algorithms

This section summarizes our extraction algorithms and key ingredients used therein.

▶ **Definition 10** (δ-Balanced classifier). *A classifier c is said to be* δ-balanced *if both $|c^0| > \sqrt{2\delta}|\mathcal{X}|$ and $|c^1| > \sqrt{2\delta}|\mathcal{X}|$.*

We let $\mathcal{B}$ denote the set of all δ-balanced classifiers of $\mathcal{X}$; henceforth, we simply call these the balanced classifiers.

Two classifiers are said to be aligned if they have relatively few disagreements. Recall that, for a classifier $c \in \{0,1\}^{|\mathcal{X}|}$, $c^0 = \{i \,|\, c_i = 0\}$ and c^1 is defined analogously.

▶ **Definition 11** (Close alignment). *Classifiers p and q in $\{0,1\}^{|\mathcal{X}|}$ are* in close alignment *if $|p^0 \cap q^1|, |q^0 \cap p^1|| \leq \sqrt{\frac{\delta}{2}}|\mathcal{X}|$.*

Furthermore, define the following sets (Figure 1) for any $v, c \in \{0,1\}^{|\mathcal{X}|}$,

$$G_{00}^{v \to c} = \{i \,|\, v_i = 0, c_i = 0\}, \qquad G_{01}^{v \to c} = \{i \,|\, v_i = 0, c_i = 1\}, \tag{5}$$

$$G_{10}^{v \to c} = \{i \,|\, v_i = 1, c_i = 0\}, \qquad G_{11}^{v \to c} = \{i \,|\, v_i = 1, c_i = 1\}. \tag{6}$$

The proof of Theorem 9 and the description of the algorithms require the following lemma.

▶ **Lemma 12.** *Let u, v be arbitrary balanced classifiers accepted by the weak oracle $\tilde{O}$ in Theorem 9. Then exactly one of the following must hold:*

1. *There exists $t \in \mathcal{T}$ such that $u, v \in \Gamma_\delta(t)$. In this case,*

$$\min\{|G_{00}^{u\to v}|, |G_{01}^{u\to v}|\} \leq \sqrt{2\delta}n \leq \min\{|G_{10}^{u\to v}|, G_{11}^{u\to v}\}. \qquad \text{Situation A}$$

2. *There does not exist any $t \in \mathcal{T}$ such that $u, v \in \Gamma_\delta(t)$. In this case, there must exist $t_1, t_2 \in \mathcal{T}$ such that $t_1 \neq t_2^{\text{flip}}$, $u \in \Gamma_\delta(t_1), v \in \Gamma_\delta(t_2)$, and that at least one of the following holds*

$$\min\{|G_{00}^{u\to v}|, |G_{01}^{u\to v}|\} > \sqrt{2\delta}n, \qquad \min\{|G_{10}^{u\to v}|, G_{11}^{u\to v}\} > \sqrt{2\delta}n. \qquad \text{Situation B}$$

Our main algorithm, Sharp Extraction, builds on Fuzzy Extraction, presented in Algorithm 1.

Informal Description of Fuzzy Extraction Algorithm

The fuzzy extraction algorithm seeks to associate candidate classifiers c with elements of $\mathcal{T}$, roughly, guided by the transportation cost $\mathcal{C}(t \to c)$. In particular, the algorithm aims to recover $\Gamma_\delta(t)$ for each $t \in \mathcal{T}$. As with the reconstruction from a strong oracle, by the boundedness requirement for contexts, we can find $\boldsymbol{V} = \{c \in \{0,1\}^{|\mathcal{X}|} \mid \exists \mathcal{A}_\mathcal{W} : \tilde{\mathcal{O}}(\mathcal{A}_\mathcal{W}, c) = 1\}$ by enumeration. The algorithm starts by finding $\boldsymbol{V}$. The algorithm then prunes out all unbalanced classifiers, setting $\boldsymbol{V}_\mathcal{B} = \boldsymbol{V} \cap \mathcal{B}$. This is the starting point for recovering $\mathcal{T}$, the classifiers associated with the strong oracle $\mathcal{O}$ of which $\tilde{\mathcal{O}}$ is a relaxation.

