From the Real Towards the Ideal: Risk Prediction in a Better World

Cynthia Dwork
Harvard University, Cambridge, MA, USA

Omer Reingold
Stanford University, CA, USA

Guy N. Rothblum
Apple, Cupertino, CA, USA

Abstract

Prediction algorithms assign scores in $[0,1]$ to individuals, often interpreted as "probabilities" of a positive outcome, for example, of repaying a loan or succeeding in a job. Success, however, rarely depends only on the individual: it is a function of the individual's interaction with the environment, past and present. Environments do not treat all demographic groups equally.

We initiate the study of corrective transformations τ that map predictors of success in the real world to predictors in a better world. In the language of algorithmic fairness, letting p^* denote the true probabilities of success in the real, unfair, world, we characterize the transformations τ for which it is feasible to find a predictor $\tilde{q}$ that is indistinguishable from $\tau(p^*)$. The problem is challenging because we do not have access to probabilities or even outcomes in a better world. Nor do we have access to probabilities p^* in the real world. The only data available for training are outcomes from the real world.

We obtain a complete characterization of when it is possible to learn predictors that are indistinguishable from $\tau(p^*)$, in the form of a simple-to-state criterion describing necessary and sufficient conditions for doing so. This criterion is inextricably bound with the very existence of uncertainty.

2012 ACM Subject Classification Theory of computation → Theory and algorithms for application domains

Keywords and phrases Algorithmic Fairness, Affirmative Action, Learning, Predictions, Multicalibration, Outcome Indistinguishability

Digital Object Identifier 10.4230/LIPIcs.FORC.2023.1

Funding *Cynthia Dwork*: Partially supported by Microsoft, by the Simons Foundation Collaboration on the Theory of Algorithmic Fairness and a Sloan Foundation Grant 2020-13941.
Omer Reingold: Part of this work was done while the author was visiting Microsoft Research. Supported by the Simons Foundation Investigators Award 689988, Simons Foundation Collaboration on the Theory of Algorithmic Fairness and a Sloan Foundation Grant 2020-13941
Guy N. Rothblum: Part of this work was done while the author was at the Weizmann Institute of Science and while visiting Microsoft Research. 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), from the Israel Science Foundation (grant number 5219/17), from the U.S-Israel Binational Science Foundation (grant No. 2018102) and from the Simons Collaboration on The Theory of Algorithmic Fairness.

1 Introduction

Prediction algorithms assign scores in $[0,1]$ to individuals, often interpreted as "probabilities" of a positive outcome, for example, of repaying a loan or succeeding in a job. Success, however, rarely depends only on the individual: it is a function of the individual's interaction with the

© Cynthia Dwork, Omer Reingold, and Guy N. Rothblum;
licensed under Creative Commons License CC-BY 4.0

4th Symposium on Foundations of Responsible Computing (FORC 2023).
Editor: Kunal Talwar; Article No. 1; pp. 1:1–1:17
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

environment, past and present. If we think of an individual x as a collection of features, *past* interaction affects those very features; that is, the accomplishments that individuals bring to a potential new job depend heavily on the opportunities afforded to the them and their families in the past. In addition, given a collection of features x, an individual's chance of a positive outcome depends heavily on the *future* environment in which the individual will be operating; for example, a woman with a given degree of talent and experience is less likely to succeed at a news organization that is hostile to women than at an organization supportive of women.

We initiate the study of corrective transformations τ that map predictors of success in the real world to predictors in a better world. In the language of algorithmic fairness, letting p^* denote the true probabilities of success in the real, unfair, world, we characterize the transformations τ for which it is feasible to find a predictor $\tilde{q}$ that is indistinguishable from $\tau(p^*)$. The problem is challenging because we do not have access to probabilities or even outcomes in a better world. Nor do we have access to probabilities p^* in the real world. The only data available for training are outcomes from the real world.

The meaning of a "probability" for a non-repeatable event is the subject of much debate [1], giving rise to the question of what we should want from an *ideal* scoring function. In one view, known as *Outcome Indistinguishability*, the scores offer a model for the real world, and we want the modeled world to be indistinguishable from the real world; this leads to a hierarchy of demands, according to the degree of access to the scoring function that is granted to the distinguisher [3]. A different, but compatible, view arises from the perspective of algorithmic fairness. Speaking informally, a scoring function is multi-calibrated with respect to a collection $\mathcal{C}$ of arbitrarily intersecting subsets of the population if it is calibrated simultaneously on each $S \in \mathcal{C}$ when viewed in isolation [6]. The sets in $\mathcal{C}$ need not be restricted to the demographic groups often described as "protected sets," but can (and should) capture conditions that are predictive of positive or negative outcomes. With this flexibility in mind, it is perhaps not surprising that multi-calibration has been shown to be equivalent to the second level of the outcome-indistinguishability hierarchy [3]. We use the term "MC/OI" to denote these equivalent properties.

Happily, MC/OI predictors can be learned from real-world Boolean outcomes data $o^*(x) \sim \mathrm{Ber}(p^*(x))$, without access to p^* [6]. Now, consider a corrective transformation τ mapping individual-score pairs $(x, p^*(x))$ to $[0,1]$, where the intuition is that $q^*(x) = [\tau(p^*)](x)$ is the probability of a positive outcome in a better world for the individual whose features in the real world are given by x. Not only do we not have access to q^*, but we do not even have outcomes data for the better world – that world does not exist! How, then, can we hope to construct a predictor that is indistinguishable from q^*? That is the problem studied in this work: for what kinds of corrective transformations τ can we obtain a predictor $\tilde{q}$ that is MC/OI with respect to q^*?

Taxonomy of transformations. We consider three kinds of corrective transformations. The conceptually simplest is fully deterministic transformations τ that are specified with no access to the underlying distribution $\mathcal{D}^*$. Due to the deterministic nature of the transformation, the transformed predictor $\tau(p)$ is completely and uniquely defined for any given predictor p. For example, the transformation that raises scores for members of a set S, setting $[\tau(p^*)](x) = \min\{p^*(x) + 0.2, 1\}$ for $x \in S$, is fully deterministic.

More generally, we consider *parameterized* transformations τ_π, where the parameters π are obtained via an efficient parameter-learning algorithm that operates on instance-outcome samples $(x, o^*(x))$ for $x \sim \mathcal{D}_\mathcal{X}$, where $o^*(x) \sim \mathrm{Ber}(p^*(x))$. Here we must be careful in defining $\tau_\pi(p)$, as different randomness – in the samples seen by the parameter-learner and in random coins it may use – will lead to different choices of π. We also allow the resulting transformation τ_π to be randomized. We informally and implicitly cover all these sources of randomness when we say that the transformation is randomized.

For example, suppose we have disjoint groups A and B and the goal of the transformation is to ensure statistical parity, so that in the transformed world the probabilities of a positive outcome are equalized between the two groups. The exact transformation depends the disparity in the real world, p^*, between the two group, *i.e.*, the difference between $p_A \overset{\text{def}}{=} \mathbf{E}_{x\in A}[p^*(x)] = \mathbf{E}_{x\in A}[o^*_x]$, and $p_B \overset{\text{def}}{=} \mathbf{E}_{x\in B}[p^*(x)] = \mathbf{E}_{x\in B}[o^*_x]$. Both of these quantities can be estimated from real-world outcomes data during the parameter-learning phase, and from these one can approximately determine $\alpha \in [0,1] = \frac{p_A - p_B}{1-p_B}$ such that the transformation τ_α that leaves scores unchanged for members of A and sets the new score for members $x \in B$ to $[\tau_\alpha(p^*)](x) = \alpha + (1-\alpha)p^*(x)$ satisfies $E_{x\in A}[(\tau(p^*))(x)] \approx E_{x\in B}[(\tau(p^*))(x)]$.

In a third type of transformation the parameter-learner $\mathcal{L}$ has access to p^*. For example, consider a population with two disjoint subgroups S, T. A predictor achieves *balance for the positive class* [9] if the average score assigned to positive instances in S equals the average score assigned to positive instances in T. Now, consider a transformation that takes an arbitrary predictor p as input and produces a transformed $\tau(p)$ satisfying the balance condition. To do this, the parameter-learner needs access to the average p^* values for the members of T and of S. For example, suppose that $\forall x \in T, p^*(x) = 0.8$, and $\forall x \in S, p^*(x) = 0.2$. Ensuring balance for the positive class can then be achieved by setting $[\tau(p^*)](x) = 0.8$ for all members of S and setting $[\tau(p^*)](x) = p^*(x)$ for all members of T. Of course, our algorithms cannot have access to p^*, but the prospect of building a predictor that is multicalibrated with respect to $\tau(p^*)$ remains compelling.

