Proceedings of Machine Learning Research vol 132:1–39, 2022 33rd International Conference on Algorithmic Learning Theory

Beyond Bernoulli: Generating Random Outcomes that cannot be Distinguished from Nature

Cynthia Dwork DWORK@SEAS.HARVARD.EDU
Harvard University

Michael P. Kim MPKIM@BERKELEY.EDU
UC Berkeley, Miller Institute

Omer Reingold REINGOLD@STANFORD.EDU
Stanford University

Guy N. Rothblum ROTHBLUM@ALUM.MIT.EDU
Weizmann Institute of Science

Gal Yona GAL.YONA@WEIZMANN.AC.IL
Weizmann Institute of Science

Abstract

Recently, Dwork et al. (STOC 2021) introduced Outcome Indistinguishability as a new desideratum for binary prediction tasks. Outcome Indistinguishability (OI) articulates the goals of prediction in the language of computational indistinguishability: a predictor is Outcome Indistinguishable if no computationally-bounded observer can distinguish Nature's outcomes from outcomes that are generated based on the predictions. In this sense, OI suggests a generative model for binary outcomes that cannot be refuted given the empirical evidence and computational resources at hand. In this work, we extend Outcome Indistinguishability beyond Bernoulli, to outcomes that live in a large discrete or continuous domain.

While the idea of OI for non-binary outcomes is natural for many applications, defining OI in generality is not simply a syntactic exercise. We introduce and study multiple definitions of OI—each with its own semantics—for predictors that completely specify each individuals' outcome distributions, as well as predictors that only partially specify the outcome distributions through statistics, such as moments. With the definitions in place, we provide learning algorithms for producing OI generative outcome models for general random outcomes. Finally, we study the relation of Outcome Indistinguishability and Multicalibration of statistics (beyond the mean) and relate our findings to the recent work of Jung et al. (COLT 2021) on Moment Multicalibration. We find an equivalence between Outcome Indistinguishability and Multicalibration that is more subtle than in the binary case and sheds light on the techniques employed by Jung et al. to obtain Moment Multicalibration.

Keywords: Indistinguishability, Computational Learning Theory, Generative Models

1. Introduction

Typically, the goal of individual outcome prediction is framed as finding a "best-fit" hypothesis. For instance, the agnostic PAC model (Valiant, 1984; Kearns et al., 1994) formalizes the goal of learning by fixing a hypothesis class $\mathcal{H}$, and then asking to find a function f that achieves loss that competes with the best hypothesis $h^* \in \mathcal{H}$ within the given class. While this paradigm has been remarkably effective for developing the theoretical foundations of supervised machine learning, a downside of the PAC model is that it only considers the overall performance of a predictor, and not the performance across subpopulations.

In contexts where models are trained to make predictions about people, a major concern with overall loss minimization is that all of the errors may concentrate on a historically-marginalized subpopulation. These concerns are far from hypothetical, and have been observed empirically in diverse application domains including commercial facial recognition software (Buolamwini and Gebru, 2018) and widely-used medical risk predictors (Obermeyer et al., 2019; Barda et al., 2021). Given the increasing use of machine learning in making consequential decisions, there is a growing need to develop formal tools to reason about the quality of predictors that go beyond overall loss minimization.

Recently, Dwork et al. (2021) introduced a new abstraction for supervised learning—*Outcome Indistinguishability*—in the context of predicting boolean outcomes. Risk assessment tools produce scores for individual instances, such as the chance that *this* student will graduate within 4 years, or the likelihood that *this* tumor will metastasize under a given course of treatment, that are treated as probabilities. But what is the probability of a non-repeatable event? How should we specify the goal of these predictive algorithms? Motivated to ground and strengthen the guarantees of individual risk prediction, Outcome Indistinguishability (OI) articulates the goals of learning in the language of computational indistinguishability. Informally, OI requires that the distribution over outcomes suggested by an outcome predictor be computationally indistinguishable from the true outcome distribution.

In more detail, OI reasons about the similarity of two joint distributions on individual-outcome pairs. The first distribution $(X, Y^*) \sim \mathcal{D}^*$, referred to as *Nature*, is the true distribution of individuals X and associated outcomes Y^* in the world. The second distribution is induced by a predictor $\tilde{p} : \mathcal{X} \to [0, 1]$ that estimates the probability of positive outcome for each individual. To draw a sample from the modeled distribution $(X, \tilde{Y}) \sim \mathcal{D}(\tilde{p})$, the individual X is sampled from Nature's true marginal distribution over individuals; then, conditioned on the individual X, the outcome $\tilde{Y} \sim \mathrm{Ber}(\tilde{p}(X))$ is resampled according to the Bernoulli distribution with probability according to the prediction $\tilde{p}(X)$. Given these two distributions, OI defines a requirement on predictors: a predictor $\tilde{p}$ is outcome indistinguishable from Nature $\mathcal{D}^*$ if the distributions $\mathcal{D}^*$ and $\mathcal{D}(\tilde{p})$ are indistinguishable.

To specify the indistinguishability condition formally, OI is parameterized by a family of efficient distinguisher algorithms $\mathcal{A}$. The distinguisher algorithms $\mathcal{A} \subseteq \{\mathcal{X} \times \mathcal{Y} \times [0, 1] \to \{0, 1\}\}$ receive as input an individual $x \in \mathcal{X}$, an outcome $y \in \mathcal{Y}$, and the prediction $\tilde{p}(x) \in [0, 1]$ on the sampled individual. Formally, for some $\varepsilon \geq 0$, a predictor $\tilde{p}$ is $(\mathcal{A}, \varepsilon)$-outcome indistinguishable[1] if for all $A \in \mathcal{A}$ the acceptance probability of A on samples from $\mathcal{D}^*$ and $\mathcal{D}(\tilde{p})$ are within ε of one another.

$$\left| \mathop{\mathbf{Pr}}_{(X,Y^*)\sim\mathcal{D}^*} \left[\, A(X, Y^*; \tilde{p}(X)) = 1 \,\right] - \mathop{\mathbf{Pr}}_{(X,\tilde{Y})\sim\mathcal{D}(\tilde{p})} \left[\, A(X, \tilde{Y}; \tilde{p}(X)) = 1 \,\right] \right| \leq \varepsilon$$

Importantly, OI requires that the predictions $\tilde{p}$ fool all distinguishers $A \in \mathcal{A}$ simultaneously. In this way, OI provides a computational-theoretic perspective on the subtle issue of individual-level outcome probabilities: an OI predictor $\tilde{p}$ defines a model for "individual probabilities" of positive outcomes ($\tilde{p}(x)$ for each individual $x \in \mathcal{X}$) that cannot be refuted by efficient tests, captured by $\mathcal{A}$.

[1] We focus on the variant of OI that Dwork et al. (2021) refer to as "sample-access" OI.

On a technical level, Dwork et al. (2021) investigate the complexity of learning OI predictors, formalizing the close relationship to a notion of fairness in prediction called *multicalibration* (Hébert-Johnson et al., 2018). They show that OI is capable of capturing the agnostic PAC model, and that learning OI predictors is closely related to the task of (weak) agnostic learning. Recent works even demonstrate a formal sense in which the OI guarantee is stronger than an arbitrary agnostic learning guarantee (Gopalan et al., 2021a). But perhaps most importantly, the work on OI establishes indistinguishability as an effective qualitative tool for reasoning about formal properties of predictors. The OI perspective has already seen application in answering questions in computational learning theory that, on the surface, do not look like questions about indistinguishability (Rothblum and Yona, 2021). In all, Outcome Indistinguishability defines an extensible framework for studying formal guarantees in supervised learning.

1.1. Our Contributions

In this work, we define and study a generalized framework for modeling random outcomes through the lens of Outcome Indistinguishability. Moving beyond Bernoulli outcomes, Outcome Indistinguishability becomes a condition about *generative outcome models* $\mathcal{M} : \mathcal{X} \to \Delta(\mathcal{Y})$ that map each individual $x \in \mathcal{X}$ to a probability distribution $\mathcal{M}(x)$ over possible outcomes $y \in \mathcal{Y}$. The generalization maintains the intuition behind the original formulation of OI: a generative outcome model $\mathcal{M}$ is Outcome Indistinguishable with respect to a family of distinguishers $\mathcal{A}$ if the joint individual-outcome distribution induced by $\mathcal{M}$ is indistinguishable from Nature's true joint distribution.

While intuitive, generalizing OI from the Bernoulli case to general random outcomes is not simply a syntactic transformation. When predicting Bernoulli outcomes, the entire distribution is captured by the estimated probability; thus, given a predictor $\tilde{p} : \mathcal{X} \to [0, 1]$, the generative model is implicitly specified by sampling $\tilde{Y} \sim \mathrm{Ber}(\tilde{p}(X))$. In contrast, without parametric assumptions on the outcome distributions, to obtain OI in full generality, we need to reason explicitly about the choice of generative model. The choice of how to generate outcomes given a set of predictions results in different possible definitions of OI.

Our contributions include formalizing the "right" generalization of OI to non-Bernoulli random outcomes, establishing the feasibility of these notions via learning algorithms, and drawing a further connection between OI and *multicalibration* of statistics. We define a sequence of OI variants, which formalize the intuition behind outcome indistinguishability for different "plausible" models of Nature. These models of Nature differ in how they quantify over the outcome generation procedure, ranging from generative outcome models that fully-specify individuals' outcome distributions to models that only require the learner to estimate certain statistics of the outcome distributions. Our results generalize the initial work of Dwork et al. (2021), and set up a general framework for discussing OI generative models and OI statistic predictors.

1.2. Generative Outcome Models

We begin by describing the most generic notion of OI we define, which we refer to as *Generative OI*. Generative OI reasons about fully-specified generative outcome models $\mathcal{M} : \mathcal{X} \to \Delta(\mathcal{Y})$, where the prediction $\mathcal{M}(X)$ gives an explicit description of the predicted probability distribution of Y given X. Because we require a complete description of the outcome probability distributions, this notion is capable of capturing any conceivable variant of OI. In particular, Generative OI can easily be adapted to handle both discrete or continuous random outcomes.

For Generative OI, we define the modeled distribution $\mathcal{D}(\mathcal{M})$ as follows: given an individual X (sampled from Nature's distribution on individuals), we sample the outcome $\tilde{Y} \sim \mathcal{M}(X)$ directly from the predicted distribution. Thus, given a family of distinguishers $\mathcal{A}$, $(\mathcal{A}, \varepsilon)$-Generative OI requires that $\mathcal{D}(\mathcal{M})$ fools each $A \in \mathcal{A}$:

$$\left| \underset{(X,Y^*)\sim\mathcal{D}^*}{\mathbf{Pr}} \left[\, A(X, Y^*; \mathcal{M}(X)) = 1 \,\right] - \underset{(X,\tilde{Y})\sim\mathcal{D}(\mathcal{M})}{\mathbf{Pr}} \left[A(X, \tilde{Y}; \mathcal{M}(X)) = 1 \right] \right| \leq \varepsilon.$$

In other words, a generative outcome model $\mathcal{M}$ satisfies the OI condition if the conditional distribution on outcomes "fools" every distinguisher $A \in \mathcal{A}$ into accepting with the same probability as it accepts on true outcomes. We require that this indistinguishability condition holds even when the distinguisher may inspect the predicted probability distribution $\mathcal{M}(X)$ on sampled individual X.

As such, this notion is very restrictive, and also depends significantly on the way that we specify $\mathcal{M}(X)$ as a probability distribution. The complexity of obtaining Generative OI is tightly coupled with the representation of outcome distributions. In Section 5, we demonstrate a generic framework for learning Generative OI models, under some key assumptions on the underlying outcome distributions. Concretely, this establishes the feasibility of Generative OI for generating outcomes over a large discrete domain.

We show that a generalization of the Multiplicative Weights algorithm can be used to learn Generative OI models. In this statement of the theorem, we give concrete bounds for the sample and time complexities for learning and evaluating the Generative OI models for general discrete outcome domains. In fact, the theorem is a special case of a more general result, where we show how to abstract out the essential components of our analysis, to learn Generative OI Models for a broad class of random variables. Given the generality at which we study this notion of OI, in many applications, it may be possible to exploit domain-specific structure in the distribution of outcomes in order to improve the complexity over the generic solution.

Parametric OI. In general, specifying a complete generative model for each individual's outcome distribution may be overly-complex. To combat this complexity, a natural approach for specifying the outcome distributions would model each outcomes to be drawn from a fixed parametric family $\mathcal{M}_{\boldsymbol{\theta}}$. Such a strategy can be very convient, because the predictor only has to specify estimates of the parameters of the family. Then, given the predicted parameters for a given individual X, the outcome can be sampled according to the distribution corresponding to these parameters. In Section 4.1, we define a special case of Generative OI that we call *Parametric OI*, which is a condition on a parameter predictor $\tilde{\theta} : \mathcal{X} \to \mathbb{R}^d$. A predictor $\tilde{\theta}$ is Parametric OI, if model that samples outcomes $\tilde{Y} \sim \mathcal{M}_{\boldsymbol{\theta}}(\tilde{\theta}(X))$ according to the parametric family with parameters $\tilde{\theta}(X)$ is Generative OI.

$$\left| \underset{(X,Y^*)\sim\mathcal{D}^*}{\mathbf{Pr}} \left[A(X, Y^*; \tilde{\theta}(X)) = 1 \right] - \underset{(X,\tilde{Y})\sim\mathcal{D}(\mathcal{M}_{\boldsymbol{\theta}}(\tilde{\theta}))}{\mathbf{Pr}} \left[A(X, \tilde{Y}; \tilde{\theta}(X)) = 1 \right] \right| \leq \varepsilon$$

As a concrete example, imagine modeling outcomes as Gaussian random variables, where the parameter predictor returns the estimated mean and variance of a Gaussian outcome; that is, where $\tilde{\theta}(X) = \left[\tilde{\mu}(X), \tilde{\sigma}^2(X)\right]$, and the generative model samples outcomes $\tilde{Y} \sim \mathcal{N}(\tilde{\mu}(X), \tilde{\sigma}^2(X))$ according to the normal distribution. Additionally, note that we can view the original formulation of OI of Dwork et al. (2021) as Parametric OI for the Bernoulli distribution, where the parameter predictor $\tilde{\theta}(X) = \tilde{p}(X)$ simply predicts the probability of positive outcome for the individual X.