At a high level, Fuzzy Extraction works as follow: for an arbitrary pair $u, v \in \boldsymbol{V}_\mathcal{B}$, the algorithm checks whether u and v are both in $\Gamma_\delta(t)$ for some $t \in \mathcal{T}$. From Lemma 12, this can be detected simply by inspection, that Situation A holds. If so, the algorithm clusters u and v into the same group, and otherwise, to different groups. In this manner, the algorithm builds up a collection of sets, each of which contains classifiers that are all in $\Gamma_\delta(t)$ for some $t \in \mathcal{T}$.

In addition, the algorithm takes special care to track when $u, v \in \boldsymbol{V}_\mathcal{B}$ are in close alignment (Definition 11; roughly speaking, they are closely aligned if they are close in Hamming distance). Using this information, the algorithm builds a collection of sets, which we call *orbits*, such that each set contains classifiers in close alignment with some $t \in \mathcal{T}$ (or with t^{flip}, where $t \in \mathcal{T}$). Thus, for each $t \in \mathcal{T}$, Fuzzy Extraction produces (1) an orbit containing t and elements in close alignment with t, and, (2) an orbit consisting of elements in close alignment with t^{flip} (but not necessarily containing t^{flip}).

Implications of Fuzzy Extraction Algorithm

The Fuzzy Extraction algorithm teases apart whether any two balanced accepted classifiers belong to the δ-neighborhood of the same or different elements of $\mathcal{T}$. Note that, $\tilde{\mathcal{O}}$ provides relatively vague information – there could be a large number of valid pairs on which $\tilde{\mathcal{O}}$ behaves arbitrarily. In the full paper, we establish that despite such imprecise information, Fuzzy Extraction distinguishes δ-neighborhoods of different elements of $\mathcal{T}$ successfully, and recovers all balanced accepted members of the neighborhoods.

Algorithm 1 Fuzzy Extraction.

input : The universe $\mathcal{X}$ and an oracle $\tilde{\mathcal{O}}$.
output : A collection of $2|\mathcal{T}|$ disjoint subsets of $\{0,1\}^{|\mathcal{X}|}$.

Find $\boldsymbol{V} = \{c \in \{0,1\}^{|\mathcal{X}|} \mid \exists \mathcal{A}_{\mathcal{W}} : \ \tilde{\mathcal{O}}(\mathcal{A}_{\mathcal{W}}, c) = 1\}$. Construct $\boldsymbol{V}_{\mathcal{B}} = \boldsymbol{V} \cap \mathcal{B}$. Set $\ell = 1$;
while $\boldsymbol{V}_{\mathcal{B}} \neq \phi$ **do**
  Choose a classifier c_ℓ from $\boldsymbol{V}_{\mathcal{B}}$ and set $\mathbf{Orb}_\ell := \{c_\ell\}$. If $c_\ell^{\text{flip}} \in \boldsymbol{V}_{\mathcal{B}}$, set $\mathbf{Orb}'_\ell = \{c_\ell^{\text{flip}}\}$, else set $\mathbf{Orb}'_\ell = \phi$;
  for $c \in \boldsymbol{V}_{\mathcal{B}} \backslash \{c_\ell\}$ **do**
    if CheckSituationA(c_ℓ, c) = "Situation A holds" **then**
      if $G_{00}^{c_\ell \to c} > \sqrt{2\delta}\mathcal{X} \geq G_{01}^{c_\ell \to c}$ **then**
        update

$$\mathbf{Orb}_\ell = \mathbf{Orb}_\ell \cup \{c\}, \ \boldsymbol{V}_{\mathcal{B}} = \boldsymbol{V}_{\mathcal{B}} \backslash \{c\},$$