Canonical transformed predictor. When the transformation is randomized, we cannot simply speak of $\tau(p^*)$, as this is a random variable. However, given all the sources of randomness and an initial predictor p, the expectation of the transformation $\tau(p)$, $\mathbb{C}[\tau(p)] \overset{\text{def}}{=} \mathbf{E}[\tau(p)]$, where the expectation is taken over the samples fed to the parameter-learner, as well as it randomness, and any randomness in the transformed predictor, is well defined. We refer to this as the *canonical* transformed predictor, and use the special symbol $\mathbb{C}$.

Uncertainty and randomized instantiations. A deep and unresolved question is whether uncertainty exists, or if instead it only appears to exist because of insufficient information about the state of the world and insufficient computing power to determine future outcomes. Thus, when we talk about real-life probabilities $p^*(x)$, we cannot know whether $p^*(x)$ must lie in $\{0,1\}$ (determinism) or whether values in $(0,1)$ are possible (uncertainty). In the real world, we only observe outcomes, not individual probabilities. If uncertainty exists, then real-world outcomes are consonant with a deterministic world p^{**} that is a specific *random instantiation* of the real-world probabilities p^* in which each x is assigned a probability $p^{**}(x) \sim \mathrm{Ber}(p^*(x)) \in \{0,1\}$.

If uncertainty exists, there are many different possible random instantiations of p^*. The central concept in a transformation τ is its robustness (or not) to random instantiations: Does $\mathbb{C}[\tau(p^*)]$ look like $E_{p^{**}\leftarrow \mathsf{RI}(p^*)}[\mathbb{C}[\tau(p^{**})]]$? For example, are their average values, over elements in a large set S, close in expectation? Example 1 above, in which scores of members of S are increased by 0.2 but capped at 1, is *not* robust to random instantiations. To see this, consider two possible choices of the real world p^*. In the first, $p^*_1(x) = 1/2$ for all $x \in S$; in the second, $p^*_2(x) = 0$ for a random half of the $x \in S$ and $p^*_2(x) = 1$ for the remainder of S. Note that p_2^* is a random instantiation of p_1^*. The average scores for members of S are different under these two transformations: $\mathbf{E}_{x \in S}[(\tau(p_1^*))(x)] = 0.7$, but $\mathbf{E}_{x \in S}[(\tau(p_2^*))(x)] = 0.6$. The Balance for the Positive Class transformation described above also fails to be robust to random instantiations; in a nutshell, this is because in a random instantiation there is no uncertainty, and all positive members of S have $p^{**}(x) = 1$.

In contrast, the parameterized statistical parity transformation described above *is* robust to random instantiations. Roughly speaking this is because every random instantiation of p^* yields (almost) the same value of the parameter α, and for any large set S the average value $\mathbf{E}_{x \in S}[(\tau(p^*))(x)] \approx \mathbf{E}_{x \in S}[(\tau(p^{**}))(x)]$ depends only on α and the expectations $E_{x \in S \cap A}[p^*(x)]$ and $E_{x \in S \cap B}[p^*(x)]$. These expectations are invariant under random instantiations (assuming the sizes of $S \cap A, S \cap B$ are sufficiently large).

It is mathematically impossible, given only real-world instance-outcome pairs, to distinguish a real-world p^* in which probabilities are real-valued (uncertainty exists) and a real world which is a random instantiation p^{**} of such a p^* (no uncertainty), an epistemic state of affairs we summarize as follows.
Unresolvability Axiom: The question of whether uncertainty exists cannot be resolved by computing on finitely many samples from $\mathcal{D}^*$.

A Complete Characterization. Quite surprisingly, the concept of robustness to random instantiations provides a complete characterization of when it is possible to learn predictors that are indistinguishable from $q^* = \tau(p^*)$:

▶ **Theorem 1** (Main Theorem – informal). *There is a multiaccurate learning algorithm, and a multi-calibrated learning algorithm, with respect to $q^* = \tau(p^*)$, if and only if τ is robust to random instantiations.*

Thus, not only is it sometimes possible to build predictors for a transformed world, but there is a simple-to-state criterion describing necessary and sufficient conditions for doing so, and this criterion is inextricably bound with the very existence of uncertainty.

To prove sufficiency, we show how to exploit robustness to random instantiation to create samples of outcomes in the better world of q^*. This sample generation process involves partitioning samples from $\mathcal{D}^*$ into groups, viewing each group as samples from an independent random instantiation of p^*, and using these capture, on average, the behavior of $\mathbf{E}_{p^{**} \leftarrow \mathsf{RI}(p^*)}[\mathbb{C}[\tau(p^{**})]]$. By employing known algorithms we can build the desired predictors using these samples. We note that at no point does our multicalibration algorithm have access to the probabilities p^* or q^*; everything is done given access only to *real-world outcomes data.*

To prove necessity, we argue that any transformation that is not robust to random instantiations must behave very differently on p^* than it behaves on random instantiations $p^{**} \leftarrow \mathsf{RI}(p^*)$. In principle, this is detectable (although not efficiently!), which would resolve the question of whether uncertainty exists, contradicting the unresolvability axiom.

Stability. A final important *stability* notion tells us when multicalibration with respect to the transformed world $q^* = \tau(p^*)$ is meaningful. *Globally stable transformations* have the property that for every fixed distribution $\mathcal{D}^*$ on instance-outcome pairs, $\tau(p^*)$ is *close* to its expectation $\mathbb{C}[\tau(p)]$. There is some flexibility in defining closeness; a natural choice is L_1 norm. In fact, a weaker condition suffices for our purposes. *Large-set stability* requires only that for any set S fixed *a priori*, with high probability over the samples and random bits fed to the learner (and the randomness of the transformed predictor, if it, too, is randomized), the average prediction of $\tau(p^*)$ on $x \sim \mathcal{D}^*|_S$ is close to its expectation $\mathbb{C}[\tau(p^*)](x)$. In consequence, given a candidate q, large-set stability ensures that the average values of $q^* = \tau(p^*)$ on the level sets S_v of $q(S)$ are well-defined. This is crucial for reasoning about whether or not q is a multicalibrated with respect to q^*.

On related work. A vast body of work spanning many disciplines has studied corrective transformations to real-life (for example, works that study affirmative action). This body of work is too vast for us to survey here. Our work studies this question in the context of risk prediction and through the lens of algorithmic fairness. While fairness in risk prediction is a widely-studied topic in algorithmic fairness, the focus has been on learning a predictor that satisfies fairness desiderata while maintaining fidelity to the underlying distribution (e.g. [2, 6–8]), or on applying corrective transformations to learned risk predictors (e.g. [5]). Our work, on the other hand, initiates a study of learning about (probabilities in) a better world, where the better world is obtained by applying a corrective transformation on the real world itself.

2 Preliminaries, Setup and Definitions

Notation. For a distribution $\mathcal{D}$ over domain $\mathcal{X}$, we use $\mathrm{Supp}(\mathcal{D})$ to refer to the support of the distribution (the set of elements in $\mathcal{X}$ that have non-zero probability). For $x \in \mathcal{X}$ we use $\mathcal{D}[x]$ to refer to x's probability. For a subset $S \subseteq \mathcal{X}$ we use $\mathcal{D}[S]$ to refer to the aggregate probability of the set S under $\mathcal{D}$ (i.e. $\mathcal{D}[S] = \sum_{x \in S} \mathcal{D}[x]$). For a set S with non-zero probability, we use $(\mathcal{D}|S)$ to refer to the conditional distribution of $\mathcal{D}$, conditioned on landing in S.

Underlying all of this is a modeling assumption, in which “Nature” assigns a probability $p^*(x)$ to each individual x. We are agnostic as to whether $p^*(x) \in \{0, 1\}$ for all x or $p^*(x)$ can be arbitrary in $[0, 1]$. Since we cannot have access to p^* (we don't even know if it is real-valued!), the OI/MC literature builds scoring functions trained on outcomes $o^*(x)$ that Nature provides. However, the nomenclature “Nature” (inherited from a long literature on forecasting) is singularly inapt when viewed from a perspective of social justice, where one's “probability” of success and actual outcome are not solely intrinsic to the individual but are influenced – positively or negatively – by family wealth, structural racism, antisemitism, sexism, ableism, hetero-normativity, (lack of) availability of contraception and access to abortion, and so on. These are not forces of “Nature”, they are social forces that shape the reality in which we live.

We model real-life as a joint distribution over individuals and outcomes, denoted $\mathcal{D}^*$. An individual is described by a d-dimensional boolean string representing their “features”, and we focus on Boolean outcomes. Thus, $\mathcal{D}^*$ is supported on $\{0, 1\}^d \times \{0, 1\}$. We refer to $\mathcal{X} = \{0, 1\}^d$ as the *feature space*, and use $x \sim \mathcal{D}_{\mathcal{X}}$ to denote a sample from real-life's marginal distribution over individuals.