While Parametric OI offers an appealing way to compress the generative outcome model into a small number of predicted parameters, unfortunately, it is not always feasible. In the case where our choice of parametric family is misspecified, it could be the case that some $A \in \mathcal{A}$ is always capable of distinguishing Nature's outcomes from the modeled outcomes. For instance, if we elect to model outcomes as Gaussian random variables, but Nature's outcomes are far from following a Gaussian, there will be efficient distinguishers that always have nontrivial distinguishing advantage.

1.3. OI Predictors of Statistics

The challenges associated with Generative OI (and infeasibility of Parametric OI) motivates additional notions of OI, that are meaningful in the context of predictors that only partially specify individuals' outcome distributions. We formalize the idea of partial specification through statistic prediction. We think of a statistic $\boldsymbol{\rho}$ simply as some function of the distribution of outcomes. The most basic statistic considered is the mean $\boldsymbol{\rho} = \boldsymbol{\mu}$; the original characterization of OI can be viewed as OI for mean predictors $\tilde{\mu} : \mathcal{X} \to \mathbb{R}$. We generalize this basic predictor to handle d-dimensional statistics. For concreteness, a useful example to have in mind is a 2-dimensional statistic predictor that estimates the mean and variance of the outcome distribution.

$$\tilde{\rho}_2(x) = \big(\tilde{\mu}(x), \tilde{\sigma}^2(x)\big)$$

Given a choice of d-dimensional statistic $\boldsymbol{\rho}$ of the outcome distribution, we consider what it means for a statistic predictor $\tilde{\rho} : \mathcal{X} \to \mathbb{R}^d$ to be OI. Given a statistic predictor, we consider the set of generative outcome models that exhibit the statistics predicted by $\tilde{\rho}(x)$ for all individuals $x \in \mathcal{X}$. Formally, we say that a model $\mathcal{M}$ is *individually-consistent* with a statistic predictor $\tilde{\rho}$ if for each individual $x \in \mathcal{X}$, the statistics $\boldsymbol{\rho}$ of $\mathcal{M}(x)$ equal $\tilde{\rho}(x)$.

Existential OI. With this setup, we can define the first notion of OI for statistic predictors, which we call *Existential OI*. The intuition for the definition is to define OI for predictors by (existentially) quantifying over models that are invidually-consistent with the predictor. Specifically, we say that a statistic predictor $\tilde{\rho}$ is Existential-OI if there exists a generative outcome model $\mathcal{M}_{\tilde{\rho}}$, which is individually-consistent with $\tilde{\rho}$, that fools the distinguishers $A \in \mathcal{A}$:

$$\left| \mathop{\mathbf{Pr}}_{(X,Y^*)\sim\mathcal{D}^*}\left[\, A(X,Y^*;\tilde{\rho}(X)) = 1 \,\right] - \mathop{\mathbf{Pr}}_{(X,\tilde{Y})\sim\mathcal{D}(\mathcal{M}_{\tilde{\rho}})}\left[\, A(X,\tilde{Y};\tilde{\rho}(X)) = 1 \,\right] \right| \le \varepsilon$$

Existential OI is a condition defined for statistic predictors, but implies that the statistics are consistent with some generative outcome model that simultaneously fools every test $A \in \mathcal{A}$. In other words, the predictions globally "explain" the statistics of the distributions on outcomes in a way that plausibly could come from Nature's model. In the case of predicting the mean and variance $\tilde{\rho}_2$, Existential OI requires that there exists a mechanism for generating individual's outcomes $\tilde{Y} \sim \mathcal{M}_{\tilde{\rho}_2}(x)$ for $x \in \mathcal{X}$ such that the mean and variance of $\tilde{Y}$ equals the predicted values specified by $\tilde{\rho}_2(x) = (\tilde{\mu}(x), \tilde{\sigma}^2(x))$ and no distinguisher $A \in \mathcal{A}$ can meaningful tell the difference between samples Y^* and $\tilde{Y}$.

In Section 4.2, we develop an understanding the properties of Existential OI. We establish that Existential OI is a relaxation of Generative OI, showing how to turn any procedure for learning Generative OI models into one for learning Existential OI predictors. The argument is intuitive, but is also subtle because the "types" of Existential OI predictors and Generative OI models differ.

Oblivious Distinguishers. In Existential OI, even though the condition is defined for predictors of statistics, the distinguishers $A \in \mathcal{A}$ can, in principle, test properties of the resulting generative outcome model that are not captured by the predicted statistics. This fact makes it important to quantify over the choice of individually-consistent model. In our final notion of OI, we consider a natural restriction on the family of distinguishers that, intuitively, restricts the distinguishers to only "care about" the statistic being predicted. For a statistic $\boldsymbol{\rho}$, for any probability distribution $F_Y \in \Delta(\mathcal{Y})$, we denote the true value of the statistic over $Y \sim F_Y$ as $\boldsymbol{\rho}\{F_Y\}$. We say that a distinguisher A is $\boldsymbol{\rho}$-oblivious if for all individuals $x \in \mathcal{X}$ and predicted statistics $\nu \in \mathbb{R}^d$, the acceptance probability of $A(x, Y; \nu)$ on an outcome $Y \sim F_Y$ is a function of $\boldsymbol{\rho}\{F_Y\}$. That is, there exists a function $h_{x,\nu} : \mathbb{R}^d \to [0, 1]$ such that for any outcome distribution $F_Y \in \Delta(\mathcal{Y})$,

$$\Pr_{F_Y}[\, A(x, Y; \nu) = 1 \,] = h_{x,\nu}(\boldsymbol{\rho}\{F_Y\}).$$

For instance, if our statistic predictor $\tilde{\rho}_2$ estimates the mean and the variance of the outcomes, then an oblivious distinguisher's acceptance probability can be parameterized by the individual X and the predictions $\tilde{\rho}_2(X) = (\tilde{\mu}(X), \tilde{\sigma}^2(X))$, but may only depend on the sampled outcome $Y \sim F_Y$ through its mean and variance $\boldsymbol{\mu}\{F_Y\}, \boldsymbol{\sigma^2}\{F_Y\}$. We say that a statistic predictor $\tilde{\rho}$ is $(\mathcal{A}, \varepsilon)$-Oblivious OI if it fools every $\boldsymbol{\rho}$-oblivious $A \in \mathcal{A}$. Intuitively, oblivious distinguishers focus all of their attention on the statistics, and not other aspects of the outcome distribution. As such, obliviousness is a natural restriction to make on statistics predictors.

We show that Oblivious OI is a strict relaxation of Existential OI: non-oblivious tests can enforce global consistency of statistic predictors that are not captured by oblivious predictors. While weaker, Oblivious OI gives the intuitively strong guarantee that no distinguisher $A \in \mathcal{A}$ can refute the statistic predictions on its own. For each such distinguisher, there exists a consistent generative model that produces outcomes that are indistinguishable from Nature, so the statistics can only appear inconsistent when considering at multiple distinguishers at once.‘

Oblivious OI and Multicalibrated Statistics. Our main technical result investigates connections between OI and multicalibration in the context of general random outcomes. The work of Dwork et al. (2021) established a tight computational equivalence between (the main variant of) Outcome Indistinguishability and multicalibration, a notion introduced in the study of algorithm fairness of predictors (Hébert-Johnson et al., 2018). Given this equivalence, it is natural to ask whether there is a similar connection between the generalized version of OI and an appropriate generalization of multicalibration. Typically, multicalibration has been studied in the context of predicting the probability of a binary outcome. A key exception to this general trend is the work of Jung et al. (2021), who initiated a study of multicalibration for statistics beyond mean estimation. Jung et al. (2021) show how their generalization—moment multicalibration—suffices to provide Chebyshev-style inequalities for uncertainty quantification based on the *predicted moments*, rather than the true moments of the underlying distribution on outcomes.

Building on the work of Jung et al. (2021), we generalize moment multicalibration and define multicalibration in the context of estimating general statistics. A key algorithmic and analytic step in achieving moment multicalibration involves conditioning on the predicted mean, in order to obtain *linearization* of the moments. We demonstrate an equivalence between statistic predictors that satisfy Oblivious OI and our novel generalization of multicalibration, that crucially relies on linearization.

Theorem 1 (Informal) *For any linearizing statistic, for any class of functions $\mathcal{C}$, there exists a family of oblivious distinguishers $\mathcal{A}$ such that $(\mathcal{A}, \varepsilon)$-Oblivious OI implies $(\mathcal{C}, \alpha)$-multicalibration for $\alpha \leq O(\varepsilon)$. For any family of oblivious distinguishers $\mathcal{A}$, there exists a class of functions $\mathcal{C}$, such that $(\mathcal{C}, \alpha)$-multicalibration implies $(\mathcal{A}, \varepsilon)$-OI for $\varepsilon \leq O(\alpha)$.*

The analysis of Oblivious OI and multicalibrated statistics further clarifies the approach used by Jung et al. (2021) to achieve moment multicalibration, showing in a sense, their techniques are necessary. In particular, we show that without mean-conditioning, OI (even Existential OI) is incapable of providing any recovery guarantee for central moments.

2. Technical Overview

The formal definitions of each variant of OI and the relationship between notions is presented in Section 4. Here, we begin with a high-level overview of the algorithm for learning Generative OI models. We focus our overview on the setting where the outcomes come from a large discrete domain. Then, we discuss the connection between Oblivious OI and Multicalibrated statistics. In particular, we highlight the technical concept of linearization and how it plays a key role in establishing the equivalence. The results are presented in full detail in Section 5 and Section 6, respectively.

2.1. Learning OI Models.

Our goal is as follows: for any family of distinguishers $\mathcal{A}$ and constant $\varepsilon > 0$, given a small set of samples from Nature's distribution $\mathcal{D}^*$, return an $(\mathcal{A}, \varepsilon)$-Generative OI Model $\tilde{\mathcal{M}} : \mathcal{X} \to \Delta(\mathcal{Y})$. First, we recall the high-level approach that Dwork et al. (2021) use to learn Bernoulli OI predictors $\tilde{p} : \mathcal{X} \to [0, 1]$. The learning algorithm follows a simple intuition: if there is some $A \in \mathcal{A}$ that distinguishes between Nature $\mathcal{D}^*$ and the model of Nature $\mathcal{D}(\tilde{p})$, use A to update $\tilde{p}$; else, $\mathcal{D}^*$ and $\mathcal{D}(\tilde{p})$ are indistinguishable, so we are done. This strategy can be viewed through different lenses, as a form of boosting or gradient descent, and is closely connected to the strategy used by Hébert-Johnson et al. (2018) to learn multicalibrated predictors.

An Abstraction for Learning OI Models. Intuitively, when we move beyond Bernoulli OI, the specifics of the learning strategy and the complexity of operations may depend intimately on the characteristics of the outcome space $\mathcal{Y}$ (e.g., discrete vs. continuous, dimensionality, smoothness of density, etc.) and on the particular OI variant we aim to achieve (e.g., generative OI vs. parametric OI). One of our contributions is presenting a general and flexible framework that can be adapted to the many variants of OI that we study in this work (and, hopefully, beyond).

A key issue is the *representation* of outcome distributions assigned for each individual, as we discuss below. Once a representation is fixed, we identify two main algorithmic tasks that suffice for running our learning algorithm: sampling from an individual's outcome distribution, and reweighting an individual's outcome distribution (for the algorithm's update step, see above).

Fixing an outcome domain $\mathcal{Y}$, we define a collection $\mathcal{R}$ of *representations* of distributions in $\Delta(\mathcal{Y})$. For a distribution $F_Y \in \Delta(\mathcal{Y})$, we take $R(F_Y)$ to be its representation (which is in the set $\mathcal{R}$). We also assume that the representation can be used to sample from the distribution, using a sample generation procedure $\mathcal{G}$ (see below). For instance, the representation may be an explicit histogram, listing the probability of each element in $\mathcal{Y}$ (up to some discretization). Or the representation may be implicit, e.g. a small circuit that can be used to sample from F_Y.

The following two algorithmic tasks are central to our OI learning algorithm:

- **Sample Generation:** Given the representation of $F_Y \in \Delta(\mathcal{Y})$, sample from F_Y.

  $\mathcal{G} : \mathcal{R} \to \mathcal{Y}$ is a randomized map, where $\mathcal{G}(R(F_Y))$ should produce an outcome distribution that is equal to F_Y (or very close to it in total variation distance, e.g. because of discretization issues).

- **Reweighting:** Given the representation of a distribution over $\mathcal{Y}$ and a predicate $B : \mathcal{Y} \to \{0, 1\}$, produce the representation of a reweighted distribution, where the probabilities of elements in $\mathcal{Y}$ that satisfy the predicate are increased, and the probabilities of elements that do not satisfy the predicate are decreased. We use a multiplicative reweighting, whose magnitude is controlled by a parameter η.

  We take $\mathcal{W}^B : R(\Delta(\mathcal{Y})) \times [-1, 1] \to R(\Delta(\mathcal{Y}))$ to be an oracle procedure, where $B : \mathcal{Y} \to \{0, 1\}$ computes the characteristic function of a subset of $\mathcal{Y}$. Given a reweighting parameter $\eta \in [-1, 1]$ and a representation $r \in \mathcal{R}$ of the distribution $F_Y = \mathcal{G}(r)$, $\mathcal{W}^B(r; \eta)$ returns a representation $r' \in \mathcal{R}$ of a new distribution $G_Y = \mathcal{G}(r')$, where:

$$\Pr_{G_Y}[\,Y \in B\,] \propto e^{\eta} \cdot \Pr_{F_Y}[\,Y \in B\,] \qquad\qquad \Pr_{G_Y}[\,Y \notin B\,] \propto \Pr_{F_Y}[\,Y \notin B\,].$$

Concretely, one can think of $\mathcal{W}$ as abstracting away the computational procedure needed to normalize distributions. In the discrete case, this may involve a summation over domain elements. In the continuous case, the reweighting procedure may involve numerical integration. In either case, it is important that $\mathcal{W}$ is given as a *uniform* computational procedure, which is fixed for all updates we may want to make; this allows us to build up the generative outcome model. The reweighting assumption is a direct generalization of the multiplicative weights update used in the learning algorithm for Bernoulli OI (Dwork et al., 2021). Abstracting these two tasks allows us to present a unified treatment of learning for the many variants OI variants we consider in this work. We discuss different representations below, but we begin with a concrete instantiation: a learning algorithm for Generative OI models that uses the above framework.