        and if $c^{\text{flip}} \in \boldsymbol{V}_B$, update

$$\mathbf{Orb}'_\ell = \mathbf{Orb}'_\ell \cup \{c^{\text{flip}}\}, \boldsymbol{V}_{\mathcal{B}} = \boldsymbol{V}_{\mathcal{B}} \backslash \{c^{\text{flip}}\};$$

      else
        update

$$\mathbf{Orb}'_\ell = \mathbf{Orb}'_\ell \cup \{c\}, \ \boldsymbol{V}_{\mathcal{B}} = \boldsymbol{V}_{\mathcal{B}} \backslash \{c\},$$

        and if $c^{\text{flip}} \in \boldsymbol{V}_{\mathcal{B}}$, update

$$\mathbf{Orb}_\ell = \mathbf{Orb}_\ell \cup \{c^{\text{flip}}\}, \boldsymbol{V}_{\mathcal{B}} = \boldsymbol{V}_{\mathcal{B}} \backslash \{c^{\text{flip}}\}.$$

      end
    end
  end
  Set $\boldsymbol{V}_{\mathcal{B}} = \boldsymbol{V}_{\mathcal{B}} \backslash \{c_\ell\}$, $\ell = \ell + 1$;
end
Return $\mathbf{Orb}_1, \ldots, \mathbf{Orb}_{\ell-1}, \mathbf{Orb}_1, \ldots, \mathbf{Orb}'_{\ell-1}$, and additional sets $\mathbf{Orb}_\ell = \{(c_i = 1, \forall i \in \mathcal{X})\}$ $\mathbf{Orb}'_\ell = \{(c_i = 0, \forall i \in \mathcal{X})\}$, whenever a constant classifier is in $\boldsymbol{V}$.

Algorithm 2 CheckSituationA: Checks whether Situation A from Lemma 12 holds.

input : An ordered pair of classifiers $(c, u) \in \{0,1\}^{|\mathcal{X}|}$.
output : Either "Situation A holds" or "Situation A does not hold".

if $\min\{|G_{00}^{c \to u}|, |G_{01}^{c \to u}|\} \leq \sqrt{2\delta}\mathcal{X} < \max\{|G_{00}^{c \to u}|, |G_{01}^{c \to u}|\}$
and
$\min\{|G_{10}^{c \to u}|, |G_{11}^{c \to u}|\} \leq \sqrt{2\delta}\mathcal{X} < \max\{|G_{10}^{c \to u}|, |G_{11}^{c \to u}|\}$ **then**
  return "Situation A holds"
else
  "Situation A does not hold"
end

Intuition for the Sharp Extraction Algorithm

We now focus on the Sharp Extraction algorithm. Fix a strong oracle $\mathcal{O}$ with associated set $\mathcal{T}$ of classifiers, and let $\tilde{\mathcal{O}}$ be a δ-weak relaxation of $\mathcal{O}$. Run Algorithm 1 with weak oracle $\tilde{\mathcal{O}}$ to obtain a collection of orbits.

For this informal discussion, fix $t \in \mathcal{T}$ and let **Orb** be the orbit from Algorithm 1 that contains t, possibly together with some classifiers in close alignment with t. Algorithm 3 applies a screening procedure to the elements of each orbit. Applied to the members of **Orb**, the procedure may screen out some $u \in$ **Orb**, meaning that it determines definitively that $u \notin \mathcal{T}$, but other vectors $v \in$ **Orb** may remain; in particular, t will remain.

The screening procedure invokes a primitive MergeOnes (Definition 13) with the key property that $\forall v \in$ **Orb**, MergeOnes$(t, v) \in$ **Orb**. Thus, to test if $v \in$ **Orb** is a valid candidate for t, the algorithm tests whether MergeOnes$(u, v) \in$ **Orb** for all $u \in$ **Orb**.