A *predictor* is a function $p : \mathcal{X} \to [0, 1]$ that maps individuals to an estimate of the conditional probability of the individual's outcome being 1. For ease of notation, we use $p_x = p(x)$ to denote a predictor's estimate for individual x. The marginal distribution over individuals $\mathcal{D}_{\mathcal{X}}$ paired with a predictor induce a joint distribution over $\mathcal{X} \times \mathcal{Y}$. Given a predictor p, we use $(x, y) \sim \mathcal{D}(p)$ to denote an individual-outcome pair, where $x \sim \mathcal{D}_{\mathcal{X}}$ is sampled from real-life's distribution over individuals, and the outcome $y \sim \mathrm{Ber}(p_x)$ is sampled – conditional on x – according to the Bernoulli distribution with parameter p_x. We use $p^* : \mathcal{X} \to [0, 1]$ to denote the marginal distribution on outcomes of real-life's distribution $\mathcal{D}^*$.

A *randomized instantiation* of a predictor p is the randomized process of fixing the prediction on each $x \in \mathcal{X}$ to be boolean, where the probability of 1 is exactly $p(x)$ (the boolean prediction for each x is drawn independently). We denote the (probabilistic) outcome of this process by $\mathsf{RI}(p)$.

2.1 Multicalibration and Multiaccuracy

We start with the notion of multi-accuracy. Given a collection of subpopulations $\mathcal{C}$, multi-accuracy requires that a predictor $\tilde{p}$ reflect the expectations of p^* correctly over each subpopulation $S \in \mathcal{C}$.

▶ **Definition 2** (Multi-Accuracy [6]). *Fix a feature distribution $\mathcal{D}_{\mathcal{X}}$ and a predictor $p^* : \mathcal{X} \to [0,1]$. For a collection of sets $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}}$ and $\alpha, \gamma \geq 0$, a predictor $\tilde{p} : \mathcal{X} \to [0,1]$ satisfies $(\mathcal{C}, \alpha, \gamma)$-multi-accuracy w.r.t. p^* (under the feature distribution $\mathcal{D}_{\mathcal{X}}$) if for every $S \in \mathcal{C}$ s.t. $\mathcal{D}_{\mathcal{X}}[S] \geq \gamma$:*

$$\left| \mathop{\mathbf{E}}_{x \sim \mathcal{D}_{\mathcal{X}}} [\, p^*(x) \mid x \in S \,] - \mathop{\mathbf{E}}_{x \sim \mathcal{D}_{\mathcal{X}}} [\, \tilde{p}(x) \mid x \in S \,] \right| \leq \alpha \tag{1}$$

Multi-calibration is a stronger notion, requiring the predictor $\tilde{p}$ to be calibrated with respect to p^* over each $S \in \mathcal{C}$. Here, a set of predictions is calibrated if amongst the individuals $x \in \mathcal{X}$ who receive prediction $\tilde{p}(x) = v$, their actual expectation is v. For a set S and a value $v \in [0,1]$, let S_v be the subset of S to which $\tilde{p}$ assigns value v. We use $\mathrm{supp}_S(\tilde{p}) = \{v \in [0,1] : \mathbf{Pr}_{x \sim \mathcal{D}_{\mathcal{X}}} [\, \tilde{p}(x) = v \mid x \in S \,] > 0\}$ to denote the support of $\tilde{p}$ on S (the set of values v s.t. $S_{v'}$ has non-zero mass).

▶ **Definition 3** (Multi-Calibration [6]). *Fix a feature distribution $\mathcal{D}_{\mathcal{X}}$ and a predictor $p^* : \mathcal{X} \to [0,1]$. For a collection of sets $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}}$ and parameters $\alpha, \gamma > 0$, a predictor $\tilde{p} : \mathcal{X} \to [0,1]$ satisfies $(\mathcal{C}, \alpha, \gamma)$-multi-calibration w.r.t. p^* (under the feature distribution $\mathcal{D}_{\mathcal{X}}$) if for every set $S \in \mathcal{C}$ s.t. $\mathcal{D}_{\mathcal{X}}[S] \geq \gamma$, there exists a set $S' \subseteq S$ with $\mathcal{D}_{\mathcal{X}}[S'] \geq (1-\alpha)\mathcal{D}_{\mathcal{X}}[S]$ where:*

$$\forall v \in \mathrm{supp}_{S'}(\tilde{p}) : \left| \mathop{\mathbf{E}}_{x \sim (\mathcal{D}_{\mathcal{X}} | S_{v'})} [p^*(x)] - v \right| \leq \alpha. \tag{2}$$

When p^* is real-life's distribution, we simply refer to the predictor $\tilde{p}$ as multi-calibrated or multi-accurate, but we will also discuss these requirements w.r.t predictors that are not real-life. We often assume that the predictor $\tilde{p}$ is discretized to precision $\lambda = \Theta(\alpha)$ (see [6]).

3 Corrective Transformations

We study corrective transformations that will be applied to risk predictors. The transformation may include an optional *parameter-learning phase*. If the transformation does not use a learning phase, then we say that it is *fully explicit*. Otherwise, the transformation specifies a parameter-learner that can observe individual-outcome pairs drawn from the underlying distribution, or even observe individual-prediction pairs (see Definition 5). The learning phase outputs parameters π that are plugged into the transformation τ, which can be deterministic or probabilistic.

We begin by defining fully explicit and deterministic corrective transformations.

▶ **Definition 4** (Fully explicit and deterministic corrective transformation.). *A* fully explicit (and deterministic) transformation *is a mapping $\tau : \mathcal{X} \times [0,1] \to [0,1]$ that transforms a predictor p into a new predictor $\tau(p)$, where $\forall x \in \mathcal{X}$, $(\tau(p))(x) = \tau(x, p(x))$.*

Parameterized transformations (see above) also include a parameter-learning phase:

▶ **Definition 5** (Parameterized transformation τ). *A transformation is a pair $(\mathcal{L}, \tau)$, where $\mathcal{L}$ is a parameter-learning algorithm that gets access to training data (see below) and outputs parameters π. For any fixing of the parameters π, the mapping τ, using those parameters, transforms a predictor p into a new predictor $\tau_\pi(p)$, where $\forall x \in \mathcal{X}$, $(\tau_\pi(p))(x) = \tau_\pi(x, p(x))$.*

We consider different options for the parameter-learning algorithm $\mathcal{L}$ and its training data:

- Fully-explicit transformation: *There is no parameter learning. The learner $\mathcal{L}$ always outputs the empty string (if τ is deterministic, then this equivalent to Definition 4).*
- Outcome-based parameters: *The transformation is applied to a predictor p with respect to an underlying feature distribution $\mathcal{D}_\mathcal{X}$. The learner $\mathcal{L}$ gets access to individual-outcome examples (x, o), where $x \sim \mathcal{D}_\mathcal{X}$ and $o \sim \mathrm{Ber}(p(X))$, and outputs parameters π.*
- Prediction-based parameters: *The transformation is applied to a predictor p with respect to an underlying feature distribution $\mathcal{D}_\mathcal{X}$. The learner $\mathcal{L}$ gets access to individual-prediction examples $(x, p(x))$, where $x \sim \mathcal{D}_\mathcal{X}$, and outputs parameters π.*

We use $\pi \leftarrow \mathcal{L}^{\mathcal{D}_\mathcal{X}, p}$ to denote the process of running the parameter learner w.r.t a feature distribution $\mathcal{D}_\mathcal{X}$ and a predictor p, producing learned parameters π. We allow both the learner and the mappting τ to be randomized, and denote the random strings they use by $r_\mathcal{L}$ and r_τ (respectively).

We sometimes abuse notation and refer to the transformation as τ, where the parameter-learning algorithm is implicit. We also use $\tau(p)$ as shorthand for $\tau_\pi(p)$, where the parameters π are learned by the parameter-learning process.

3.1 Stable Transformations

Our primary focus is on transformations that are *stable* with respect to the choice of samples and random coins used by the learner, as well as the coins used by τ. We consider two definitions of stability: global stability, which requires that the resulting predictor is close to its expectation (globally, in L_1 distance). The more relaxed property of Large-set stability only requires that for any sufficiently large set (fixed a-priori), w.h.p. the average prediction is close to the expectation (the latter expectation is over the learner's and τ's random choices).

▶ **Definition 6** (Canonical transformed predictor). *Fix a feature distribution $\mathcal{D}_\mathcal{X}$, a corrective transformation $(\mathcal{L}, \tau)$, and a predictor p. The* canonical transformed predictor *is defined as:*

$$\mathbb{C}[\tau(p)] \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{\pi \leftarrow \mathcal{L}^{\mathcal{D}_\mathcal{X}, p}, r_\tau} [\tau_{\pi, r_\tau}(p)].$$

For the remainder of this writeup, We will reserve the special symbol "$\mathbb{C}$" to remind the reader that we are referring to the canonical predictor.