Learning Generative OI Models. Intuitively, the algorithm follows the same high-level strategy of that of Dwork et al. (2021). The algorithm starts with a naive predictor $\tilde{\mathcal{M}}$, which maps individuals in $\mathcal{X}$ to representations of distributions over $\mathcal{Y}$. The distributions are all initialized to return a chosen prior $\tilde{\mathcal{M}}(x) = \mathcal{P}_Y$ (say, the uniform distribution over $\mathcal{Y}$). Then, the algorithm iteratively identifies whether there is any $A \in \mathcal{A}$ that distinguishes between Nature $\mathcal{D}^*$ and the current modeled distribution $\mathcal{D}(\tilde{\mathcal{M}})$ with absolute advantage greater than ε. If not, then we are done: the failed search certifies that $\tilde{\mathcal{M}}$ is already $(\mathcal{A}, \varepsilon)$-OI. If we find some $A \in \mathcal{A}$ that distinguishes successfully, then we update the model $\tilde{\mathcal{M}}$ to bring A's acceptance probability closer to its acceptance probability on $\mathcal{D}^*$.

With this algorithm in mind, several remarks are in order. First, the algorithm accesses Nature's distribution $\mathcal{D}^*$ and the distribution of the learned model $\tilde{\mathcal{M}}$ in a fairly restricted way: in each iteration, the algorithm evaluates the acceptance probability of each $A \in \mathcal{A}$ on $\mathcal{D}^*$ and on $\mathcal{D}(\tilde{\mathcal{M}})$. This estimation can be implemented using random sampling. For Nature, the algorithm can use true samples from $\mathcal{D}^*$, and for the model of Nature, it uses the sample generation procedure $\mathcal{G}$ to obtain samples from $\mathcal{D}(\tilde{\mathcal{M}})$. If $\tilde{\mathcal{M}}$ is not OI, then the algorithm updates $\tilde{\mathcal{M}}$ using the reweighting procedure. We emphasize that these two steps involving Sample Generation and reweighting are the only ways that the algorithm interacts with $\tilde{\mathcal{M}}$.

```
Generative OI Learning Algorithm (Overview)
Given. Family of distinguishers $\mathcal{A}$; advantage $\varepsilon$
Initialization. For all $x \in \mathcal{X}$, initialize $\tilde{\mathcal{M}}(x)$ to prior $\mathcal{P}_Y$
Iterate. for $t = 1, \ldots, T$
    • Let $\varepsilon_A$ be the (signed) distinguishing advantage
          $$\varepsilon_A = \mathop{\mathbf{Pr}}_{\mathcal{D}^*}\left[ A(X, Y^*; \tilde{\mathcal{M}}(X)) = 1 \right] - \mathop{\mathbf{Pr}}_{\mathcal{D}(\tilde{\mathcal{M}})}\left[ A(X, \tilde{Y}; \tilde{\mathcal{M}}(X)) = 1 \right]$$
      If $\max_{A \in \mathcal{A}} \ |\varepsilon_A| \leq \varepsilon$, return $\tilde{\mathcal{M}}$.
      Else, let $A_t \leftarrow \mathrm{argmax}_{A \in \mathcal{A}} \ |\varepsilon_A|$.
    • Implicitly update the predictor $\tilde{\mathcal{M}}$. For each $x \in \mathcal{X}$, the reweighted predictor outputs the following distribution:
        1. Let $B_x = \left\{ y \in \mathcal{Y} : A_t(x, y; \tilde{\mathcal{M}}(x)) = 1 \right\}$
        2. $\tilde{\mathcal{M}}(x) \leftarrow \mathcal{W}^{B_x}(\tilde{\mathcal{M}}(x), \eta)$
```

The algorithm does *not* maintain an explicit representation of the outcome distribution for each $x \in \mathcal{X}$—and this is crucial! Instead, the distribution for each $x \in \mathcal{X}$ is described implicitly using the list of distinguishers $(A_1, \ldots, A_k)$ found in each update iteration, as well as the magnitudes (and directions) of the updates $(\varepsilon_{A_1}, \ldots, \varepsilon_{A_k})$. Thus, in each intermediate iteration, and in its final output, the algorithm uses a model $\tilde{\mathcal{M}} : \mathcal{X} \to \Delta(\mathcal{Y})$, where the distribution described by $\tilde{\mathcal{M}}(x)$ is obtained by starting with the prior $\mathcal{P}_Y$ and reweighting according to the distinguishers and magnitudes in prior iterations. In this sense, we will really incorporate copies of $\mathcal{W}^B$ into the learned model $\tilde{\mathcal{M}}$.

In all, we show that with minimal assumptions on Nature's true outcome distribution, it is possible to obtain the very strong outcome indistinguishability guarantee of Generative OI. Still, to obtain such a strong notion, we require fairly strong requirements on the ability to generate samples and reweight the underlying distribution. To obtain Generative OI in full generality, the costs can be bounded modestly in terms of data, but can be costlier in terms of computation. Thus, in future applications of OI, it is advisable to bring domain-specific assumptions (e.g., smoothness or sparsity) that may aid in the computational efficiency. Exploring the efficiency of Generative OI in more structured outcome distributions seems to be an interesting direction for future research.

In Section 5, we describe the algorithm formally. Then, we analyze key quantities, like the iteration complexity and how this informs quantities like the sample complexity, time complexity of evaluation, and time complexity of learning. While a direct generalization of the approach of Dwork et al. (2021), the analysis in the generic version is much more subtle. We analyze the algorithm in generality, in terms of the generation and reweighting time complexities. Then, we instantiate the general bound for the special case of learning generative outcome models for outcomes that are drawn from a large discrete distribution.

2.2. Oblivious OI captures Multicalibration

Multicalibration was originally introduced by Hébert-Johnson et al. (2018) as a notion of fairness in binary prediction. Informally, a predictor $\tilde{p} : \mathcal{X} \to [0, 1]$ is multicalibrated over a collection of subpopulations, if for every subpopulation $S \subseteq \mathcal{C}$ in the collection, $\tilde{p}$ is well-calibrated even when restricting our attention to the individuals in S. More technically, we work with the following generalization of multicalibration[2] that is defined in terms of functions, instead of subpopulations. For a class of functions $\mathcal{C} \subseteq \{\mathcal{X} \times [0, 1] \to [0, 1]\}$ and an approximation parameter $\alpha \geq 0$, a predictor $\tilde{p}$ is $(\mathcal{C}, \alpha)$-multicalibrated if for all $c \in \mathcal{C}$

$$\left| \mathop{\mathbf{E}}_{\mathcal{D}^*} \left[\, c(X, \tilde{p}(X)) \cdot (Y - \tilde{p}(X)) \, \right] \right| \leq \alpha.$$

That is, the predictions $\tilde{p}$ are α-accurate in expectation, even when we restrict our attention to the images of functions in $\mathcal{C}$. Equivalently, by linearity of expectation we can write this condition in terms of the true probability of positive outcomes, $p^*(x) = \mathbf{Pr}\left[\, Y = 1 \mid X = x \, \right]$.

$$\left| \mathop{\mathbf{E}}_{\mathcal{D}^*_X} \left[\, c(X, \tilde{p}(X)) \cdot (p^*(X) - \tilde{p}(X)) \, \right] \right| \leq \alpha.$$

In the case of Bernoulli OI, Dwork et al. (2021) establish a tight computational equivalence between OI and multicalibration. The reductions show how to translate between calibration tests and distinguishers. Intuitively, the key idea is to design distinguishers that accept with probability proportional to the statistic of interest, e.g., the expectation of the outcome on the image of a given function $c \in \mathcal{C}$

$$\mathbf{Pr}\left[\, A^c(X, Y; \tilde{p}(X)) = 1 \, \right] \propto \mathbf{E}\left[\, c(X, \tilde{p}(X)) \cdot Y \, \right].$$

Similarly, given a distinguisher A, the goal is to construct a function c_A such that the multicalibration condition on c enforces indistinguishability by A.

Multicalibrated Statistics. When moving to predicting statistics beyond Bernoulli parameters, we need to formalize the idea of calibration and multicalibration of general statistics. This simple definitional question turns out to be subtle.

Suppose we are interested in learning a statistic predictor $\tilde{\rho} : \mathcal{X} \to \mathbb{R}^d$ for some d-dimensional statistic $\boldsymbol{\rho}$. Denote by $\rho^* : \mathcal{X} \to \mathbb{R}^d$ true statistics according to Nature. Then, a natural generalization of multicalibration to general statistics would be the following condition that checks the accuracy of $\tilde{\rho}$ in expectation over a class of functions $\mathcal{C}$. To test calibration of d-dimensional statistics, it is natural to take $\mathcal{C} \subseteq \left\{\mathcal{X} \times \mathbb{R}^d \to \mathbb{R}^d\right\}$ to be a collection of vector-valued functions, assumed to be bounded in ℓ_1-norm. We say that $\tilde{\rho}$ is $(\mathcal{C}, \alpha)$-multicalibrated, if for all $c \in \mathcal{C}$,

$$\left| \mathop{\mathbf{E}}_{\mathcal{D}^*} \left[\, \langle c(X, \tilde{\rho}(X)), (\rho^*(X) - \tilde{\rho}(X)) \rangle \, \right] \right| \leq \alpha.$$

[2] In the subsequent technical sections, we show how to implement the original framework of multicalibration as an instance of this generalization.

While natural, this generalization of multicalibration encounters issues when working with general statistics. For arbitrary statistics $\boldsymbol{\rho}$, this multicalibration condition is information-theoretically infeasible, even to certify. In particular, it is not obvious—from a small sample—how to evaluate $\rho^*(X)$. Crucially, the Bernoulli case relied on linearity of expectation, to use outcomes Y, as a surrogate for the probability parameters $p^*(X)$.

Building off of the work on moment multicalibration by Jung et al. (2021), we show how a technical condition on the statistic $\boldsymbol{\rho}$—*linearization*—allows us to make progress. In effect, linearization ensures that the generalization of calibration to general statistics works in the same manner as standard Bernoulli calibration. Specifically, for any joint distribution $\mathcal{D}$ over individual-outcome pairs, we require that the expected statistic value (averaged over individuals) is equal to the statistic on the marignal distribution of outcomes,

$$\mathop{\mathbf{E}}_{\mathcal{D}_X}\left[\ \boldsymbol{\rho}\left\{\mathcal{D}_{Y|X}\right\}\ \right] = \boldsymbol{\rho}\left\{\mathcal{D}_Y\right\}$$

where $\mathcal{D}_X$ is the marginal on individuals, $\mathcal{D}_{Y|X}$ is the conditional outcome distribution given an individual, and $\mathcal{D}_Y$ is the marginal on outcomes. Linearization allows us to reason about the predicted statistics, not just in terms of individuals, but in terms of the average value across a group of individuals, which we can estimate from a small sample of outcomes.

An equivalence with Oblivious OI. As in the Bernoulli case, we connect the idea of multicalibration to OI. In particular, for linearizing statistics $\boldsymbol{\rho}$, we show a computational equivalence between $(\mathcal{C}, \alpha)$-multicalibration and $(\mathcal{A}, \varepsilon)$-oblivious OI. The first direction—implementing multicalibration using OI—follows similarly to the Bernoulli case. Given a function $c \in \mathcal{C}$, we design a distinguisher A^c to accept with probability proportional to $\mathbf{E}\left[\ \langle c(X, \tilde{\rho}(X)), \rho^*(X)\rangle\ \right]$, when given samples from Nature, and $\mathbf{E}\left[\ \langle c(X, \tilde{\rho}(X)), \tilde{\rho}(X)\rangle\ \right]$ when given modeled samples. As such, the distinguishing advantage for any such A^c upper bounds the multicalibration violation. Note also that the acceptance probability is a function of the statistic of interest, and thus, the distinguishers are oblivious.

To gain intuition for the construction, consider a concrete example where $\boldsymbol{\rho}$ is simply the mean $\boldsymbol{\mu}$ of the outcome. Then, as in the Bernoulli case, given some $c \in \mathcal{C}$, we can define a randomized distinguisher $A^c_{\boldsymbol{\mu}}$ that accepts with probability proportional to the expected mean. For simplicitly, assume that $c(x, \tilde{\rho}(x)) \in [0, 1]$ and $y \in [0, 1]$. Then, we define $A^c_{\boldsymbol{\mu}}$ as follows:

$$A^c_{\boldsymbol{\mu}}(x, y; \tilde{\rho}(x)) = \begin{cases} 1 & \text{w.p. } c(x, \tilde{\rho}(x)) \cdot y \\ 0 & \text{o.w.} \end{cases}$$

Similarly, consider a different example where $\boldsymbol{\rho}$ is the second (non-central) moment $\boldsymbol{\mu_2}$. We can implement a distinguisher that accepts with probability proportional to this moment.

$$A^c_{\boldsymbol{\mu_2}}(x, y; \tilde{\rho}(x)) = \begin{cases} 1 & \text{w.p. } c(x, \tilde{\rho}(x)) \cdot y^2 \\ 0 & \text{o.w.} \end{cases}$$

Given a collection of functions $\mathcal{C}$, we can build the corresponding family of distinguishers $\mathcal{A} = \{A^c : c \in \mathcal{C}\}$. It's not hard to see that any mean or moment predictor that fools these distinguishers must also be multicalibrated with respect to the original class $\mathcal{C}$. Following the intuition in these examples, our reduction shows how to implement multicalibration for any linearizing statistic into an OI condition. As with these examples, the distinguishers are randomized, but efficient: each

$A^c \in \mathcal{A}$ makes a single oracle call to the associated $c \in \mathcal{C}$. We present and analyze this reduction formally in Section 6.2.1.

The reverse direction—implementing OI using multicalibration—requires analytic tools that differ significantly from the Bernoulli case. In particular, it is not immediately obvious how to write the acceptance probability of an arbitrary $\boldsymbol{\rho}$-oblivious distinguisher as a calibration condition. To this end, we show that for a linearizing statistic $\boldsymbol{\rho}$, the acceptance probability of a $\boldsymbol{\rho}$-oblvious distinguisher satisfies a certain separability condition. In particular, we show that without loss of generality, we can assume that the acceptance probability of such an A is a linear function of the outcome statistic. Specifically, for any fixed individual $x \in \mathcal{X}$ and prediction $\tilde{\rho}(x) = \nu$, for an outcome distribution $F_Y \in \Delta(\mathcal{Y})$, the acceptance probability of A can be written as follows.

$$\Pr_{F_Y}[\, A(x, Y; \nu) = 1 \,] \propto \langle c_A(x, \nu), \boldsymbol{\rho}\{F_Y\} \rangle$$

The argument itself is technical, but follows from some simple observations. First, the acceptance probability for any distinguisher can be re-written as an expectation over F_Y.

$$\Pr_{F_Y}[\, A(x, Y; \nu) = 1 \,] = \mathbf{E}[\, A(x, Y; \nu) \,]$$

Further, by $\boldsymbol{\rho}$-obliviousness, this acceptance probability must also be a function of the statistic of the outcome distribution $\boldsymbol{\rho}\{F_Y\}$. We show that linearization of $\boldsymbol{\rho}$ implies that $\boldsymbol{\rho}\{F_Y\}$ can be written as an expectation over F_Y. As such, the acceptance probability is an expectation over F_Y, which is a function of $\boldsymbol{\rho}\{F_Y\}$. In combination, these observations imply the that the acceptance probability must be a linear function of $\boldsymbol{\rho}\{F_Y\}$, implying the separability result.