Algorithm 3 Sharp Extraction.

input : The universe $\mathcal{X}$ and the $\tilde{\mathcal{O}}$ considered in Theorem 9.
output : A list $P_1, \ldots, P_m, Q_1, \ldots, Q_m$ of classifiers of the universe $\mathcal{X}$.

Run Fuzzy Extraction with the inputs $\mathcal{X}$ and $\tilde{\mathcal{O}}$. Let $\boldsymbol{Orb}_1, \ldots, \boldsymbol{Orb}_m, \boldsymbol{Orb}'_1, \ldots, \boldsymbol{Orb}'_m$ *denote the output. For $j = 1, \ldots, m$, define* $\Pi_j, \Pi'_j, \Gamma_j, \Gamma'_j = \phi$. ;

for $j = 1, \ldots, m$ **do**
  for $c \in \boldsymbol{Orb}_j$ **do**
    if for all $c' \in \mathbf{Orb}_j$, MergeOnes$(c, c') \in \mathbf{Orb}_j$, update $\Pi_j = \Pi_j \cup c$;
    if for all $c' \in \mathbf{Orb}_j$, MergeZeros$(c, c') \in \mathbf{Orb}_j$, update $\Gamma_j = \Gamma_j \cup c$;
  end
  for $c' \in \boldsymbol{Orb}'_j$ **do**
    if for all $c'' \in \mathbf{Orb}'_j$, MergeZeros$(c', c'') \in \mathbf{Orb}'_j$, update $\Pi'_j = \Pi'_j \cup c'$;
    if for all $c'' \in \mathbf{Orb}'_j$, MergeOnes$(c', c'') \in \mathbf{Orb}'_j$, update $\Gamma'_j = \Gamma'_j \cup c'$;
  end
end
for $j = 1, \ldots, m$ **do**
  set $u_j =$ MergeOnes$(\Pi_j), v_j =$ MergeZeros$(\Pi'_j), w_j =$ MergeZeros$(\Gamma_j), x_j =$ MergeOnes(Γ'_j), and define

$$P_j = \begin{cases} u_j & \text{if } |{u_j}^0| \geq |{v_j}^1| \\ v_j^{\text{flip}} & \text{if } |{v_j}^1| > |{u_j}^0| \end{cases}, \qquad Q_j = \begin{cases} w_j & \text{if } |w_j^1| \geq |x_j^0| \\ x_j^{\text{flip}} & \text{if } |x_j^0| > |{w_j}^1| \end{cases}.$$

end
Return $P_1, \ldots, P_m, Q_1, \ldots, Q_m$;
// The classifier c with ${P_j}^0 \subseteq c^0$ and ${Q_j}^1 \subseteq c^1$ accurately recovers t^0, t^1 (or $[t^{\text{flip}}]^0, [t^{\text{flip}}]^1$) upto a small error margin (Theorem 9).

Let **In** $\subseteq$ **Orb** be the subset of **Orb** screened in. Let us arbitrarily name these elements $u_1, u_2, \ldots, u_k$. If there is exactly one element in **In**, then $u_1 = t$. This is excellent: the algorithm found an element of $\mathcal{T}$. Consider now the more general case of multiple elements. Choose any ordering of **In**, say, $(u_1, u_2, \ldots, u_k)$. Then since $u_1 \in$ **In**, we have $w =$ MergeOnes$(u_1, u_2) \in$ **Orb**. Now, since $u_3 \in$ **In**, MergeOnes$(w, u_3) \in$ **Orb**. We can continue in this way until we have merged in u_k, and we see by induction that at every step the merge remains in **Orb**.

MergeOnes is commutative and associative, so we can assume without loss of generality that $u_1 = t$, which will simplify the description of the properties of the merging of all the classifiers in **In**. We first define the operation.