▶ **Definition 7** (Globally stable transformation). *Fix a feature distribution $\mathcal{D}_\mathcal{X}$. A transformation $(\mathcal{L}, \tau)$ is (α, β)-*globally stable *w.r.t. $\mathcal{D}_\mathcal{X}$ if for any predictor p, w.h.p. its (randomized) transformation $\tau(p)$is close to the canonical transformed predictor in L_1 distance:*

$$\mathop{\mathbf{Pr}}_{\pi \leftarrow \mathcal{L}^{\mathcal{D}_\mathcal{X}, p}, r_\tau} \left[\mathop{\mathbf{E}}_{x \sim \mathcal{D}_\mathcal{X}} [|[\tau_{\pi, r_\tau}(p)](x) - \mathbb{C}[\tau(p)](x)|] > \alpha \right] < \beta.$$

If $(\mathcal{L}, \tau)$ is (α, β)-globally stable for every *distribution $\mathcal{D}_\mathcal{X}$ then we say that it is* universally *globally stable.*

▶ **Definition 8** (Large-set stable (LSS) transformation). *Fix a feature distribution $\mathcal{D}_{\mathcal{X}}$ and let $\alpha, \beta : [0,1] \to [0,1]$ be functions bounding the magnitude and probability of instability as a function of the set size (see below). A transformation $(\mathcal{L}, \tau)$ is (α, β)-*large set stable (LSS) *w.r.t. $\mathcal{D}_{\mathcal{X}}$ if for any predictor p and for any fixed set $S \subseteq \mathcal{X}$, taking $\gamma = \mathbf{Pr}_{x \sim \mathcal{D}_{\mathcal{X}}}[S]$:*

$$\mathop{\mathbf{Pr}}_{\pi \leftarrow \mathcal{L}^{\mathcal{D}_{\mathcal{X}}, p}, r_\tau} \left[\left| \mathop{\mathbf{E}}_{x \sim (\mathcal{D}_{\mathcal{X}} | S)} [[\tau_{\pi, r_\tau}(p)](x) - \mathbb{C}[\tau(p)](x)] \right| > \alpha(\gamma) \right] < \beta(\gamma).$$

We emphasize that the absolute value in the above equation is external*: we compare the expectation of $\tau(p)$ on the entire set S with the expectation of the canonical transformed predictor on that set.*

If $(\mathcal{L}, \tau)$ is (α, β, γ)-LSS for every *distribution $\mathcal{D}_{\mathcal{X}}$ then we say that it is* universally *LSS.*

The error probability β will usually be exponentially small, so we can take a Union bound over large collections of sets, and conclude that w.h.p. for all of them simultaneously, the expectation of the transformed predictor is close to the expectation of the canonical transformed predictor.

We omit the "universally" or "w.r.t a particular distribution" suffix when they are clear from the context, simply referring to a corrective transformation as globally or large-set stable.

3.2 Our Goal: Evidence-Based Corrective Action

Once a corrective transformation is specified, our goal is learning a risk predictor that is "close to" the probabilities specified by the transformation, when it is applied to real-life's probabilities p^*, i.e. close to $\tau(p^*)$. However, we can only observe *outcomes* by real-life's distribution: the *probabilities* p^* are unknowable. Thus, we study the relaxed (but still significant!) goals of obtaining predictors that are multicalibrated or multiaccurate with respect to $\tau(p^*)$.

Here the importance of *stability* (see Section 3.1) becomes apparent: parameter learners are inherently randomized (as they draw samples), and there can also be additional randomization in $\mathcal{L}$ or in τ. We want to be "close" to the transformed predictor, but which of the many possibly predictors in the support of $\tau(p^*)$'s output distribution should we aim to be close to? For stable transformations, the behavior of $\tau(p^*)$ on any (large enough) set is close to its expectation w.h.p. Thus, it is natural to aim to be close to the canonical transformed predictor $\mathbb{C}[\tau(p^*)]$:

▶ **Definition 9** (multiaccurate/multicalibrated learning algorithm for $(\mathcal{L}, \tau)$). *Let $(\mathcal{L}, \tau)$ be a transformation. An algorithm $\mathcal{A}$ for learning a multi-calibrated (respectively, multi-accurate) predictor for the transformation gets as input a collection of subsets $\mathcal{C} \subseteq 2^{\mathcal{X}}$, an error bound $\alpha \in [0,1]$, a failure probability $\beta \in [0,1]$, a set size $\gamma \in [0,1]$, and labeled individual-outcomes pairs drawn from a distribution $\mathcal{D}^*$. Let $\mathbb{C}[\tau(p^*)]$ be the canonical transformation of p^* (see Definition 6).*

We say that $\mathcal{A}$ is a $(\mathcal{C}, \alpha, \beta, \gamma)$-multicalibration (respectively, multi-accuracy) learning algorithm for the transformation $(\mathcal{L}, \tau)$ if, when we run $\mathcal{A}$ on input $(\mathcal{C}, \alpha, \beta, \gamma)$, with all but β probability over $\mathcal{A}$'s random coin tosses and the training samples drawn i.i.d. from $\mathcal{D}^*$, it outputs a predictor $\tilde{q}$ that is (C, α, γ) multi-calibrated (respectively, (C, α, γ) multi-accurate) w.r.t $\mathbb{C}[\tau(p^*)]$ (under the distribution $\mathcal{D}^*_{\mathcal{X}}$).

Discussion. If $(\mathcal{L}, \tau)$ satisfies large-set stability (or the more stringent requirement of global stability), then multi-calibration w.r.t. $\mathbb{C}[\tau(p^*)]$ is quite meaningful: suppose $\tilde{q}$ is a $\mathcal{C}$-multicalibarted predictor w.r.t. $\mathbb{C}[\tau(p^*)]$. Large-set stability implies that w.h.p. over the coins and samples of the transformation, for each set S in the collection $\mathcal{C}$, and for each (sufficiently large) level set S_v of $\tilde{q}$ in S, the expectation of $\tau(p^*)$ (with the above random choices and samples) is close to the expectation by the canonical transformed predictor. Thus, with high probability over the coins and samples of the transformation, the predictions of $\tilde{q}$ will be calibrated on all the sets in $\mathcal{C}$ w.r.t. the (probabilistic) outcome of the corrective transformation applied to real-life. We find this to be a strong guarantee. Note that we assume here that the high probability guarantee is strong enough to allow union bounding over the sets in the collection and their prediction categories.

Multi-calibration with respect to $\mathbb{C}[\tau(p^*)]$ is not appropriate for corrective transformations that make arbitrary randomized distinctions between members of a protected class S, because random but "baseless" distinctions can nonetheless be averaged out in $\mathbb{C}[\tau(p^*)]$. For example, consider a protected group S where $p^* = 0.5$ for all members of S, because the data representation fails to capture appropriate features for members of S that permit accurate prediction[1]. Suppose further that on $T = S^c$, half the elements have $p^*(x) = 1$ and half have $p^*(x) = 0$. One might consider a corrective τ that addresses the situation by arbitrarily assigning a random value in $\{0, 1\}$ to each member of S. This transformation is large-set stable (though it is very much *not* globally stable). However, we have that $\mathbb{C}[\tau(p^*)] = p^*$, so the effect of the transformation is "washed out" in the canonical transformed predictor, and in any $\tilde{q}$ that is multicalibrated w.r.t. $\mathbb{C}[\tau(p^*)]$. One can argue that a corrective transformation, aiming to move the predictions towards a better world, should not make such arbitrary distinctions, and we are sympathetic to this argument. In the full version of this work we address this issue by including in the multicalibration set collect $\mathcal{C}$ sets that may depend on the randomness used by the transformation τ. Finally, we remark that the above discussion is mainly for interpreting the positive direction of our characterization (*i.e.*, how meaningful is multicalibration with respect to $\mathbb{C}[\tau(p^*)]$). The negative direction characterizes the transformations for which achieving multicalibration with respect to $\mathbb{C}[\tau(p^*)]$ is impossible, regardless of how meaningful such a guarantee would be.

4 The Characterization

As discussed in the introduction (and the literature), we are agnostic on the question of whether real-life's outcomes are deterministic (binary) or probabilistic. Our view is that this question is unanswerable, and thus corrective transformations should also be agnostic to it. We formalize this as a *robustness* property from the transformation $(\mathcal{L}, \tau)$: we require that the canonical transformed predictor should be "similar" regardless of whether p^* is binary (deterministic) or not (probabilistic). Similarity is captured by requiring that $\mathbb{C}[\tau(p^*)]$ is close to the expectation, over a randomized instantiation p^{**} of p^*, of the canonical transformation of p^{**}. Closeness is measured in L_1 distance, and recall that each x's probability in p^{**} is binary, drawn from the Bernoulli distribution with expectation $p^*(x)$ (see Section 2). For example, this implies that (at least in expectation), the transformed probabilities should look similar regardless of whether real-life assigned a 0.5 probability to all the individuals, or whether the individuals were randomly partitioned into equally-sized sets with probability 0 and probability 1.

[1] See Chapter 4 of [4] for a real life example involving child protective services.