Given the separability condition, we can easily rewrite the OI constraints as multicalibration constraints over $\tilde{\rho}$. In particular, we construct the collection of functions $\mathcal{C} = \{c_A : A \in \mathcal{A}\}$; then, $(\mathcal{C}, \alpha)$-multicalibration implies $(\mathcal{A}, \varepsilon)$-Oblvious OI. Some care is needed to ensure that the distinguishing advantage ε translates smoothly from the approximation parameter α. The full analysis and reduction are presented in Section 6.2.2.

Note that unlike the reduction from multicalibration to OI (and the reductions that connected OI to multicalibration in the Bernoulli case), the complexity of this reduction is less obvious. Our techniques here are analytical rather than explicit, demonstrating that for every $A \in \mathcal{A}$ there exists some corresponding function $c_A \in \mathcal{C}$, such that multicalibration over $\mathcal{C}$ implies $\mathcal{A}$-OI. The constructed function c_A does not use A as an oracle, but is simply guaranteed to exist by the assumed properties of $\boldsymbol{\rho}$ and $\mathcal{A}$. Thus, it is not immediately clear how to bound the complexity of $\mathcal{C}$ compared to $\mathcal{A}$. Nevertheless, we argue that $\mathcal{C}$ cannot be significantly more complex than $\mathcal{A}$. We show that there is a way to "decode" the function c_A using oracle calls to A, assuming $\boldsymbol{\rho}$ satisfies some natural non-degeneracy conditions. The number of calls needed scales with the dimension d of the statistic and the desired approximation of c_A.

On Mean-Conditioning for Moment Multicalibration. Finally, we remark that our study of Oblivious OI and statistic Multicalibration sheds light on the work of Jung et al. (2021) on Moment Multicalibration. In language of the present work, Jung et al. (2021) study how to obtain multicalibration for a statistic $\boldsymbol{\rho}$ that contains central moments of the outcome. A key technical hurdle to obtaining moment multicalibration is that central moments do not linearize on their own. Instead, to obtain algorithms for learning multicalibrated moment predictors, Jung et al. (2021) condition not only on the predicted moments, but also on the mean. They show that mean-conditioned moments do linearize, and thus, mean-conditioned moment multicalibrated predictors can be learned from random samples. We show a sweeping converse to this result. Roughly speaking, when predicting central moments, any OI guarantee (including those implementable by multicalibration) with respect to a family of distinguishers that do not "condition" on the mean, is meaningless. We give a construction in Section 6.3 demonstrating that even Existential OI can be fooled by central moments that are predicted to be uniformly 0 when the distinguisher class $\mathcal{A}$ does not take the predicted mean of Y as input. Collectively, our results show that, in some strong sense, the seemingly-technical tools developed by Jung et al. (2021) are necessary for obtaining central moment multicalibration.

2.3. Related Works and Discussion

The study of OI grew out of the literature on fairness in prediction tasks. To address the limitations of learning fair predictors via constrained loss minimization (identified first by Dwork et al. (2012)), a recent line of work, initiated independently by Hébert-Johnson et al. (2018) and Kearns et al. (2018), has proposed an alternative paradigm for achieving "multi-group" fairness. Multicalibration has emerged as a prominent notion from this line of work, requiring that predictions be well-calibrated, not simply overall, but even when restricting attention to structured subpopulations. Rather than fixating on a singular, global objective defined by a fixed hypothesis class $\mathcal{H}$, multicalibration is parameterized by a collection of subpopulations $\mathcal{C}$, which represent groups of individuals that can be efficiently-identified by the individuals' data. By tuning the choice of $\mathcal{C}$ based on the available data and computational resources, multicalibration allows for precise control of the performance across subpopulations, and thus, the downstream fairness of the predictions.

Since the initial work of Hébert-Johnson et al. (2018) that defined and described learning algorithms for obtaining multicalibrated predictors, multicalibration has been studied in a growing number works within the literature on fairness and, more generally, computational learning theory (Kim et al., 2018, 2019; Dwork et al., 2019; Shabat et al., 2020; Dwork et al., 2021; Jung et al., 2021; Gupta et al., 2021; Gopalan et al., 2021b; Rothblum and Yona, 2021; Gopalan et al., 2021a). Unlike most notions of fairness, multicalibration does not exhibit an accuracy-fairness tradeoff and can be used to obtain high-quality predictors that perform well across diverse subpopulations (Kim et al., 2019). In fact, recent work of Gopalan et al. (2021a) demonstrates that multicalibration is sufficient to guarantee a novel learning desideratum, dubbed *omniprediction*, which requires a predictor be an agnostic learner not for a single loss function, but simultaneously for a collection of loss functions. They show that any multicalibrated predictor is also an omnipredictor, formally establishing a sense in which multicalibration can be viewed as a strengthening of the standard agnostic PAC solution concept. Their work also recognizes a connection between multicalibration and the boosting by branching programs learning paradigm (Mansour and McAllester, 2002; Kalai, 2004; Kalai and Servedio, 2005).

3. Preliminaries

To begin, we establish notation and review the prior work on Outcome Indistinguishability.

Individuals and outcomes. Let $\mathcal{X}$ denote the space of individual inputs. Throughout, we assume that $\mathcal{X}$ is discrete and that each individual $x \in \mathcal{X}$ has a finite, known representation; without loss of generality, we may assume $\mathcal{X} \subseteq \{0,1\}^n$ for some $n \in \mathbb{N}$. Let $\mathcal{Y}$ denote the outcome sample space and $\Delta(\mathcal{Y})$ denote probability distributions over elements in $\mathcal{Y}$. We consider both discrete and continuous outcome distributions. For discrete outcome spaces, $\Delta(\mathcal{Y})$ is the set of all discrete distributions. For continuous outcomes, we typically restrict our attention to the case where $\mathcal{Y} = [-B, B] \subseteq \mathbb{R}$ for some bound $B \in \mathbb{R}$. For any continuous domain, we consider events defined by half-open intervals $(a, b]$ for $a, b \in \mathbb{R} \cup \{-\infty, \infty\}$ and take $\Delta(\mathcal{Y})$ to be Borel measures. Generically, we use $\mathcal{D}$ to denote a joint distribution supported on $\mathcal{X} \times \mathcal{Y}$. We use $\mathcal{D}_X$ to denote the marginal distribution over individuals, $\mathcal{D}_Y$ to denote the marginal distribution over outcomes, and $\mathcal{D}_{Y|X}$ to denote the conditional outcome distribution given X.

Generative outcome models. The key objects of study in OI are generative outcome models

$$\mathcal{M} : \mathcal{X} \to \Delta(\mathcal{Y})$$

that map individual inputs to outcome probability distributions. In particular, for a given individual $x \in \mathcal{X}$, a model evaluates to $\mathcal{M}(x) = F_Y$ for some outcome distribution $F_Y \in \Delta(\mathcal{Y})$. The underlying representation of outcome distributions in $\Delta(\mathcal{Y})$, and thus models $\mathcal{M}$, is important but varies throughout the presentation based on the setting. We defer details of representing outcome distributions to the subsequent technical sections.

In addition to mapping from individuals to outcome distributions, we need a way to map from outcome distributions to random outcomes. To this end, we let the *generator*

$$\mathcal{G} : \Delta(\mathcal{Y}) \to \mathcal{Y}$$

be a randomized map from probability distributions to sampled outcomes. In particular, given any distribution $F_Y \in \Delta(\mathcal{Y})$, we use $Y = \mathcal{G}(F_Y)$ to denote a random draw of an outcome $Y \sim \mathcal{F}_Y$ sampled according to the specified distribution. We assume that each call to $\mathcal{G}$ uses independent internal randomness.

Statistics. Throughout, we specify and estimate properties of outcome distributions, which we refer to as statistics. We typically denote the statistics of interest as $\boldsymbol{\rho}$. Formally, for $d \in \mathbb{N}$, a d-dimensional statistic

$$\boldsymbol{\rho} : \Delta(\mathcal{Y}) \to \mathbb{R}^d$$

is a function mapping distributions to real-valued vectors. Examples of statistics include functions of distributions like the mean, median, or variance. A natural d-dimensional statistic $\boldsymbol{\rho} = \langle \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_d \rangle$ is the vector of the first d moments of the distribution $\boldsymbol{\mu}_k \{F_Y\} = \ell_k \cdot \mathbf{E}_{F_Y} \left[Y^k \right] + \tau_k$, for some choice of $\ell_k, \tau_k \in \mathbb{R}$ to scale and shift the moments to live in $[-1, 1]$.

Generically, we can define statistics over $\mathbb{R}^d$, but for the sake of computational estimation, it is useful to have some prior bound on the magnitude of the statistics' values. Throughout, we assume that statistics are coordinate-wise scaled and shifted to live in the ℓ_∞ unit ball.[3] We use the following notation for the ℓ_1 and ℓ_∞ balls for d-dimensional vectors.

$$\mathcal{B}_1^d = \left\{ \omega \in \mathbb{R}^d : \|\omega\|_1 \le 1 \right\} \qquad \mathcal{B}_\infty^d = \left\{ \omega \in \mathbb{R}^d : \|\omega\|_\infty \le 1 \right\} = [-1, 1]^d$$

For any outcome distribution $F_Y \in \Delta(\mathcal{Y})$ we use $\boldsymbol{\rho}\{F_Y\} \in \mathcal{B}_\infty^d$ to denote the true value of the statistic on the distribution.

[3] The choice to make the range of the prior on $\boldsymbol{\rho}$ symmetric around 0 is arbitrary and for convenience. Equally, we could imagine scaling and shifting statistics to lie in $[0, 1]^d$ or to have a different bound for each individual statistic.

3.1. Outcome Indistinguishability

Outcome Indistinguishability is a requirement on a generative outcome model that stipulates that the generated outcomes should be "indistinguishable" from the true outcome model, which we refer to as *Nature*. In fact, OI refers to a family of related definitions that vary based on the way distinguishers may access the predictions and samples from the outcome distributions. Dwork et al. (2021) originally studied the notion of OI in the context of predicting binary outcomes. We discuss the high-level framework of OI and review the variants defined in the prior work (Dwork et al., 2021).

Models of Nature. Outcome indistinguishability considers two joint distributions on individual-outcome pairs, Nature and a Model of Nature. Throughout, we denote samples from Nature as follows.

$$(X, Y^*) \sim \mathcal{D}^*$$

We assume Nature $\mathcal{D}^*$ is a valid joint distribution over individuals $\mathcal{X}$ and outcomes $\mathcal{Y}$, but as in the agnostic PAC learning model, we make no further assumptions about the relationship between X and the distribution over outcomes $\mathcal{D}^*_{Y|X}$.

Given Nature, the goal of outcome indistinguishability is to learn a generative outcome model $\mathcal{M} : \mathcal{X} \to \Delta(\mathcal{Y})$ that "looks like" Nature. Specifically, every generative outcome model $\mathcal{M}$ induces a model of Nature where we draw an individual $X \sim \mathcal{D}^*_X$ from Nature's marginal distribution over individuals, then resample their outcome $\tilde{Y} = \mathcal{G}(\mathcal{M}(X))$ according to the modeled outcome distribution. In this way, we can represent modeled samples as

$$(X, \tilde{Y}) \sim \mathcal{D}(\mathcal{M})$$

where $X \sim \mathcal{D}^*_X$ and $\tilde{Y} \sim \mathcal{M}(X)$.

Distinguishers and OI. If our goal is to model Nature, then ideally, the distributions $\mathcal{D}^*$ and $\mathcal{D}(\mathcal{M})$ would be similar. OI formalizes this intuitive goal for learning through the language of computational indistinguishability as follows. Consider a distinguisher algorithm $A : \mathcal{X} \times \mathcal{Y} \to \{0, 1\}$ that takes as input an individual and outcome pair and maps the input to either 0 or 1. For any given distinguisher A, we say the distinguishing advantage of A between Nature $\mathcal{D}^*$ and the modeled distribution $\mathcal{D}(\mathcal{M})$ is the difference in the acceptance probabilities

$$\varepsilon_A = \left| \Pr_{\mathcal{D}^*} [\, A(X, Y^*) = 1 \,] - \Pr_{\mathcal{D}(\mathcal{M})} \left[A(X, \tilde{Y}) = 1 \right] \right|$$

when the outcome is generated by Nature Y^* and by the model of Nature $\tilde{Y} = \mathcal{G}(\mathcal{M}(X))$. We say that the two distributions are ε-indistinguishable to A, when the advantage is upper bounded by $\varepsilon_A \leq \varepsilon$. Outcome Indistinguishability requires indistinguishability, not just for a single distinguisher A, but instead simultaneously for a rich family $\mathcal{A} \subseteq \{A : \mathcal{X} \times \mathcal{Y} \to \{0, 1\}\}$ of distinguisher algorithms.

Definition 2 (Dwork et al. (2021)) *For a collection of distinguisher algorithms $\mathcal{A}$, a model $\mathcal{M}$ is $(\mathcal{A}, \varepsilon)$-outcome indistinguishable from Nature if $\mathcal{D}^*$ and $\mathcal{D}(\mathcal{M})$ are ε-indistinguishable to each $A \in \mathcal{A}$.*

If $\mathcal{A}$ is taken to be the class of all (possibly-inefficient) statistical distinguishers, then the only way to obtain $\mathcal{A}$-OI would be to learn Nature's probability law in statistical distance. Without any assumptions on the complexity of $\mathcal{D}^*$, however, learning in statistical distance is information-theoretically impossible from a bounded sample. To obtain a feasible notion, Dwork et al. (2021) take $\mathcal{A}$ to be a class of distinguishers that can be implemented within some bounded complexity. On a technical level, to achieve OI with a bounded sample complexity, the main requirement is that the cardinality (or some equivalent measure of complexity of the class of $\mathcal{A}$) is bounded.