▶ **Definition 13** (MergeOnes)**.** *For a set of classifiers $\mathcal{C} = \{c_1, \ldots, c_k\}$, define the* MergeOnes *operation applied to $\mathcal{C}$ by* MergeOnes$(\{c_1, \ldots, c_k\})$ *to be a new classifier w such that*

$$w^1 = \cup_{j \in [k]} {c_j}^1, \qquad w^0 = \mathcal{X} \backslash w^1 \tag{7}$$

MergeZeros *is defined analogously and denoted* MergeZeros$(\{c_1, \ldots, c_k\})$.

Let $w = \text{MergeOnes}(t, u)$ for an arbitrary $u \in \{0,1\}^{|\mathcal{X}|}$. Then w^1 contains all the elements with positive classification under t (and other elements that are positive under u). Since w^0 contains none of these, we have that $w^0 \subseteq t^0$.

Let $w = \text{MergeOnes}(\{u \in \mathbf{In}\})$. Then by the above reasoning also for this w, we have $w^0 \subseteq t^0$. To argue that w^0 recovers most of t^0, we show that $t^0 \cap w^1$ must be small. This follows from the fact that $w \in \mathbf{Orb}$, as argued above. This ensures that $t^0 \cap w^0$, is in fact, large. A similar reasoning applies to the MergeZeros procedure, which is used to create another subset $\mathbf{In}' \subseteq \mathbf{Orb}$ with the property that $\text{MergeZeros}(\mathbf{In}')^1$ has a large intersection with t^1.

In its final phase, still focusing on a single t and its corresponding orbit **Orb**, the algorithm combines the left side of the output of the MergeOnes procedure and the right side of the output of the MergeZeros procedure to create a classifier that substantially agrees with t (for elements $i \notin (\text{MergeOnes}(\mathbf{In}))^0 \cup (\text{MergeZeros}(\mathbf{In}'))^1$ it assigns an arbitrary value). This completes the high level intuition for the Sharp Extraction algorithm. Theorem 9 provides the formal guarantees.

5 Discussion on Assumptions

Our main theorem relies on the following crucial assumption regarding the structure of $\mathcal{T}$.

▶ **Assumption 14.** *Assume that every $t \in \mathcal{T}$ obeys the following structure.*

1. *Every $t \in \mathcal{T}$ that is not a constant classifer, must be 4δ-balanced.*
2. *For any $t, u \in \mathcal{T}$, if $t^{\text{flip}} \neq u$, then at most one of $G_{00}^{t \to u}, G_{01}^{t \to u}, G_{10}^{t \to u}$ and $G_{11}^{t \to u}$ can be empty. Furthermore, whenever one of these sets is non-empty, it must contain strictly more than $2\sqrt{2\delta}\mathcal{X}$ elements (Figure 1).*
3. *For every $t \in \mathcal{T}$ such that $t^{\text{flip}} \notin \mathcal{T}$, let $\tilde{\mathcal{O}}(\mathcal{A}_{\mathcal{W}}, c) = 0$ whenever c is in close alignment with t^{flip}.*

We provide some intuition for the assumptions, deferring details to the full paper. Note that, if $t \in \mathcal{T}$ is imbalanced, then most $c \in \Gamma_\delta(t)$ will also be imbalanced. However, imbalanced classifiers are problematic: they belong to δ-neighborhoods of more than one $t \in \mathcal{T}$, even when these are far apart. The presence of several imbalanced classifiers confuses our algorithms. By requiring that all $t \in \mathcal{T}$ be well balanced (Assumption 14(1)), we ensure that sufficiently many $c \in \Gamma_\delta(t)$ are also balanced.