▶ **Definition 10** (Robustness to RI.). *Fix a feature space $\mathcal{X}$ and a distribution $\mathcal{D}_{\mathcal{X}}$ over features. A transformation $(\mathcal{L}, \tau)$ is (ε, δ)*-robust to random instantiations *w.r.t $\mathcal{D}_{\mathcal{X}}$ if for every predictor p:*

$$\Pr_{x \sim \mathcal{D}_{\mathcal{X}}}\left[\left|\mathbb{C}[\tau(p)](x, p(x)) - \left(\mathop{\mathbf{E}}_{p' \leftarrow \mathsf{RI}(p)}[\mathbb{C}[\tau(p')](x, p'(x))]\right)\right| > \varepsilon\right] \leq \delta$$

▶ **Theorem 11** (Main theorem: transformation characterization). *Fix a feature space $\mathcal{X}$ and a distribution $\mathcal{D}_{\mathcal{X}}$. Let $(\mathcal{L}, \tau)$ be a transformation. Then for every $\varepsilon, \delta > 0$:*

- *If $(\mathcal{L}, \tau)$ is (ε, δ)-robust to random instantiations (as per Definition 10), then there is an algorithm $\mathcal{A}$ s.t. for every collection $\mathcal{C}$, and every $(\bar{\alpha}, \bar{\beta}, \bar{\gamma})$ s.t. $\bar{\alpha} = O((\delta/\bar{\gamma}) + \varepsilon)$, $\mathcal{A}$ is a $(\mathcal{C}, \bar{\alpha}, \bar{\beta}, \bar{\gamma})$ multi-calibration learning algorithm for the tranformation $(\mathcal{L}, \tau)$. The sample complexity of $\mathcal{A}$ is $\mathrm{poly}(\log|\mathcal{C}|, 1/\bar{\alpha}, \log(1/\bar{\beta}), 1/\bar{\gamma})$.*
- *If $(\mathcal{L}, \tau)$ is* not *(ε, δ)-robust to random instantiations w.r.t $\mathcal{D}_{\mathcal{X}}$, then there exists a set S s.t. for any α, β s.t. $(\alpha + \beta) < (\varepsilon/2 - \mathsf{negl})$ where* negl *bounds the probability that there is a feature-collision in the algorithm's training sample (some feature vector appears more than once), there is no $(\mathcal{C} = \{S\}, \alpha, \beta, \gamma = \delta/2)$ multi-accurate learning algorithm for the transformation.*

Theorem 11 characterizes the transformations for which, for any given finite collection of sets $\mathcal{C}$, it is sample-theoretically possible to learn a predictor that is $\mathcal{C}$-multi-calibrated (or multi-accurate) with respect to $\mathbb{C}[\tau(p^*)]$. The positive direction constructs an algorithm whose sample complexity is logarithmic in $|\mathcal{C}|$, whereas the negative direction shows a *singleton* collection for which even multi-accuracy is impossible to obtain. The impossibility holds unless the algorithm uses sufficiently many samples to start observing "collisions" or repeated events (i.e. multiple instances of the same feature vector), whereas we are interested in the setting where events are non-repeatable. Thus, we think of the collision probability as negligible. Finally, the theorem does not assume the transformation is stable; our study of stability (Section 3.1) elucidates the qualitative *significance* of being multicalibrated with respect to $\mathbb{C}[\tau(p^*)]$, finding that the concept is meaningful under large-set stability.

Proof of Theorem 11.

Direction I: Non-Robustness ⇒ no multiaccuracy. If $(\mathcal{L}, \tau)$ is not δ-robust to random instantiations w.r.t $\mathcal{D}_{\mathcal{X}}$, then there exists a predictor $p : \mathcal{X} \to [0, 1]$ s.t.:

$$\Pr_{x \sim \mathcal{D}_{\mathcal{X}}}\left[\left|\mathbb{C}[\tau(p)](x, p(x)) - \left(\mathop{\mathbf{E}}_{p' \leftarrow \mathsf{RI}(p)}[\mathbb{C}[\tau(p')]]\,(x, p'(x))\right)\right| > \varepsilon\right] \geq \delta.$$

The above probability considers the absolute value of the difference between the two terms. Since the absolute value is large at least δ probability, there must be a subset $S \subseteq X$ (defined ex-post) where the predictions of the canonical transformed predictor are either significantly larger or significantly smaller than those of the canonical transformation of a randomized instantiation of p. Suppose w.l.o.g that the former is true, i.e. we have that:

$$\mathcal{D}_{\mathcal{X}}[S] \geq \frac{\delta}{2}, \tag{3}$$

and that:

$$\forall x \in S : \mathbb{C}[\tau(p)](x, p(x)) - \mathop{\mathbf{E}}_{p' \leftarrow \mathsf{RI}(p)}[\mathbb{C}[\tau(p')]\,(x, p'(x))] > \varepsilon. \tag{4}$$

Suppose towards contradiction that $\mathcal{A}$ is an algorithm for learning a multi-accurate transformed predictor $\tilde{q}$. We run $\mathcal{A}$ with parameters α, β (see below) and $\gamma = \delta/2$ and on the collection of sets $\{S\}$ (i.e. the collection is a singleton). $\mathcal{A}$ gets i.i.d. feature-outcome samples $\{(x_i, y_i)\}$, where $x_i \in \mathcal{X}$ is sampled from $\mathcal{D}_{\mathcal{X}}$ and $y_i \in \{0, 1\}$ is Bernoulli with expectation $p^*(x)$. Consider two experiments of running $\mathcal{A}$ with different p^*'s:

1. In Experiment 1, we set $p^* = p$.
2. In Experiment 2, we draw $p^* \leftarrow \mathsf{RI}(p)$.

In both experiments we run $\mathcal{A}$ on outcomes drawn by p^*, and let $\tilde{q}$ be the predictor that $\mathcal{A}$ outputs.

Consider the random variables Q_1 and Q_2, where Q_c is defined to be the value $\mathbf{E}_{x\sim(\mathcal{D}_\mathcal{X}|S)}[\tilde{q}(x)]$ in Experiment c (the RVs Q_1, Q_2 are over the domain $[0,1]$). If $\mathcal{A}$ is an $(\alpha, \beta, \gamma = \delta/2)$-multiaccuracy learning algorithm for the transformation $(\mathcal{L}, \tau)$, then since $\mathcal{D}_\mathcal{X}[S] \geq \gamma$ (see Equation (3)), by Definition 9:

$$\mathbf{Pr}\left[\left|Q_1 - \mathop{\mathbf{E}}_{x\sim(\mathcal{D}_\mathcal{X}|S)}[\mathbb{C}[\tau(p)](x, p(x))]\right| > \alpha\right] < \beta. \tag{5}$$

On the other hand, consider Experiment 2 and consider a *fixed* randomized instantiation p' (Experiment 2 includes the random process of drawing the randomized instantiation, whereas here we consider a fixed instantiation that has positive probability). Let $(Q_2|p')$ be the RV obtained by conditioning Q_2 on this fixed p'. Again, since $\mathcal{D}_\mathcal{X}[S] \geq \gamma$, by Definition 9:

$$\mathbf{Pr}\left[\left|(Q_2|p') - \mathop{\mathbf{E}}_{x\sim(\mathcal{D}_\mathcal{X}|S)}[\mathbb{C}[\tau(p')](x, p'(x))]\right| > \alpha\right] < \beta.$$

Experiment 2 consists of choosing a random instantiation p', and then running the learning algorithm. By the above, adding an expectation over the randomized instantiation p', we have that:

$$\left|\mathbf{E}[Q_2] - \mathop{\mathbf{E}}_{x\sim(\mathcal{D}_\mathcal{X}|S), p'\leftarrow\mathsf{RI}(p)}[\mathbb{C}[\tau(p')](x, p'(x))]\right| \leq \alpha + \beta. \tag{6}$$

Thus, by Equation (4), the value of Q_1 is w.h.p. higher than the expectation of Q_2. This implies a lower bound on the statistical distance between Q_1 and Q_2

▷ Claim 12. $\Delta(Q_1, Q_2) > \frac{\varepsilon}{2} - \alpha - \beta$.

Proof. The proof follows by the fact that the expectations of two random variables supported on $[0,1]$ cannot differ by more than their statistical distance:

$$\mathbf{E}[Q_1] - \mathbf{E}[Q_2] = \sum_{v\in[0,1]} (Q_1[v] \cdot v - Q_2[v] \cdot v) \leq \sum_{v\in[0,1]} |Q_1(v) - Q_2(v)| = 2\Delta(Q_1, Q_2).$$

Further, putting together Equations (4), (5) and (6) we conclude that:

$$E[Q_1] - E[Q_2] > \varepsilon - 2(\alpha + \beta).$$

The claim follows. ◁

The only difference between the two experiments is in the distributions of the feature-outcome samples fed to the learning algorithm. In particular, the difference is in the distribution of the binary outcomes: by p, or by a randomized instantiation of p. The feature-vectors are identically distributed in both experiments (i.i.d. from $\mathcal{D}_\mathcal{X}$). If the feature-vectors sampled by the learning algorithm are all distinct, then the conditional distributions on the outcomes in the two experiments (for those fixed feature vectors) are also identical: for each x, the outcome is Bernoulli with expectation $p(x)$. In Experiment 1 this is by design. In Experiment 2, this is due to the choice of a randomized instantiation p' of p, and so long as the samples are all distinct, the outcomes are drawn i.i.d. from the above distribution.