Definition 2 is actually a generalization of the original formulation of Dwork et al. (2021). In this work, Outcome Indistinguishability is defined in terms of predictors of a Bernoulli probability $\tilde{p} : \mathcal{X} \to [0, 1]$, which implicitly induces a generative outcome model. Specifically, any predictor $\tilde{p}$ implies a model of Nature $\mathcal{D}(\tilde{p})$: sample an individual $X \sim \mathcal{D}^*_X$, then resample their outcome $\tilde{Y} \sim \mathrm{Ber}(\tilde{p}(X))$. With this notion of modeled distribution, we recover the basic definition of OI for binary outcomes.

Definition 3 (Bernoulli OI) *For a collection of distinguisher algorithms $\mathcal{A}$, a predictor $\tilde{p} : \mathcal{X} \to [0, 1]$ is $(\mathcal{A}, \varepsilon)$-outcome indistinguishable from Nature if*

$$\left| \Pr_{\mathcal{D}^*}\left[\, A(X, Y^*; \tilde{p}(X)) = 1 \,\right] - \Pr_{\mathcal{D}(\tilde{p})}\left[\, A(X, \tilde{Y}; \tilde{p}(X)) = 1 \,\right] \right| \leq \varepsilon.$$

Variants of OI. With this general framework in place, Dwork et al. (2021) defined a number of specific variants of OI that differ in strength and complexity. The different variants of OI differ based on their access to the predictions of the model of Nature. For the following four variants of distinguishers, a predictor $\tilde{p}$ is $(\mathcal{A}, \varepsilon)$-outcome indistinguishable if:

- *No-Access:* Distinguishers only observe the individual and outcome and no prediction $\tilde{p}(X)$.

$$\left| \Pr_{\mathcal{D}^*}\left[\, A(X, Y^*) = 1 \,\right] - \Pr_{\mathcal{D}(\tilde{p})}\left[\, A(X, \tilde{Y}) = 1 \,\right] \right| \leq \varepsilon$$

- *Sample-Access:* OI as defined in Definition 3.

- *Oracle-Access:* Distinguishers have oracle access to $\tilde{p}$.

$$\left| \Pr_{\mathcal{D}^*}\left[\, A^{\tilde{p}}(X, Y^*) = 1 \,\right] - \Pr_{\mathcal{D}(\tilde{p})}\left[\, A^{\tilde{p}}(X, \tilde{Y}) = 1 \,\right] \right| \leq \varepsilon$$

- *Code-Access:* Distinguishers receive an explicit description of the code (circuit) implementing $\tilde{p}$.

$$\left| \Pr_{\mathcal{D}^*}\left[\, A(X, Y^*; \langle \tilde{p} \rangle) = 1 \,\right] - \Pr_{\mathcal{D}(\tilde{p})}\left[\, A(X, \tilde{Y}; \langle \tilde{p} \rangle) = 1 \,\right] \right| \leq \varepsilon$$

In each level, the distinguishers receive increasing degrees of access to the predictor $\tilde{p}$. The primary results of Dwork et al. (2021) characterize the complexity of obtaining each level of OI.

In addition to defining and characterizing variants of OI based on access to $\tilde{p}$, Dwork et al. (2021) also show that the definitions extend naturally when distinguishers are given access to multiple individual-outcome pairs. For sufficiently expressive classes of distinguishers $\mathcal{A}$, a standard hybrid argument shows that single-sample OI implies multiple-sample OI with an increase in ε that grows linearly with the number of samples.

Multicalibration and OI. Beyond defining Outcome Indistinguishability and its variants, the main result of Dwork et al. (2021) shows a tight connection between OI and the notion of multicalibration, defined by Hébert-Johnson et al. (2018) in the context of learning fair predictors. The results show a tight computational equivalence between the concepts of multiaccuracy and multicalibration and No-Access and Sample-Access OI, respectively. These notions of multi-group fairness have been studied in a growing list of works in the algorithmic fairness literature and beyond (Hébert-Johnson et al., 2018; Kim et al., 2018, 2019; Dwork et al., 2019; Shabat et al., 2020; Jung et al., 2021; Gupta et al., 2021; Gopalan et al., 2021b; Rothblum and Yona, 2021; Gopalan et al., 2021a).

Multiaccuracy and multicalibration define a set of accuracy constraints on a predictor that need to hold, not simply overall, but even when we restrict our attention to structured subpopulations of the domain. A convenient generalization that captures both multiaccuracy and multicalibration[4] is the following technical reformulation of multicalibration, that defines the "subpopulations" in terms of a class of functions $\mathcal{C} \subseteq \{\mathcal{X} \times [0,1] \to [0,1]\}$.

Definition 4 (Multicalibration, generalized) *For a class of functions $\mathcal{C}$ and an approximation $\alpha > 0$, a predictor $\tilde{p} : \mathcal{X} \to [0,1]$ is $(\mathcal{C}, \alpha)$-multicalibrated over a distribution $\mathcal{D}^*$ if for all $c \in \mathcal{C}$*

$$\left| \mathop{\mathbf{E}}_{\mathcal{D}^*} \left[\, c(X, \tilde{p}(X)) \cdot (Y^* - \tilde{p}(X)) \, \right] \right| \leq \alpha.$$

Intuitively, we can strengthen the guarantees of multicalibration by increasing the complexity of the functions $c \in \mathcal{C}$. For example, we may think of $\mathcal{C}$ as the class of functions implementable within some concrete complexity class (decision trees of depth 3, linear functions with bounded weights, or Neural Networks of bounded size, etc.). Dwork et al. (2021) show that how to efficiently reduce any class of subpopulations $\mathcal{C}$ into a class of distinguishers $\mathcal{A}$ (and vice versa) such that $(\mathcal{C}, \alpha)$-multicalibration is equivalent to $(\mathcal{A}, \varepsilon)$-Sample-Access-OI.

4. Defining Outcome Indistinguishability

Here, we define variants of OI for general random outcomes. The variants differ in the way they quantify over the outcome generation process in the modeled distribution.

4.1. Outcome Indistinguishability for Generative Models

The first notion of OI which we define requires the most of the predictor. In this notion of *Generative OI*, we require the predictor to be a complete generative outcome model $\mathcal{M} : \mathcal{X} \to \Delta(\mathcal{Y})$ that for each individual, returns a fully-specified probability distribution over outcomes. Intuitively, the model $\mathcal{M}$ is OI if $\mathcal{M}$ is indistinguishable from Nature's true conditional outcome distribution $\mathcal{D}^*_{Y|X}$ given individuals.

[4] For the reader familiar with multiaccuracy and multicalibration, as defined by Hébert-Johnson et al. (2018), we argue that this generalization is without loss. To implement multiaccuracy, we can define a class of functions $\mathcal{C}$, where for each subpopulations $S \subseteq \mathcal{X}$, we define a function c_S that ignores $\tilde{p}(X)$ and returns 1 if and only if $X \in S$. To implement the original subpopulation formulation of multicalibration, we can use a similar class of functions that returns 1 if and only if $X \in S$ and $\tilde{p}(X) \approx v$. Finally, we must also scale the choice of α to account for the fact that Definition 4 requires bounded absolute error, rather than error normalized by the probability of $X \in S$ and $\tilde{p}(X) \approx v$.

Definition 5 (Generative OI) *Fix a family of distinguishers $\mathcal{A} \subseteq \{\mathcal{X} \times \mathcal{Y} \times \Delta(\mathcal{Y}) \to \{0,1\}\}$ and an advantage $\varepsilon > 0$. A generative outcome model $\mathcal{M} : \mathcal{X} \to \Delta(\mathcal{Y})$ is $(\mathcal{A}, \varepsilon)$-generative-outcome indistinguishable if for all $A \in \mathcal{A}$,*

$$\left| \mathop{\mathbf{Pr}}_{(X,Y^*)\sim\mathcal{D}^*} \left[\, A(X, Y^*; \mathcal{M}(X)) \,\right] - \mathop{\mathbf{Pr}}_{X\sim\mathcal{D}^*_X} \left[\, A(X, \mathcal{G}(\mathcal{M}(X)); \mathcal{M}(X)) \,\right] \right| \leq \varepsilon.$$

Here, we define Generative OI abstractly, for outcome models that output a probability distribution $\mathcal{M}(x) \in \Delta(\mathcal{Y})$. This notion is mathematically well-defined, but for an effective computational notion, we must also fix an explicit representation for distributions in $\Delta(\mathcal{Y})$. We delay our discussion of the choice of representation until Section 5, where we turn to learning Generative OI Models.

In a sense, Generative OI is the most natural generalization of the notion defined by Dwork et al. (2021), because it is a condition on a predictor, which specifies the complete generative model for each individual's outcome. As such, we don't need to quantify over the way we generate outcomes to be consistent with the model. Further, the notion is always definitionally feasible, as we can always take $\mathcal{M}$ to be $\mathcal{D}^*_{Y|X}$.

Parametric generative models. As discussed, a technically subtle but important point in defining Generative-OI is how to represent the outcome distribution for a given individual. Rather than working in full generality, in many applications, it may make sense to restrict our attention to distributions that have an efficient and explicit representation. Parametric families are one way to represent distributions succinctly.

Definition 6 (Parametric Family) *A model $\mathcal{M}_{\boldsymbol{\theta}} : \mathbb{R}^d \to \Delta(\mathcal{Y})$ implements a parametric family if there exists a d-dimensional statistic $\boldsymbol{\theta}$ such that for each parameter setting $t \in \mathbb{R}^d$, there exists a unique, explicit probability distribution $F_{Y;t} \in \Delta(\mathcal{Y})$ where*

$$\mathcal{M}_{\boldsymbol{\theta}}(t) = F_{Y;t}$$
$$\boldsymbol{\theta}\left\{F_{Y;t}\right\} = t.$$

In other words, given a setting $t \in \mathbb{R}^d$ of the parameter $\boldsymbol{\theta}$, the outcome distribution is perfectly specified. Common parametric families for modeling outcomes include the Bernoulli, Gaussian, and Poisson distributions (all of which are exponential families).

Remark 7 *We make two remarks about the definition of parametric families:*

- *First, note that our usage of this notion refers only to the conditional outcome distribution, given an individual. In particular, we do not make any assumptions across individuals, in contrast to parametric models of prediction (e.g., well-specified linear regression). Our assumption that the conditional outcome distribution of each individual follows a parametric family is substantially weaker.*

- *Second, in addition to the requirements listed in Definition 6, we require that a parametric family corresponds to a collection of "explicit" probability distributions. In this work, we use this term to mean there is an efficient (randomized) procedure that, given $t \in \mathbb{R}^d$, produces samples from the probability distribution $F_{Y;t}$. For instance, we could model the conditional outcome distributions as Gaussians, assuming there is an efficient procedure that given a predicted mean $\tilde{\mu}(x)$ and variance $\tilde{\sigma}^2(x)$ produces a sample $\tilde{Y} \sim \mathcal{N}(\tilde{\mu}(x), \tilde{\sigma}^2(x))$.*

Of course, if we have no prior knowledge of Nature's distribution of outcomes, parametric models may not be capable of achieving indistinguishability from Nature. In particular, infeasibility could arise due to quantitative limitations of the family (the best-fit parametric model may not fit the outcome distribution to sufficient accuracy) and qualitative differences between the family and Nature's outcomes (e.g., if we use a continuous parametric family to model a discrete Nature). That said, in the case where we do believe it's reasonable to model the outcomes via a parametric model, we may aim to learn an OI generative outcome model, by learning a parameter predictor $\tilde{\theta} : \mathcal{X} \to \mathbb{R}^d$ and generating outcomes according to the family $\mathcal{M}_{\boldsymbol{\theta}}$.

Definition 8 (Parametric OI) *For a parameteric family $\mathcal{M}_{\boldsymbol{\theta}}$ with d-dimensional parameter $\boldsymbol{\theta}$, fix a family of distinguishers $\mathcal{A} \subseteq \left\{\mathcal{X} \times \mathcal{Y} \times \mathbb{R}^d \to \{0,1\}\right\}$ and an advantage $\varepsilon > 0$. A parameter predictor $\tilde{\theta} : \mathcal{X} \to \mathbb{R}^d$ is $(\mathcal{A}, \varepsilon)$-parametric-outcome indistinguishable from Nature $\mathcal{D}^*$ if the composition $\mathcal{M}_{\boldsymbol{\theta}} \circ \tilde{\theta}$ is $(\mathcal{A}, \varepsilon)$-generative-outcome indistinguishable.*

In other words, we require that the distribution induced by the parametric family $\mathcal{M}_{\boldsymbol{\theta}}$ using parameters $\tilde{\theta}$ satisfy the OI condition,

$$\left| \Pr_{(X,Y^*)\sim\mathcal{D}^*}\left[\, A(X, Y^*; \tilde{\theta}(X)) = 1 \,\right] - \Pr_{X\sim\mathcal{D}^*_X}\left[\, A(X, \mathcal{G}(\mathcal{M}_{\boldsymbol{\theta}}(\tilde{\theta}(X))); \tilde{\theta}(X)) = 1 \,\right] \right| \leq \varepsilon.$$

In this sense, parametric-OI is a special case of generative-OI, where we fix the model to be a parametric family $\mathcal{M}_{\boldsymbol{\theta}}$. If Nature's conditional distribution over outcomes $\mathcal{D}^*_{Y|X}$ is actually determined by the parametric family $\mathcal{M}_{\boldsymbol{\theta}}$, then there is some true parameter predictor $\theta^* : \mathcal{X} \to \mathbb{R}^d$ such that $\mathcal{D}^*_{Y|X} = \mathcal{M}_{\boldsymbol{\theta}}(\theta^*)$. In such a setting, parametric OI is guaranteed to be feasible: taking $\tilde{\theta} = \theta^*$ guarantees that the constraints are satisfied.

Note also, that the original formulation of OI given in (Dwork et al., 2021), can be viewed as a special case of parametric-OI, for the parametric Bernoulli distribution. In the setting of (Dwork et al., 2021), the goal is to learn a predictor $\tilde{p} : \mathcal{X} \to [0,1]$ that predicts the probability that a given individual's outcome is positive; then, given the prediction for an individual X, modeled outcomes are sampled according to $\tilde{Y} \sim \text{Ber}\left(\tilde{p}(X)\right)$.

4.2. Outcome Indistinguishability for Statistic Predictors

While generative OI gives a strong notion of indistinguishability from Nature, the constraints may be too demanding in some circumstances. Learning a complete generative outcome model requires specifying a complete outcome distribution for each individual. In many cases, due to sampling and computational constraints, such a complete model may be out of reach, and the learner may instead elect to estimate statistics of individuals' outcome distributions. Here, we define notions of outcome indistinguishability for partial models, which consist of a predictor that for each individual returns an estimate of the statistic on the individual's outcome distribution $\boldsymbol{\rho}\left\{\mathcal{D}^*_{Y|X}\right\}$.