To recover $\mathcal{T}$, our algorithms need to tease apart the following situations for any two classifiers $u, v \in \{0,1\}^{|\mathcal{X}|}$: (a) $\exists t \in \mathcal{T}$ such that $u, v \in \Gamma_\delta(t)$, and (b) $\nexists t \in \mathcal{T}$ such that $u, v \in \Gamma_\delta(t)$. Our intuition is that if elements of $\mathcal{T}$ are well-separated as defined via transportation costs, then this should be possible; however, our proof requires Assumption 14(2) that implies separation but is not equivalent. We do not know if this stronger condition can be relaxed.

Finally, for any $t \in \mathcal{T}$, the flipped classifier t^{flip} may or may not belong to $\mathcal{T}$. But, the neighborhoods of t and t^{flip} are identical – this creates additional challenges when $t \in \mathcal{T}$ but $t^{\text{flip}} \notin \mathcal{T}$. Assumption 14(3) protects from complications arising in this case.

References

1 Benjamin R Baer, Daniel E Gilbert, and Martin T Wells. Fairness criteria through the lens of directed acyclic graphical models. *arXiv preprint*, 2019. `arXiv:1906.11333`.

2 Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.

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

4 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.

5 Cynthia Dwork and Christina Ilvento. Fairness under composition. In *10th Innovations in Theoretical Computer Science Conference, ITCS 2019, January 10-12, 2019, San Diego, California, USA*, pages 33:1–33:20, 2019.

6 Cynthia Dwork, Michael P Kim, Omer Reingold, Guy N Rothblum, and Gal Yona. Learning from outcomes: Evidence-based rankings. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 106–125. IEEE, 2019.

7 Harrison Edwards and Amos Storkey. Censoring representations with an adversary. *arXiv preprint*, 2015. `arXiv:1511.05897`.

8 Stephen Gillen, Christopher Jung, Michael Kearns, and Aaron Roth. Online learning with an unknown fairness metric. In *Advances in Neural Information Processing Systems*, pages 2600–2609, 2018.

9 Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

10 Úrsula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1944–1953, 2018.

11 Christina Ilvento. Metric learning for individual fairness. *arXiv preprint*, 2019. `arXiv:1906.00250`.

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

13 Christopher Jung, Sampath Kannan, Changwa Lee, Mallesh M. Pai, Aaron Roth, and Rakesh Vohra. Fair prediction with endogenous behavior. Manuscript shared with authors.

14 Christopher Jung, Michael Kearns, Seth Neel, Aaron Roth, Logan Stapleton, and Zhiwei Steven Wu. Eliciting and enforcing subjective individual fairness. *arXiv preprint*, 2019. `arXiv:1905.10660`.

15 Faisal Kamiran and Toon Calders. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*, pages 1–6. IEEE, 2009.

16 Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 643–650. IEEE, 2011.

17 Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2569–2577, 2018.

18 Niki Kilbertus, Philip J Ball, Matt J Kusner, Adrian Weller, and Ricardo Silva. The sensitivity of counterfactual fairness to unmeasured confounding. *arXiv preprint*, 2019. `arXiv:1907.01040`.

19 Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, pages 656–666, 2017.

20 Michael Kim, Omer Reingold, and Guy Rothblum. Fairness through computationally-bounded awareness. In *Advances in Neural Information Processing Systems*, pages 4842–4852, 2018.

21 Jon 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)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

22 Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076, 2017.

23 Himabindu Lakkaraju. Course notes for compsci 282br, harvard university: Interpretability and explainability in machine learning, 2019.

24 Zachary C Lipton. The mythos of model interpretability. *Queue*, 16(3):31–57, 2018.

25 David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. *arXiv preprint*, 2018. `arXiv:1802.06309`.

26 Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

27 Roland Neil and Christopher Winship. Methodological challenges and opportunities in testing for racial discrimination in policing. *Annual Review of Criminology*, 2:73–98, 2019.

28 Judea Pearl. *Causality*. Cambridge university press, 2009.

29 Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 560–568, 2008.

30 Gal Yona and Guy N. Rothblum. Probably approximately metric-fair learning. In *ICML*, 2018.

31 Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 325–333, 2013.