The only difference between the experiments is that if the same feature vector x is observed more than once, then in Experiment 1, the outcomes for the different occurrences of x will be independent, whereas in Experiment 2 they will be *identical* (since the predictor p' is instantiated once). The random variable Q is just a function of the algorithm's training sample. Thus, so long as the probability of observing the same feature vector more than once is negligible, we have:

▷ Claim 13. $\Delta(Q_1, Q_2) \leq \mathsf{negl}$.

Claims 12 and 13 give a contradition to the assumption that $\mathcal{A}$ is a $(\alpha, \beta, \gamma = \delta/2)$ multiaccuracy algorithm for any values of α and β for which $\alpha + \beta < \frac{\varepsilon}{2} - \mathsf{negl}$.

Direction II: Robustness implies calibration-feasibility. We construct an algorithm that learns a predictor that is multicalibrated with respect to the canonical transformation of p^* for any robust transformation. For a robust transformation, the canonical transformation of any p^* is close to the expectation, over a randomized instantiation p^{**} of p^*, of the canonical transformation of p^{**}. The main step in our algorithm is using outcomes drawn by p^* to generate outcomes whose distributions are close to $\mathbf{E}_{p^{**} \leftarrow \mathsf{RI}(p^*)}[\mathbb{C}[\tau(p^{**})]]$. Robustness guarantees that this distribution is close to that of the canonical transfomaion of p^*. We then use a standard outcome-based multi-calibration learning algorithm (e.g. [6]), trained over the aforementioned samples, to obtain a predictor $\tilde{q}$ that is multiclibrated w.r.t. the canonical transformation of p^*. The theorem follows.

Our goal, then, is generating outcomes that are close in distribution to $\mathbf{E}_{p^{**} \leftarrow \mathsf{RI}(p^*)}[\mathbb{C}[\tau(p^{**})]]$. To do this, we treat the observed *outcomes* drawn by p^* as specifying *probabilities* according to a fictitious randomized instantiation p^{**} of p^*. These probabilities are fed into the (probability-based) parameter learner $\mathcal{L}$ to learn parameters π for the transformation τ, towards applying it on (the fictitious) p^{**}. The key point is that these learned parameters will be *identically* distributed to parameters learned by $\mathcal{L}$ on an actual randomized instantiation of p^*. Algorithm 1 details the sample-generation procedure.

The predictor q. Step 1 of the sample generation algorithm produces a set of learned parameters $\{\pi_i\}$. These parameters are then used in Step 2 to generate new samples, where we also take care (both in training and in sample generation) to ensure that the unstransformed outcome for each feature vector $x \in X$ is consistent across all its appearances in training the i-th parameters and in generating samples. Fixing a run of Step 1 of the sample generator, for any fixed feature vector $x \in \mathcal{X}$ that is in the support of $\mathcal{D}_{\mathcal{X}}$, let $q(x)$ denote the conditional probability that Step 2 produces the sample $(x, y' = 1)$ (conditioned on the feature vector x). The following claim shows that w.h.p. over the coins used in Step 1, for almost all x drawn from $\mathcal{D}_{\mathcal{X}}$, the conditional probability $q(x)$ is close to the expectation, over a randomized instantiation p^{**} of p^*, of the probability assigned by the canonical transformed predictor. The notation $E_{q \leftarrow \text{Step 1}}$ emphasizes that we are taking expectation only over the randomness in the first step, in which the parameters $\{\pi_1, i \in [\ell]\}$ are learned, and not over the randomness in Step 2 in which a random $i \in [\ell]$ is selected.

▷ Claim 14. Fix parameters $\mu, \rho \in [0, 1]$. For the sample-generation algorithm (Algorithm 1) it holds that:

$$\Pr_{q \leftarrow \text{Step } 1, x \sim \mathcal{D}^*_{\mathcal{X}}} \left[\left| q(x) - \mathop{\mathbf{E}}_{p^{**} \leftarrow \mathsf{RI}(p^*)} [\mathbb{C}[\tau(p^{**})](x, p^{**}(x))] \right| \geq \mu \right] < \rho$$

Algorithm 1 Sample Generation for Robust Transformations.

```
Input: feature-outcome pairs, outcomes by $p^*$, error parameters $\mu, \rho \in [0,1]$
Output: feature-outcome pairs, outcomes close to $\mathbf{E}_{p^{**}\leftarrow\mathsf{RI}(p^*)}[\mathbb{C}[\tau(p^{**})]]$

1. Run $\ell = O(\sqrt{\log(1/\rho)}/\mu^2)$ indep. executions of the parameter learner $\mathcal{L}$. For each $i \in [\ell]$:
    a. For every $x \in \mathcal{X}$, the $i$-th untransformed outcome $o_x^i$ of $x$ is initialized to be "undefined".
    b. The $i$-th execution uses freshly drawn random coins $r_{\mathcal{L},i}$.
    c. To produce the $j$-th feature-probability sample requested by the $i$-th execution of $\mathcal{L}$, sample $(x_{i,j}, y_{i,j} \in \{0,1\}) \sim \mathcal{D}(p^*)$. If $x_{i,j}$'s $i$-th untransformed outcome $o^i_{x_{i,j}}$ is defined, then proceed to the next step. Otherwise, set it to $y_{i,j}$.
    d. Use $(x_{i,j}, o^i_{x_{i,j}})$ as the $j$-th sample in the $i$-th execution of the parameter-learner.
    e. The parameter-learner outputs parameters $\pi_i$.

2. Produce each new feature-outcome output sample as follows:
    a. Draw $(x, y \in \{0,1\}) \sim \mathcal{D}(p^*)$. Pick $i \in [\ell]$ uniformly at random.
    b. If $x$'s $i$-th untransformed outcome $o_x^i$ is defined, then proceed to the next step. Otherwise, set it to $y$.
    c. Draw $y' \in \{0,1\}$ from the Bernoulli distribution with expectation $\tau_{\pi_i}(x, o_x^i)$ and output the sample $(x, y')$.
```

Proof. In Step 1 of the algorithm, consider a single execution i of the parameter-learning algorithm: the distribution of the learned parameters π_i is *identical* to the distribution of the parameters that would be learned by taking a randomized instantiation p^{**} of p^*: the randomized instantiation is simply determined by the observed binary outcomes (which are drawn by $p^*(x)$)), where we take care to make sure that if a feature-vector x appears more than once in the training examples, then it is always "assigned" the binary outcome with which it first appeared (the i-th untrasnformed outcome is set only once). Moreover, we also take care that for any feature vector x that appears in Step 2, its untransformed outcome is set only once (when it first appeared, in training or in sample-generation for the i-th learned parameters).

Thus, for each $i \in [\ell]$, the distribution of outcomes that are generated in Step 2, conditioned on that using the i-th learned parameters, is identical to the distribution that would be obtained in a mental experiment, where we take a randomized instantiations $p_i^{**} \leftarrow \mathsf{RI}(p^*)$, and learn the parameters π_i by training on examples drawn by p_i^{**}.

For x in the support of $\mathcal{D}_{\mathcal{X}}$, recall that $q(x)$ denotes the probability that the sample generator assigns outcome 1 to x. We conclude that q is in fact the average of ℓ predictors q_i, where each q_i is drawn by choosing a random instantiation of p^* and transforming it using $(\mathcal{L}, \tau)$. Thus:

$$\begin{aligned}
&\Pr_{q\leftarrow \text{Step } 1}\left[\left|q(x) - \mathop{\mathbf{E}}_{p^{**}\leftarrow \mathsf{RI}(p^*)}[\mathbb{C}[\tau(p^{**})](x,p^{**}(x))]\right| > \mu\right]\\
=&\Pr_{\{p_i^{**}\leftarrow \mathsf{RI}(p^*),\pi_i\}_{i\in[\ell]}}\left[\left|\mathop{\mathbf{E}}_{i\in[\ell]}[\tau_{\pi_i}(x,p_i^{**}(x))] - \mathop{\mathbf{E}}_{p^{**}\leftarrow \mathsf{RI}(p^*)}[\mathbb{C}[\tau(p^{**})](x,p^{**}(x))]\right| > \mu\right]\\
<&\rho,
\end{aligned}$$

where the first equality is by the mental experiment discussed above, and the second inequality is by a Chernoff bound. The above holds for any fixed x in the support of $\mathcal{D}_{\mathcal{X}}$, and thus it also holds for a randomly drawn $x \sim \mathcal{D}_{\mathcal{X}}$. ◁

Speaking intuitively, Claim 14 tells us that, with high probability over the randomness in defining the building blocks of q, the resulting predictor is close to the expectation, over randomness in $p^{**} \leftarrow \mathsf{RI}(p^*)$, of the canonical transformation of p^{**}. By the robustness of τ, this in turn is close to the canonical transformed $\mathbb{C}[\tau(p^*)]$. Hence, q is close to $\mathbb{C}[\tau(p^*)]$. The remainder of the proof will show that this closeness is maintained under multicalibration; that is, multicalibrating with respect to q yields a predictor that is close to something multicalibrated with respect to $\mathcal{C}[\tau(p^*)]$. Before proceding with that argument, we first state a corollary that follows directly from Claim 14 via a standard argument.