Intuitively, for a statistic $\boldsymbol{\rho}$, a predictor $\tilde{\rho} : \mathcal{X} \to \mathbb{R}^d$ is outcome indistinguishable from Nature if it is consistent with some generative model that produces indistinguishable outcomes. Unlike generative OI, a predictor $\tilde{\rho}$ only specifies values for statistics $\boldsymbol{\rho}$ of the individuals' outcome distributions, rather than a full probability distribution. To discuss this notion of OI, we need to formalize the idea of models that are consistent with a predictor $\tilde{\rho}$. Such models take as input both an individual $x \in \mathcal{X}$ and a prediction $\tilde{\rho}(x) = \nu \in \mathbb{R}^d$, and satisfy the following definition.

Definition 9 (Individually-Consistent Model) *For a d-dimensional statistic $\boldsymbol{\rho}$, a model $\mathcal{M} : \mathcal{X} \times \mathbb{R}^d \to \Delta(\mathcal{Y})$ is individually consistent over $\boldsymbol{\rho}$ if for all $x \in \mathcal{X}$, for all $\nu \in \mathbb{R}^d$*

$$\boldsymbol{\rho} \left\{ \mathcal{M}(x; \nu) \right\} = \nu.$$

In other words, given a predictor $\tilde{\rho}$, for each individual $x \in \mathcal{X}$, an individually-consistent model $\mathcal{M}$ evaluated on the individual-prediction pair $\mathcal{M}(x; \tilde{\rho}(x))$, returns a distribution where the statistic $\boldsymbol{\rho}$ of the modeled outcome distribution equals the predicted statistic $\tilde{\rho}(x)$.

Existential OI. With this definition in place, we can define notions of "existential" OI: a predictor $\tilde{\rho}$ satisfies this variant of OI if there exists a consistent model that generates indistinguishable outcomes based on the statistics specified in $\tilde{\rho}$. Here, the family of distinguishers $\mathcal{A} \subseteq \left\{ \mathcal{X} \times \mathcal{Y} \times \mathbb{R}^d \to \{0,1\} \right\}$ take in the individual, outcome, and d-dimensional statistic.

Definition 10 (Existential OI) *For a d-dimensional statistic $\boldsymbol{\rho}$, fix a family of distinguishers $\mathcal{A}$ and an advantage $\varepsilon > 0$. A predictor $\tilde{\rho} : \mathcal{X} \to \mathbb{R}^d$ is $(\mathcal{A}, \varepsilon)$-existential-outcome indistinguishable from Nature $\mathcal{D}^*$ if there exists an individually-consistent model $\mathcal{M} : \mathcal{X} \to \Delta(\mathcal{Y})$, such that for all $A \in \mathcal{A}$,*

$$\left| \Pr_{(X,Y^*) \sim \mathcal{D}^*} \left[A(X, Y^*; \tilde{\rho}(X)) = 1 \right] - \Pr_{X \sim \mathcal{D}^*_X} \left[A(X, \mathcal{G}(\mathcal{M}(X; \tilde{\rho}(X))); \tilde{\rho}(X)) = 1 \right] \right| \leq \varepsilon.$$

In other words, a statistic predictor is Existential OI if there exists a single, global generative outcome model $\mathcal{M}$ that is consistent with the statistics and satisfies the indistinguishability conditions. In this sense, it suggests that there is a generative outcome model that exhibits the predicted statistics that fools the distinguishers.

In existential OI, the distinguishers receive only the predicted statistics $\tilde{\rho}(X)$, not the full predicted distribution $\mathcal{M}(X)$. It is tempting to think that the restricted access to predictions implies that the distinguishers are simply a function of the statistic of interest, but this intuition is wrong. Recall, the distinguishers receive the outcomes directly. By using these outcomes, distinguishers can depend nontrivially on the distribution of the outcome, not just the predicted statistics. For instance, consider a family of distinguishers that implement threshold tests, where for $\tau \in \mathbb{R}$

$$A_\tau(x, y; \nu) = \mathbf{1} \left[y \leq \tau \right].$$

Such thresholding tests can effectively be used to the marginal outcome distribution for closeness in CDF. As the CDF characterizes the entire distribution of the outcomes, then such tests depend on aspects of the distribution that go beyond any statistic that does not (at least approximately) characterize the entire distribution.

We observe that Generative OI can be used to obtain Existential OI, by taking a generative outcome model that fools the distinguishers in $\mathcal{A}$ and "flattening" it into a statistic predictor $\tilde{\rho}$.

Proposition 11 *Suppose $\mathcal{A} \subseteq \left\{ \mathcal{X} \times \mathcal{Y} \times \mathbb{R}^d \to \{0,1\} \right\}$ is a family of existential-distinguishers. There is an explicit family of generative-distinguishers $\mathcal{A}' \subseteq \{ \mathcal{X} \times \mathcal{Y} \times \Delta(\mathcal{Y}) \to \{0,1\} \}$ such that for any $\varepsilon > 0$, if a model $\mathcal{M} : \mathcal{X} \to \Delta(\mathcal{Y})$ is $(\mathcal{A}, \varepsilon)$-generative-OI, then, the statistic predictor $\tilde{\rho} = \boldsymbol{\rho} \left\{ \mathcal{M}(\cdot) \right\}$ is $(\mathcal{A}, \varepsilon)$-existential-OI.*

Proof Fix a family of distinguishers $\mathcal{A} \subseteq \left\{\mathcal{X} \times \mathcal{Y} \times \mathbb{R}^d \to \{0,1\}\right\}$. We derive a new family $\mathcal{A}'$ of distinguishers that take as input full probability distributions $F_Y \in \Delta(\mathcal{Y})$, but simulate the distinguishers in $\mathcal{A}$ using the relevant statistics of F_Y. Specifically, for each $A \in \mathcal{A}$, consider a distinguisher A', where for $x \in \mathcal{X}$, $y \in \mathcal{Y}$, and $F_Y \in \Delta(\mathcal{Y})$, A' is defined as follows:

$$A'(x, y, F_Y) = A(x, y, \boldsymbol{\rho}\left\{F_Y\right\}).$$

Then, if a model $\mathcal{M} : \mathcal{X} \to \Delta(\mathcal{Y})$ that is $(\mathcal{A}', \varepsilon)$-Generative OI, the statistics of $\mathcal{M}$

$$\tilde{\rho}(x) = \boldsymbol{\rho}\left\{\mathcal{M}(x)\right\}$$

are $(\mathcal{A}, \varepsilon)$-existential-OI. ■

In practice, to bound the complexity of the distinguishers and the computation of $\tilde{\rho}$, we need an estimation procedure for $\boldsymbol{\rho}\left\{\mathcal{M}(x)\right\}$. This introduces some technical complications. In particular, we either need the family of distinguishers $\mathcal{A}$ to be Lipschitz in the estimated statistics, so that small changes in the estimated statistics do not result in large changes in acceptance probability, or we need an effective deterministic procedure for computing $\boldsymbol{\rho}\left\{F_Y\right\}$, given the representation of F_Y.

An alternative quantification. We can better understand the guarantee of existential OI by considering an alternative quantification. Existential OI says that there exists an individually-consistent generative outcome model that fools every test in $\mathcal{A}$. A weaker notion would flip the quantification: for every test in $\mathcal{A}$, there exists some individually-consistent model that fools the test. Intuitively, this framing views each distinguisher $A \in \mathcal{A}$ as some an independent audit of the predicted statistics. Following this intuition, we define the following variant of OI, that captures the idea that no distinguisher $A \in \mathcal{A}$ can refute the predicted statistics on its own.

Definition 12 ($\forall\exists$-OI) *For a d-dimensional statistic $\boldsymbol{\rho}$, fix a family of distinguishers $\mathcal{A}$ and an advantage $\varepsilon > 0$. A predictor $\tilde{\rho} : \mathcal{X} \to \mathbb{R}^d$ is $(\mathcal{A}, \varepsilon)$-$\forall\exists$-outcome indistinguishable from Nature $\mathcal{D}^*$ if for all $A \in \mathcal{A}$, there exists an individually-consistent model $\mathcal{M}_A : \mathcal{X} \to \Delta(\mathcal{Y})$ such that,*

$$\left| \underset{(X,Y^*)\sim\mathcal{D}^*}{\mathbf{Pr}}\left[\, A(X, Y^*; \tilde{\rho}(X)) = 1 \,\right] - \underset{X\sim\mathcal{D}^*_X}{\mathbf{Pr}}\left[\, A(X, \mathcal{G}(\mathcal{M}_A(X; \tilde{\rho}(X))); \tilde{\rho}(X)) = 1 \,\right] \right| \leq \varepsilon.$$

Immediately, we can observe that existential OI is at least as strict as $\forall\exists$-OI: if there exists a single model that fools all distinguishers, this model can witness the constraint for each distinguisher individually. That is, a single consistent model $\mathcal{M}$ from Definition 10 satisfies the conditions required of the models $\mathcal{M}_A$ for each $A \in \mathcal{A}$ in Definition 12.

In fact, we argue that existential OI derives significant power by requiring global consistency: fooling many tests simultaneously is strictly harder than fooling them individually. That is, existential OI is a strictly stronger notion than $\forall\exists$-OI. This separation holds even if we simply consider distinguishers of outcomes: in the construction, we can assume that for every individual $x \in \mathcal{X}$ the outcome y is identically distributed.

Consider a pair of distinguishers $A_{1/2}$ and $A_{1/4}$, with the behaviors:

$$A_{1/2}(Y) = \mathbf{1}\left[\, Y \leq 1/2 \,\right]$$
$$A_{1/4}(Y) = \mathbf{1}\left[\, Y \leq 1/4 \,\right].$$

$A_{1/2}$ tests the probability that the outcome is at most $1/2$ and $A_{1/4}$ tests the probability that the outcome is at most $1/4$. Suppose our goal is to estimate the mean of the outcome, and suppose that Nature's outcome Y^* satisfies

$$\mathbf{Pr}\left[\, Y^* \leq 1/2 \,\right] = 1$$
$$\mathbf{Pr}\left[\, Y^* \leq 1/4 \,\right] = 1/2.$$

Suppose we estimate $\tilde{\mu} = 1/2$. It is not hard to see that $\tilde{\mu}$ does not fool both distinguishers simultaneously: there is no single model $\mathcal{M}$ that generates outcomes with mean $\boldsymbol{\mu}\left\{\tilde{Y}\right\} = 1/2$ where $\mathbf{Pr}\left[\, \tilde{Y} \leq 1/2 \,\right] = 1$ and $\mathbf{Pr}\left[\, \tilde{Y} \leq 1/4 \,\right] = 1/2$. Nevertheless, it is possible to fool each test separately while respecting $\boldsymbol{\mu}\left\{\tilde{Y}\right\} = 1/2$: to fool $A_{1/2}$, take the constant model that always outputs $\tilde{Y} = 1/2$; to fool $A_{1/4}$, take the model that outputs $\tilde{Y} \sim \mathrm{Ber}(1/2)$.

4.3. Oblivious Distinguishers

While existential OI is a condition on the predicted statistics, we have seen how it can enforce constraints that go beyond the statistics themselves. A natural restriction of existential OI would only consider distinguishers that constrain the statistics themselves. Here, we formalize the idea that a distinguisher only depends on the statistic of interest. We say that a distinguisher is *oblivious* with respect to a statistic $\boldsymbol{\rho}$ if its acceptance probability on an outcome sampled $Y \sim F_Y$ is a function of $\boldsymbol{\rho}\left\{F_Y\right\}$.

Definition 13 (Oblivious Distinguisher) *For a d-dimensional statistic $\boldsymbol{\rho}$, a distinguisher A is $\boldsymbol{\rho}$-oblivious if for all individuals $x \in \mathcal{X}$ and predicted statistics $\nu \in \mathbb{R}^d$, there exists a function $h_{x,\nu} : \mathbb{R}^d \to [0,1]$ such that for any outcome distribution $F_Y \in \Delta(\mathcal{Y})$,*

$$\mathop{\mathbf{Pr}}_{F_Y}\left[\, A(x, Y; \nu) = 1 \,\right] = h_{x,\nu}(\boldsymbol{\rho}\left\{F_Y\right\}).$$

For $\ell \in \mathbb{R}$, a $\boldsymbol{\rho}$-oblivious distinguisher is $(\boldsymbol{\rho}, \ell)$-Lipschitz if for all $x \in \mathcal{X}$ and $\nu \in \mathbb{R}^d$, for any $F_Y, F_{Y'} \in \Delta(\mathcal{Y})$,

$$\left|\, \mathop{\mathbf{Pr}}_{F_Y}\left[\, A(x, Y; \nu) = 1 \,\right] - \mathop{\mathbf{Pr}}_{F_{Y'}}\left[\, A(x, Y'; \nu) = 1 \,\right] \,\right| \leq \ell \cdot \left\| \boldsymbol{\rho}\left\{F_Y\right\} - \boldsymbol{\rho}\left\{F_{Y'}\right\} \right\|_\infty$$

Obliviousness captures the idea that the only aspect of the outcome distribution of Y that the distinguisher tests is the statistic $\boldsymbol{\rho}\left\{F_Y\right\}$. In other words, the acceptance probability of an oblivious distinguisher must be constant on individually-consistent models that are fed the same estimate of $\boldsymbol{\rho}$. Specifically, if A is $\boldsymbol{\rho}$-oblivious, then for any pair of individually-consistent models $\mathcal{M}, \mathcal{M}'$, for all $x \in \mathcal{X}$ and any $\nu, \nu' \in \mathbb{R}^d$

$$\mathbf{Pr}\left[\, A(x, \mathcal{G}(\mathcal{M}(x;\nu)); \nu') = 1 \,\right] = \mathbf{Pr}\left[\, A(x, \mathcal{G}(\mathcal{M}'(x;\nu)); \nu') = 1 \,\right]$$

over the randomness of $\mathcal{G}$. The observation motivates a final notion of outcome indistinguishability, that can be viewed as a $\forall\forall$-variant of OI.