▶ **Corollary 15.** *Fix parameters $\alpha', \beta', \sigma', \rho' \in [0,1]$. For the sample-generation algorithm (Algorithm 1), run with parameters $\mu = \alpha'$ and $\rho = (\alpha' \cdot \beta' \cdot \sigma' \cdot \rho')$ it holds that:*

$$\Pr_{q \leftarrow \textit{Step 1}}\left[\Pr_{x \sim \mathcal{D}^*_{\mathcal{X}}}\left[\left|q(x) - \mathop{\mathbf{E}}_{p^{**} \leftarrow \mathsf{RI}(p^*)}[\mathbb{C}[\tau(p^{**})](x, p^{**}(x))]\right| > \alpha'\right] > (\alpha' \cdot \sigma' \cdot \rho')\right] < \beta'$$

Proof. Plugging the values of μ, ρ into Claim 14, we conclude that:

$$\Pr_{q \leftarrow \text{Step } 1, x \sim \mathcal{D}_{\mathcal{X}}}\left[\left|q(x) - \mathop{\mathbf{E}}_{p^{**} \leftarrow \mathsf{RI}(p^*)}[\mathbb{C}[\tau(p^{**})](x, p^{**}(x))]\right| > \alpha'\right] < (\alpha' \cdot \beta' \cdot \sigma' \cdot \rho').$$

By a standard argument, it follows that it cannot be that with probability larger than β' over the q that is defined by Step 1, the probability, over $x \sim \mathcal{D}_{\mathcal{X}}$, that $q(x)$ is far from its "target" in the above equation is larger than $(\alpha' \cdot \sigma' \cdot \rho')$:

$$\Pr_{q \leftarrow \text{Step } 1}\left[\Pr_{x \sim \mathcal{D}_{\mathcal{X}}}\left[\left|q(x) - \mathop{\mathbf{E}}_{p^{**} \leftarrow \mathsf{RI}(p^*)}[\mathbb{C}[\tau(p^{**})](x, p^{**}(x))]\right| > \alpha'\right] > (\alpha' \cdot \beta' \cdot \sigma' \cdot \rho')\right] < \beta'.$$

◀

From MC w.r.t q to MC w.r.t. the canonical transformed predictor. Running a multicalibration algorithm on outcomes generated by the sample generation algorithm (Algorithm 1) will w.h.p. produce a predictor $\tilde{q}$ that is approximately multicalibrated w.r.t. q. We use Corollary 15 and the robustness of the transformation $(\mathcal{L}, \tau)$ to show that $\tilde{q}$ is also approximately MC w.r.t. the canonical transformation of p^*.

In more detail, let $\mathcal{C}$ be the collection of sets, and let α, β, γ be parameters to be set below. We run the sample-generation algorithm (Algorithm 1) with parameters $\alpha' = \Theta(\alpha), \beta' = \Theta(\beta), \sigma' = \gamma, \rho' = \Theta(\alpha^2)$. By Corollary 15, with all but $\Theta(\beta)$ probability over the training in Step 1, the sample generator trains a predictor q for which there exists a "bad" set $B_q \subseteq \text{Supp}(\mathcal{D}_{\mathcal{X}})$ s.t. $\mathcal{D}_{\mathcal{X}}[B_q] \leq (\alpha^3 \cdot \gamma)/100$ where:

$$\forall x \in (\text{Supp}(\mathcal{D}_{\mathcal{X}}) \setminus B_q) : \left|q(x) - \mathop{\mathbf{E}}_{p^{**} \leftarrow \mathsf{RI}(p^*)}[\mathbb{C}[\tau(p^{**})](x, p^{**}(x))]\right| \leq \alpha/100. \tag{7}$$

Further, by the (ε, δ)-robustness of the transformation (Definition 10), there exists a "bad" set $B_{\text{robust}} \subseteq \mathcal{X}$ where $\mathcal{D}_{\mathcal{X}}[B_{\text{robust}}] \leq \delta$ and

$$\forall x \in (\text{Supp}(\mathcal{D}_{\mathcal{X}}) \setminus B_{\text{robust}}) : \left|\mathbb{C}[\tau(p^*)](x, p^*(x)) - \left(\mathop{\mathbf{E}}_{p^{**} \leftarrow \mathsf{RI}(p^*)}[\mathbb{C}[\tau(p^{**})](x, p^{**}(x))]\right)\right| \leq \varepsilon \tag{8}$$

We are now ready to analyze the guarantee of the multicalibrated predictor $\tilde{q}$ w.r.t. the canonical transformation of p^*. We train $\tilde{q}$ by running an outcome-based multicalibration algorithm on samples generated by Algorithm 1, where the MC algorithm is run on a collection of sets $\mathcal{C}$, and with parameters $\alpha'' = \Theta(\alpha)$, $\beta'' = \Theta(\beta)$ and $\gamma'' = \gamma$. Let $\tilde{q}$ be the predictor trained by the MC learning algorithm. We assume w.l.o.g. that $\tilde{q}$ is discretized to precision $\lambda = \Theta(\alpha)$. In what follows, we assume both that the MC algorithm does not fail (this happens with all but β'' probability), and that q trained by the sample generator satisfies Equation (7) (happens with all but β' probability). By a Union bound, this is the case with all but β probability.

Let $S \in \mathcal{C}$ be a set in the collection s.t. $\mathcal{D}_{\mathcal{X}}[S] \geq \gamma$. For a value $v \in [0,1]$, let S_v be the subset of S to which $\tilde{q}$ assigns value v. We define the "bad" level sets to be the elements assigned values v for which the set S_v has small mass by $\mathcal{D}_{\mathcal{X}}$:

$$B_{\text{levels}}(S) = \bigcup_{v \in [0,1] : \mathcal{D}_{\mathcal{X}}(S_v) \leq (\alpha \cdot \lambda \cdot \gamma)/10} S_v, \tag{9}$$

where recall that the predictor was discretized to precision $\lambda = \Theta(\alpha)$, so there are at most $1/\lambda$ "level sets". Thus, by construction, $\mathcal{D}_{\mathcal{X}}[B_{\text{levels}}(S)] \leq (\alpha \cdot \gamma)/10$.

By Definition 3, for any set $S \in \mathcal{C}$, s.t. $\mathcal{D}_{\mathcal{X}}[S] \geq \gamma$, there is a subset $S' \subseteq S$ where Equation (2) holds. Let S'' be the subset of S' that does not contain members of B_q, of B_{robust}, or of $B_{\text{levels}}(S)$. We have that:

$$\begin{aligned}
\mathcal{D}_{\mathcal{X}}[S''] &\geq \mathcal{D}_{\mathcal{X}}[S'] - \mathcal{D}_{\mathcal{X}}[B_q] - \mathcal{D}_{\mathcal{X}}[B_{\text{robust}}] - \mathcal{D}_{\mathcal{X}}[B_{\text{levels}}(S)] \\
&\geq (1 - \alpha'')\, \mathcal{D}_{\mathcal{X}}[S] - \frac{\alpha^3 \gamma}{100} - \delta - \frac{\alpha \cdot \gamma}{10} \\
&\geq \left(1 - \alpha'' - \frac{\alpha^3}{100} - \frac{\delta}{\gamma} - \frac{\alpha}{10}\right) \mathcal{D}_{\mathcal{X}}[S] \\
&\geq \left(1 - \alpha - \frac{\delta}{\gamma}\right) \mathcal{D}_{\mathcal{X}}[S].
\end{aligned}$$

Since we removed the members of $B_{\text{levels}}(S)$ from S'', it is the case that for every $v \in [0,1]$ for which S''_v has non-zero mass, it has mass at least $(\alpha \cdot \lambda \cdot \gamma)/10$ (see Equation (9)). Thus:

$$\begin{aligned}
\left| \mathop{\mathbf{E}}_{x \sim (\mathcal{D}_{\mathcal{X}} | S''_v)} [\mathbb{C}[\tau(p^*)](x, p^*(x))] - v \right| &\leq \left| \mathop{\mathbf{E}}_{x \sim (\mathcal{D}_{\mathcal{X}} | S''_v)} [E_{p^{**} \leftarrow \mathsf{RI}(p^*)}[\mathbb{C}[\tau(p^{**})](x, p^{**}(x))]] - v \right| + \varepsilon \quad (10) \\
&\leq \left| \mathop{\mathbf{E}}_{x \sim (\mathcal{D}_{\mathcal{X}} | S''_v)} [q(x)] - v \right| + \varepsilon + \frac{\alpha}{100} \quad (11) \\
&\leq \left| \mathop{\mathbf{E}}_{x \sim (\mathcal{D}_{\mathcal{X}} | S'_v)} [q(x)] - v \right| + \varepsilon + \frac{\alpha}{100} + \Theta\left(\alpha + \frac{\delta}{\alpha^2 \cdot \gamma}\right) \quad (12) \\
&\leq \alpha'' + \varepsilon + \Theta\left(\alpha + \frac{\delta}{\alpha^2 \cdot \gamma}\right) \quad (13) \\
&= \Theta\left(\alpha + \frac{\delta}{\alpha^2 \cdot \gamma}\right) + \varepsilon. \quad (14)
\end{aligned}$$

Where in the above: Equation (10) follows by the definition of S'' (which excludes elements in B_{robust}, and by Equation (8)). Equation (11) follows because S'' excludes elements in B_q (and by Equation (7)). In Equation (12) we switch the expectation from S''_v to S'_v using Proposition 16 below, which follows by standard manipulations. Finally, Equation (13) is by the multicalibration guarantee of $\tilde{q}$ w.r.t q.

▶ **Proposition 16.** *For $v \in [0,1]$ s.t. S''_v has non-zero mass:*

$$\left| \mathop{\mathbf{E}}_{x \sim (\mathcal{D}_\mathcal{X} | S'_v)} [q(x)] - \mathop{\mathbf{E}}_{x \sim (\mathcal{D}_\mathcal{X} | S''_v)} [q(x)] \right| = \Theta\left(\alpha + \frac{\delta}{\alpha^2 \cdot \gamma} \right)$$

Proof. The proof is by a case analysis on the sign of the difference in the absolute value. Suppose that the sign is positive, i.e. the first term is larger, then the absolute value is bounded by:

$$\begin{aligned}
\mathop{\mathbf{E}}_{x \sim (\mathcal{D}_\mathcal{X} | S'_v)} [q(x)] - \mathop{\mathbf{E}}_{x \sim (\mathcal{D}_\mathcal{X} | S''_v)} [q(x)] &\leq \frac{1}{\mathcal{D}_\mathcal{X}[S''_v]} \cdot \left(\sum_{x \in S'_v} \mathcal{D}_\mathcal{X}[x] \cdot q(x) - \sum_{x \in S''_v} \mathcal{D}_\mathcal{X}[x] \cdot q(x) \right) \\
&= \frac{1}{\mathcal{D}_\mathcal{X}[S''_v]} \cdot \sum_{x \in (S'_v \setminus S''_v)} \mathcal{D}_\mathcal{X}[x] \cdot q(x) \\
&\leq \frac{\mathcal{D}_\mathcal{X}[(S'_v \setminus S''_v)]}{\mathcal{D}_\mathcal{X}[S''_v]} \\
&\leq \frac{(\alpha^3 \cdot \gamma/100) + \delta}{(\alpha \cdot \lambda \cdot \gamma)/10}.
\end{aligned}$$

If the second term is larger, then the absolute value is bounded by:

$$\begin{aligned}
\mathop{\mathbf{E}}_{x \sim (\mathcal{D}_\mathcal{X} | S''_v)} [q(x)] - \mathop{\mathbf{E}}_{x \sim (\mathcal{D}_\mathcal{X} | S'_v)} [q(x)] &= \frac{1}{\mathcal{D}_\mathcal{X}[S''_v]} \cdot \left(\sum_{x \in S''_v} \mathcal{D}_\mathcal{X}[x] \cdot q(x) - \frac{\mathcal{D}_\mathcal{X}[S''_v]}{\mathcal{D}_\mathcal{X}[S'_v]} \cdot \sum_{x \in S'_v} \mathcal{D}_\mathcal{X}[x] \cdot q(x) \right) \\
&\leq \frac{\sum_{x \in S'_v} \mathcal{D}_\mathcal{X}[x] \cdot q(x) - \frac{\mathcal{D}_\mathcal{X}[S''_v]}{\mathcal{D}_\mathcal{X}[S'_v]} \cdot \sum_{x \in S'_v} \mathcal{D}_\mathcal{X}[x] \cdot q(x)}{\mathcal{D}_\mathcal{X}[S''_v]} \\
&= \frac{\left(1 - \frac{\mathcal{D}_\mathcal{X}[S''_v]}{\mathcal{D}_\mathcal{X}[S'_v]}\right) \cdot \sum_{x \in S'_v} \mathcal{D}_\mathcal{X}[x] \cdot q(x)}{\mathcal{D}_\mathcal{X}[S''_v]} \\
&\leq \left(\frac{\mathcal{D}_\mathcal{X}[S'_v]}{\mathcal{D}_\mathcal{X}[S''_v]} - 1 \right) \cdot \frac{\mathcal{D}_\mathcal{X}[S'_v]}{\mathcal{D}_\mathcal{X}[S''_v]},
\end{aligned}$$

where the last inequality holds because $\frac{\mathcal{D}_\mathcal{X}[S''_v]}{\mathcal{D}_\mathcal{X}[S'_v]} \in (0,1]$, and thus:

$$1 - \frac{\mathcal{D}_\mathcal{X}[S''_v]}{\mathcal{D}_\mathcal{X}[S'_v]} \leq \frac{1 - \frac{\mathcal{D}_\mathcal{X}[S''_v]}{\mathcal{D}_\mathcal{X}[S'_v]}}{\frac{\mathcal{D}_\mathcal{X}[S''_v]}{\mathcal{D}_\mathcal{X}[S'_v]}} = \frac{\mathcal{D}_\mathcal{X}[S'_v]}{\mathcal{D}_\mathcal{X}[S''_v]} - 1.$$

The claim follows by observing that:

$$\begin{aligned}
\frac{\mathcal{D}_\mathcal{X}[S'_v]}{\mathcal{D}_\mathcal{X}[S''_v]} &\leq \frac{\mathcal{D}_\mathcal{X}[S''_v] + (\alpha^3 \cdot \gamma/100) + \delta}{\mathcal{D}_\mathcal{X}[S''_v]} \\
&\leq 1 + \frac{(\alpha^3 \cdot \gamma/100) + \delta}{(\alpha \cdot \lambda \cdot \gamma)/10}
\end{aligned}$$

◀

We conclude that, with all but β probability over the sample generation and learning procedures, $\tilde{q}$ is $(\Theta(\alpha + \delta/(\alpha^2 \cdot \gamma)) + \varepsilon, \gamma)$-multicalibrated w.r.t. the canonical transformation of p^*. The second direction of the theorem follows by setting $\beta = \bar{\beta}$, $\gamma = \bar{\gamma}$ and setting:

$$\alpha = \bar{\alpha} - \Theta\left((\delta/\gamma)^{1/3} \right) - \varepsilon.$$

The restriction on $\bar{\alpha}$ implies that $\alpha = \Omega(\bar{\alpha})$ (so the sample complexity of the multicalibrated learning algorithm will be polynomial in $(1/\bar{\alpha})$), and that $\alpha > (\delta/\gamma)^{1/3}$. Thus:

$$\Theta(\alpha + \delta/(\alpha^2 \cdot \gamma)) + \varepsilon \leq \Theta(\alpha + (\delta/\gamma)^{1/3}) + \varepsilon = \bar{\alpha}.$$

We conclude that the algorithm indeed achieves $(\bar{\alpha}, \bar{\beta}, \bar{\gamma})$ multicalibration w.r.t. the transformed predictor, and (this direction of) the theorem follows. ◀

References

1 Philip Dawid. On individual risk. *Synthese*, 194(9):3445–3474, 2017.

2 Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. Fairness through awareness. In Shafi Goldwasser, editor, *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*, pages 214–226. ACM, 2012. `doi:10.1145/2090236.2090255`.

3 Cynthia Dwork, Michael P Kim, Omer Reingold, Guy N Rothblum, and Gal Yona. Outcome indistinguishability. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1095–1108, 2021.

4 Virginia Eubanks. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press, Inc., USA, 2018.

5 Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3315–3323, 2016. URL: `https://proceedings.neurips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html`.

6 Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1939–1948. PMLR, 2018.

7 Lunjia Hu, Inbal Livni Navon, Omer Reingold, and Chutong Yang. Omnipredictors for constrained optimization. *CoRR*, abs/2209.07463, 2022. `doi:10.48550/arXiv.2209.07463`.

8 Michael J. Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2569–2577. PMLR, 2018. URL: `http://proceedings.mlr.press/v80/kearns18a.html`.

9 Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint*, 2016. `doi:10.48550/arXiv.1609.05807`.