Definition 14 (Oblivous OI) *For a d-dimensional statistic $\boldsymbol{\rho}$, fix a family of distinguishers $\mathcal{A} \subseteq \{\mathcal{X} \times \mathcal{Y} \times \mathbb{R}^d \to \{0,1\}\}$ and an advantage $\varepsilon > 0$. A predictor $\tilde{\rho} : \mathcal{X} \to \mathbb{R}^d$ is $(\mathcal{A}, \varepsilon)$-oblivious-outcome indistinguishable from Nature $\mathcal{D}^*$ if for all $\boldsymbol{\rho}$-oblivious distinguishers $A \in \mathcal{A}$, for all individually-consistent $\mathcal{M} : \mathcal{X} \to \mathbb{R}^d$,*

$$\left| \Pr_{(X,Y^*)\sim\mathcal{D}^*}\left[\, A(X, Y^*; \tilde{\rho}(X)) = 1 \,\right] - \Pr_{X\sim\mathcal{D}^*_X}\left[\, A(X, \mathcal{G}(\mathcal{M}(X; \tilde{\rho}(X))); \tilde{\rho}(X)) = 1 \,\right] \right| \leq \varepsilon.$$

We study Oblivious-OI in detail in Section 6. As an initial observation, we show that Oblivious-OI is the weakest notion yet, and is implied by existential OI (even the $\forall\exists$ variant).

Proposition 15 *If a statistic predictor $\tilde{\rho} : \mathcal{X} \to \mathbb{R}^d$ is $(\mathcal{A}, \varepsilon)$-$\forall\exists$-OI, then $\tilde{\rho}$ is $(\mathcal{A}, \varepsilon)$-Oblivious-OI.*

The proposition follows by the above observations. Assuming $\tilde{\rho}$ is $(\mathcal{A}, \varepsilon)$-$\forall\exists$-OI, then for each $A \in \mathcal{A}$, there is some individually-consistent model $\mathcal{M}_A$ such that the acceptance probability on $(X, \mathcal{G}(\mathcal{M}(X; \tilde{\rho}(X))))$ for $X \sim \mathcal{D}^*_X$ is within ε of the acceptance probabililty on $(X, Y^*) \sim \mathcal{D}^*$. For any $\boldsymbol{\rho}$-oblivious distinguisher A, the acceptance probability is the same for $\tilde{Y}$ generated from any individually-consistent model $\mathcal{M}$. Thus, the acceptance probability of all $\boldsymbol{\rho}$-oblivious $A \in \mathcal{A}$ under any individually-consistent $\mathcal{M}$ must be the same as that on $\mathcal{M}_A$, and thus ε-close to the probability on Nature.

5. Learning OI Generative Models

In this section, we discuss how to learn Generative OI Models. That is, given some class of distinguishers $\mathcal{A}$ and an acceptable distinguishing advantage ε, produce a generative outcome model $\mathcal{M} : \mathcal{X} \to \Delta(\mathcal{Y})$ that is indistinguishable from Nature's conditional distribution on outcomes $\mathcal{D}^*_{Y|X}$. In order to learn a model that encodes and outputs probability distributions over the outcome space, a few key questions need to be answered first. In particular, for a given outcome space $\mathcal{Y}$, how do we represent the distributions $F_Y \in \Delta(\mathcal{Y})$. Even ignoring measure-theoretic pathologies, there can be many ways to write down a given probability distribution. Different methods of specifying probability distributions are more or less efficient for different tasks, and will change the complexity of obtaining OI.

As such, we identify an abstraction through which we will interact with the learned probability distributions. We show how the abstract algorithmic methods we define suffice to learn a generative outcome model satisfying OI. In this way, given any outcome space $\mathcal{Y}$ for which there is a reasonable representation that supports these distributional operations efficiently will lead to efficient learning of OI models.

Representing Outcome Distributions. For a given outcome space $\mathcal{Y}$ with outcome distribution space $\Delta(\mathcal{Y})$, we denote the *Representation Space* as $\mathcal{R}$. For a distribution $F_Y \in \Delta(\mathcal{Y})$, we use $R(F_Y) \in \mathbb{R}$ to denote its representation. Throughout, we will assume that $\mathcal{R}$ is of fixed complexity; that is, for any $R \in \mathcal{R}$, there exists some fixed finite upper bound on the description length. For instance, $\mathcal{R}$ may be a subset of $\mathbb{R}^M$ for some fixed $M \in \mathbb{N}$. To be concrete, for discrete random variables over a finite domain $\mathcal{Y} = [N] = \{1, \ldots, N\}$, $\mathcal{R}$ may be an explicit approximation to the probability mass function, and for continuous random variables over a bounded range, $\mathcal{R}$ may be a circuit that on input $y \in \mathcal{Y}$, outputs the probability density of y. In generality, we may think of $\mathcal{R}$ as giving an efficient (discretized) approximation to the cumulative distribution function of the probability distributions in $\Delta(\mathcal{Y})$.

The importance of working with a fixed representation $\mathcal{R}$ is algorithmic. In particular, we will define the learning algorithm for Generative OI Models to operate on and return represenations $R \in \mathcal{R}$, corresponding to $F_Y \in \Delta(\mathcal{Y})$. Importantly, we assume that the representation $\mathcal{R}$ supports the following two operations.

- **Sample Generation.** A sample generator

  $$\mathcal{G} : \mathcal{R} \to \mathcal{Y}$$

  is a randomized map that given the representation $R(F_Y)$ of some distribution, $\mathcal{G}(R(F_Y))$ returns a random outcome $\tilde{Y} \sim F_Y$ drawn from the corresponding distribution.

- **Reweighting.** The reweighting procedure is a fixed algorithm (i.e., uniform computation) that computes a multiplicative weights update. In particular, the algorithm

  $$\mathcal{W}^B : \mathcal{R} \times [-1, 1] \to \mathcal{R}$$

  is an oracle-algorithm, that has access to an oracle to the characteristic function of some subset $B \subseteq \mathcal{Y}$, assumed to have non-zero (and non-unit) measure. Given the representation $R(F_Y)$ and some constant $\eta \in [-1, 1]$, the reweighting $\mathcal{W}^B(R(F_Y); \eta)$ returns the representation $R(G_Y)$ for some distribution $G_Y \in \Delta(\mathcal{Y})$ where

  $$\Pr_{G_Y}[\, Y \in B \,] \propto e^{\eta} \cdot \Pr_{F_Y}[\, Y \in B \,] \qquad\qquad \Pr_{G_Y}[\, Y \notin B \,] \propto \Pr_{F_Y}[\, Y \notin B \,].$$

With these two methods, we can describe the learning algorithm for Generative OI Models.

```
Generative OI Learning Algorithm.
Given. Family of distinguishers $\mathcal{A}$; advantage $\varepsilon > 0$; Prior $\mathcal{P}_Y \in \Delta(\mathcal{Y})$; step size $\eta > 0$
Output. Generative outcome model $\tilde{\mathcal{M}} : \mathcal{X} \to \mathcal{R}$
Initialization.
Initialize $\tilde{\mathcal{M}}_0 : \mathcal{X} \to \mathcal{R}$ to be the constant model that returns the representation of prior $R(\mathcal{P}_Y)$.
Iterate. for $t = 0, 1, \ldots, T$
For each $A \in \mathcal{A}$, evaluate the (signed) distinguishing advantage $\varepsilon_A$,

    $\varepsilon_A = \Pr_{\mathcal{D}^*}\left[\, A(X, Y^*; \tilde{\mathcal{M}}_t(X)) = 1 \,\right] - \Pr_{\mathcal{D}(\tilde{\mathcal{M}}_t)}\left[\, A(X, \tilde{Y}; \tilde{\mathcal{M}}_t(X)) = 1 \,\right]$

using random samples from $\mathcal{D}^*$ and random generated samples from $\mathcal{D}(\tilde{\mathcal{M}}_t)$

If $\max_{A \in \mathcal{A}} \; |\varepsilon_A| \le \varepsilon$, return $\tilde{\mathcal{M}}_t$.

Else, select any $A_t \in \{A \in \mathcal{A} : |\varepsilon_A| > \varepsilon\}$.
Let $B_t(\cdot) = \left\{ y \in \mathcal{Y} : A_t(\cdot, y, \tilde{\mathcal{M}}_t(\cdot)) = 1 \right\}$
Reweight the model $\tilde{\mathcal{M}}_{t+1}(\cdot) \leftarrow \mathcal{W}^{B_t(\cdot)}(\tilde{\mathcal{M}}_t(\cdot); \mathrm{sgn}(\varepsilon_{A_t}) \cdot \eta)$
```

The algorithm proceeds as follows. Initially, the model $\tilde{\mathcal{M}}$ is initialized to return the same prior distribution for all individuals $x \in \mathcal{X}$. Importantly, we don't maintain the distributions for each x

explicitly, but rather implicitly. In particular, we build up the logic of $\tilde{\mathcal{M}}$ into a circuit that takes an input $x \in \mathcal{X}$, and returns the representation of the outcome distribution $\tilde{\mathcal{M}}(x)$.

Then, the algorithm executes the following iteration until convergence. The first stage of the iteration searches for some $A \in \mathcal{A}$ that successfully distinguishes between Nature $\mathcal{D}^*$ and the current model of Nature $\mathcal{D}(\tilde{\mathcal{M}})$. If the algorithm cannot find such an A, then $\tilde{\mathcal{M}}$ ε-fools all of the distinguishers in $\mathcal{A}$, so it is $(\mathcal{A}, \varepsilon)$-Generative OI and we are done.

The second stage occurs when we do find a distinguisher A with significant advantage. Intuitively, such a distinguisher contains information that may be useful for pulling the distributions defined by $\tilde{\mathcal{M}}$ closer to the true conditional outcome distribution $\mathcal{D}^*_{Y|X}$.

As such, we devise and update based on a successful $A \in \mathcal{A}$. Specifically, the update to the circuit for $\tilde{\mathcal{M}}$ must occur implicity, such that in a constant overhead of logic, the distributions defined by $\tilde{\mathcal{M}}(x)$ are simultaneously updated for all $x \in \mathcal{X}$. To begin, consider a fixed x. We want to run the multiplicative weights update to $\tilde{\mathcal{M}}(x)$, such that we increase/decrease the density on the set of outcomes defined as follows.

$$B[x] = \left\{ y \in \mathcal{Y} : A(x, y; \tilde{\mathcal{M}}(x)) = 1 \right\}$$

For a fixed representation $\tilde{\mathcal{M}}(x) \in \mathcal{R}$, we can use the Reweighting algorithm $\mathcal{W}^{B[x]}$ to return the representation of the distribution, where the probability of $y \in B[x]$ has been reweighted by e^{η}. Importantly, we assume that the logic of $\mathcal{W}^B$ can be implemented by a uniform oracle procedure, that runs the same code independent of the actual predicate B.

This uniformity assumption is crucial: it allows us to update the distribution for all $x \in \mathcal{X}$ simultaneously by taking x as an input to the oracle. That is, rather than defining a pre-specified $B[x]$ for each x, we instead define a parameterized oracle $B(\cdot) = A(\cdot, y; \tilde{\mathcal{M}}(\cdot))$ that takes $x \in \mathcal{X}$ as input. Then, when we wish to evaluate $\tilde{\mathcal{M}}(\cdot)$ on a specific input $x \in \mathcal{X}$, we feed the input x into $B_t(\cdot)$ corresponding to the update for each iteration t.

Returned predictor. The final predictor $\tilde{\mathcal{M}}$ will be $(\mathcal{A}, \varepsilon)$-generative OI by the fact the termination condition. The "hypothesis class" of predictors that the algorithm can output is built up from individual $A \in \mathcal{A}$ as well as from the logic of $\mathcal{W}$. In particular, the evaluation of $\tilde{\mathcal{M}}(x)$ follows a sequence of calls to the Reweighting algorithm, using a subpopulation oracle determined by A_t for each iteration t. In particular, if $\tilde{\mathcal{M}}$ is returned after T iterations, we can express $\tilde{\mathcal{M}}(\cdot)$ as

$$\mathcal{W}^{B_T(\cdot)} \left(\mathcal{W}^{B_{T-1}(\cdot)} \left(\ldots \mathcal{W}^{B_1(\cdot)}(R(\mathcal{P}_Y); \eta_1) \ldots ; \eta_{T-1} \right) ; \eta_T \right)$$

where:

- $B_t(\cdot) = \left\{ y \in \mathcal{Y} : A_t(\cdot, y, \tilde{\mathcal{M}}_t(\cdot)) = 1 \right\}$; on a fixed input $x \in \mathcal{X}$, then $B_t(x)$ is the set of outcomes that make the distinguisher $A(x, y, \tilde{\mathcal{M}}_t(x)) = 1$ given the prediction $\tilde{\mathcal{M}}_t(x)$ in the tth iteration.
- $\tilde{\mathcal{M}}_0(x) = R(\mathcal{P}_Y)$ for all $x \in \mathcal{X}$ is the representation of the prior distribution on outcomes
- $\eta_t \in \{-\eta, \eta\}$ indicates whether the tth update should be positive or negative.

To begin the analysis, we discuss the complexity of evaluating the learned model $\tilde{\mathcal{M}}$ returned by the algorithm. A key assumption that we will make is that the representation returned by $\tilde{\mathcal{M}}_t(x) \in$

$\Delta(\mathcal{Y})$ is of a fixed size; in some models, this complexity could scale as the complexity of computing $\tilde{\mathcal{M}}_t$ increased.

With this assumption in place, we can decompose the cost into costs per iterate. Evaluating the predictor requires evaluating $\mathcal{W}^{B_t}$ for each $t \in [T]$. This complexity scales with the cost of Reweighting and the cost of evaluating $A \in \mathcal{A}$, which are required to evaluate the oracle B_t.

Proposition 16 (Time Complexity of Evaluation) *When the algorithm returns $\tilde{\mathcal{M}}$ after T iterations, the complexity of evaluating $\tilde{\mathcal{M}}$ requires T executions of Reweighting, which may make calls to $A \in \mathcal{A}$ as an oracle.*

Note that in the worst case, B_t may make a call to A_t for each $y \in \mathcal{Y}$. For instance, for discrete distributions, the time complexity can be upper bounded by $O(T \cdot |\mathcal{Y}| \cdot time_{\mathcal{A}})$ plus the time to compute the reweighting. As such, a key quantity in understanding the complexity of the predictive model is the iteration complexity.

Iteration Complexity. The first step in bounding the complexity of the learning algorithm is to argue that the algorithm converges to an $(\mathcal{A}, \varepsilon)$-OI model in a bounded number of iterations T. This argument follows by identifying a potential function ϕ_{KL} based on the expected KL-divergence between $\tilde{\mathcal{M}}(X)$ and $\mathcal{D}^*_{Y|X}$, and arguing that any successful distinguisher $A \in \mathcal{A}$ suggests an update that causes the potential to drop significantly.

We measure progress in expected KL Divergence. For concreteness, we focus on the case of discrete probability distributions, but a similar analysis follows for the continuous case. For two probability mass functions f_Y, g_Y, the KL-divergence is defined as follows.

$$D_{\mathrm{KL}}(f_Y; g_Y) = \sum_{y \in \mathcal{Y}} f_Y(y) \cdot \log\left(\frac{f_Y(y)}{g_Y(y)}\right)$$

We define the potential function as the KL-divergence from $\mathcal{D}^*_{Y|X}$.

$$\phi_{\mathrm{KL}}(\tilde{\mathcal{M}}) = \mathop{\mathbf{E}}_{X \sim \mathcal{D}^*_X}\left[\ D_{\mathrm{KL}}\left(\mathcal{D}^*_{Y|X}; \tilde{\mathcal{M}}(X)\right)\ \right]$$

We argue that if there is a distinguisher $A \in \mathcal{A}$ that has nontrivial advantage, the multiplicative update under Reweighting causes the potential function to drop.

For notational convenience, we use $f^* : \mathcal{X} \times \mathcal{Y} \to [0,1]$ to denote the conditional probability mass function of Nature $\mathcal{D}^*_{Y|X}$ and use $\tilde{m} : \mathcal{X} \times \mathcal{Y} \to [0,1]$ to denote the probabiltiy mass function of $\tilde{\mathcal{M}}$, for a given individual $x \in \mathcal{X}$. Suppose there exists $A \in \mathcal{A}$ s.t. $|\varepsilon_A| > \varepsilon$, for

$$\begin{aligned}
\varepsilon_A &= \mathop{\mathbf{Pr}}_{\mathcal{D}^*}\left[\ A(X, Y^*; \tilde{\mathcal{M}}(X)) = 1\ \right] - \mathop{\mathbf{Pr}}_{\mathcal{D}(\tilde{\mathcal{M}})}\left[\ A(X, \tilde{Y}; \tilde{\mathcal{M}}(X)) = 1\ \right] \\
&= \mathop{\mathbf{E}}_{\mathcal{D}^*}\left[\ A(X, Y^*; \tilde{\mathcal{M}}(X))\ \right] - \mathop{\mathbf{E}}_{\mathcal{D}(\tilde{\mathcal{M}})}\left[\ A(X, \tilde{Y}; \tilde{\mathcal{M}}(X))\ \right] \\
&= \mathop{\mathbf{E}}_{\mathcal{D}^*_X}\left[\ \sum_{y \in \mathcal{Y}} A(X, y; \tilde{\mathcal{M}}(X)) \cdot (f^*(X; y) - \tilde{m}(X; y))\ \right]
\end{aligned}$$

Consider the update under Reweighting. Let $\tilde{\mathcal{M}}_0$ refer to the initial model, and $\tilde{\mathcal{M}}_1$ the updated. Let $Z(x)$ denote the normalization constant on individual $x \in \mathcal{X}$.

$$Z(x) = \sum_{y \in \mathcal{Y}} \tilde{m}_0(x;y) \cdot e^{\operatorname{sgn}(\varepsilon_A) \cdot \eta \cdot A(x,y,\tilde{\mathcal{M}}_0(x))}$$

By a standard argument, we can bound the difference in potential as follows.

$$\phi_{\mathrm{KL}}(\tilde{\mathcal{M}}_0) - \phi_{\mathrm{KL}}(\tilde{\mathcal{M}}_1) = \mathop{\mathbf{E}}_{\mathcal{D}_X^*}\left[\operatorname{sgn}(\varepsilon_A) \cdot \eta \sum_{y \in \mathcal{Y}} A(X,y;\tilde{\mathcal{M}}_0(X)) \cdot f^*(X;y) - \log\left(Z(X)\right) \right]$$

By the fact that $\eta \in [-1,1]$ and the fact that $\log(1-x) \le -x$, we derive the following in equality for the log partition function.

$$\begin{aligned}
\log(Z(X)) &= \log\left(\sum_{y \in \mathcal{Y}} \tilde{m}_0(X;y) \cdot e^{\operatorname{sgn}(\varepsilon_A)\cdot\eta\cdot A(X,y,\tilde{\mathcal{M}}_0(X))} \right) \\
&\le \log\left(\sum_{y \in \mathcal{Y}} \tilde{m}_0(X;y) \cdot \Big(1 - \operatorname{sgn}(\varepsilon_A) \cdot \eta \cdot A(X,y,\tilde{\mathcal{M}}_0(X)) + \eta^2 \cdot A(X,y,\tilde{\mathcal{M}}_0(X))\Big) \right) \\
&\le \log\left(1 - \operatorname{sgn}(\varepsilon_A) \cdot \eta \cdot \sum_{y \in \mathcal{Y}} \tilde{m}_0(X;y) \cdot A(X,y,\tilde{\mathcal{M}}_0(X)) + \eta^2 \right) \\
&\le \operatorname{sgn}(\varepsilon_A) \cdot \eta \cdot \sum_{y \in \mathcal{Y}} \tilde{m}_0(X;y) \cdot A(X,y,\tilde{\mathcal{M}}_0(X)) + \eta^2
\end{aligned}$$

Thus, we can bound the difference as follows.

$$\begin{aligned}
&\ge \mathop{\mathbf{E}}_{\mathcal{D}_X^*}\left[\operatorname{sgn}(\varepsilon_A) \cdot \eta \sum_{y \in \mathcal{Y}} A(X,y;\tilde{\mathcal{M}}_0(X)) \cdot (f^*(X;y) - \tilde{m}(X;y)) - \eta^2 \right] \\
&= \operatorname{sgn}(\varepsilon_A) \cdot \eta \left(\mathop{\mathbf{Pr}}_{\mathcal{D}^*}\left[A(X,Y^*;\tilde{\mathcal{M}}(X)) = 1 \right] - \mathop{\mathbf{Pr}}_{\mathcal{D}(\tilde{\mathcal{M}})}\left[A(X,\tilde{Y};\tilde{\mathcal{M}}(X)) = 1 \right] \right) - \eta^2 \\
&= |\varepsilon_A| \cdot \eta - \eta^2
\end{aligned}$$

Thus, provided we take $\eta \le \varepsilon/2$, then the potential function is guaranteed to drop by at least $\varepsilon^2/4$ in each update.

Proposition 17 (Iteration complexity) *Taking $\eta = \varepsilon/2$, the algorithm is guaranteed to return a $(\mathcal{A}, \varepsilon)$-Generative OI Model in T iterations for*

$$T \le O\left(\frac{\phi_{\mathrm{KL}}(\mathcal{P}_Y)}{\varepsilon^2} \right)$$

where $\phi_{\mathrm{KL}}(\mathcal{P}_Y)$ is the KL-divergence between $\mathcal{D}^*_{Y|X}$ and $\mathcal{P}_Y$, averaged over $X \sim \mathcal{D}^*_X$.

That is, the worst-case number of iterations is upper bounded in terms of the quality of the prior. For discrete probability distributions of finite support, this quantity can be bounded finitely; for distributions of unbounded support, then to achieve convergence in the worst-case, we need to assume a prior that is bounded in KL for all individuals' outcome distributions.

Sample Complexity. Using the bound on the iteration complexity, we can derive an upper bound on the sample complexity. In particular, we can bound the sample complexity generically by obtaining a per-iteration bound and then resampling to estimate the acceptance probabilities each iteration.

Note that the algorithm's only interaction with Nature's samples is in determining the acceptance probability of each $A \in \mathcal{A}$ for the current $\tilde{\mathcal{M}}$. Given that we care about ε-indistinguishability, it suffices to obtain estimates that are ε/c-accurate for some constant c for all $A \in \mathcal{A}$. Thus, it suffices to take enough samples from $\mathcal{D}^*$ to guarantee uniform convergence over $\mathcal{A}$. By standard arguments, we can bound the sample complexity by $m_0 = O((\log|\mathcal{A}| + \log(T))/\varepsilon^2)$.

Proposition 18 (Sample complexity) *With success probability* $1-\delta$, *the algorithm can be implemented using* m *samples, for*

$$m \le T \cdot O\left(\frac{\log(T\,|A|\,/\delta)}{\varepsilon^2}\right) \le \tilde{O}\left(\phi_{\mathrm{KL}}(\mathcal{P}_Y) \cdot \frac{\log(|A|\,/\delta)}{\varepsilon^4}\right).$$

Time Complexity of Learning. To bound the time complexity of running the learning algorithm, we decompose each iteration into three tasks: (1) Sample Generation, (2) Searching, and (3) Reweighting. We first state the bounds generically, then analyze it concretely for the case for discrete outcome distributions.

To bound the cost of sampling, we note that the sample complexity analysis of Nature's samples equally bounds the number of samples we need from the modeled distribution $\mathcal{D}(\tilde{\mathcal{M}})$. In particular, over the course of the algorithm, we need to generate $m \le \tilde{O}\left(\phi_{\mathrm{KL}}(\mathcal{P}_Y) \cdot \frac{\log(|A|/\delta)}{\varepsilon^4}\right)$ modeled samples. This involves drawing m unlabeled samples from $X \sim \mathcal{D}^*_X$, then as needed, calling $\mathcal{G}(\tilde{\mathcal{M}}(X))$ to obtain a modeled label for each sample. Thus, the sample generation cost of the algorithm is proportional to the sample complexity times the time complexity of generating a sample, which we denote $time_{\mathcal{G}}$.

The next key cost is searching in each iteration for some $A \in \mathcal{A}$ that distinguishes between $\mathcal{D}^*$ and $\mathcal{D}(\tilde{\mathcal{M}})$ for the current model $\tilde{\mathcal{M}}$. Here, we perform an exhaustive search by iterating through the $|\mathcal{A}|$ distinguishers, and evaluating ε_A for each $A \in \mathcal{A}$. This complexity will be dominated by iterating through $\mathcal{A}$, with $O(m_0)$ evaluations of the distinguisher A per iterate to estimate ε_A. We denote an upper bound on the time required to evaluate $A \in \mathcal{A}$ as $time_{\mathcal{A}}$.

Finally, at the end of the iteration, the algorithm incorporates the reweighting algorithm into the generative outcome model based on the identified distinguisher. This process implicitly updates $\tilde{\mathcal{M}}(\cdot)$ to have a new normalized probability function, given any input $x \in \mathcal{X}$. In order to feed representations of the predicted outcome distributions into the distinguishers at the next level, we need to evaluate the Reweighting process $\mathcal{W}^{B(\cdot)}$ in each iteration, which we assume uses $time_{\mathcal{W}}$.

Proposition 19 (Time Complexity of Learning) *The time complexity to learn* $(\mathcal{A}, \varepsilon)$*-generative OI models can be upper bounded by* $T \le O\left(\phi_{\mathrm{KL}}(\mathcal{P}_Y)/\varepsilon^2\right)$ *iterations where each iteration makes*

- $O(m_0)$ *calls to Sample Generation* $\mathcal{G}$ *from* $\tilde{\mathcal{M}}_t$
- $O(m_0)$ *evaluations of each distinguisher* $A \in \mathcal{A}$
- *One call to Reweighting* $\mathcal{W}^{B_t(\cdot)}$

totally time bounded by

$$time \le O\left(m \cdot time_{\mathcal{G}} + m \cdot |\mathcal{A}| \cdot time_{\mathcal{A}} + T \cdot time_{\mathcal{W}}\right).$$

Implementing Sample Generation and Reweighting. In all, to learn Generative OI Models for any Nature, it suffices to exhibit a concrete representation of outcome distributions from $\Delta(\mathcal{Y})$ that supports efficient Sample Generation and Reweighting. For concreteness, we consider how to instantiate the abstraction for outcomes that come from some large, but finite, discrete domain $\mathcal{Y} = [N] = \{1, \ldots, N\}$. In this case, we can represent the distribution explicitly as a discrete CDF, where $\tilde{\mathcal{M}} : \mathcal{X} \times \mathcal{Y} \to [0, 1]$ is used to approximate the true conditional probability law,[5] where for each $y \in [N]$,

$$\tilde{\mathcal{M}}(x; y) \approx \Pr_{\mathcal{D}^*}[\, Y \leq y \mid X = x \,] \,.$$

With this representation fixed, we can analyze the complexity necessary to implement Sample Generation and Reweighting for a given predicted distribution $\tilde{\mathcal{M}}(x) \in \Delta(\mathcal{Y})$.

- **Sample Generation:** Draw a uniform random $\tau \in [0, 1]$, and binary search through $\tilde{\mathcal{M}}(x)$ for the maximum $y \in [N]$, such that $\tilde{\mathcal{M}}(x; y) \leq \tau$. Then, return y with probability

$$\frac{\tau - \tilde{\mathcal{M}}(x; y)}{\tilde{\mathcal{M}}(x; y+1) - \tilde{\mathcal{M}}(x; y)}$$

  and $y + 1$ otherwise.

- **Reweighting:** Compute the probability mass function $\tilde{m} : \mathcal{X} \times \mathcal{Y}$ corresponding to $\tilde{\mathcal{M}}$, by differencing. Then, multiply $\tilde{m}(\cdot; y)$ by e^{η} for all $y \in B$, compute the normalization factor by summing over all $y \in \mathcal{Y}$, and divide each entry by the normalization.

As such, using the CDF representation, Sample Generation is possible using $\text{polylog}(N)$ time (in a RAM model, with two calls to uniform randomness), and Reweighting is possible using $\tilde{O}(N)$ arithmetic operations. The main computational cost of sample generation, then, involves evaluating $\tilde{\mathcal{M}}(x) \in \Delta(\mathcal{Y})$ to obtain the CDF representation. By the proposition above, this procedure requires at most T calls to Reweighting. In the worst case, each of these calls could make $|\mathcal{Y}|$ calls to some $A \in \mathcal{A}$ to implement the oracle B. Thus, we can bound $time_{\mathcal{G}}$ as follows.

$$time_{\mathcal{G}} \leq T \cdot time_{\mathcal{A}} \cdot O(\tilde{N})$$

To bound the complexity, we use the fact that the KL-divergence for a discrete distribution is upper bounded by $\log |\mathcal{Y}|$, so the iteration complexity is bounded by $T \leq O(\log |\mathcal{Y}| / \varepsilon^2)$ and the sample complexity $m \leq \frac{\log|\mathcal{Y}| \cdot \log|\mathcal{A}|}{\varepsilon^4}$. In all, we can bound the expression for $time$ as

$$time \leq m \cdot T \cdot time_{\mathcal{A}} \cdot \tilde{O}(|\mathcal{Y}|) + m \cdot |\mathcal{A}| \cdot time_{\mathcal{A}} + T \cdot \tilde{O}(|\mathcal{Y}|)$$

Subsituting in the bounds, we obtain the following guarantee, where the $\tilde{O}$ notation suppresses polylogarithmic factors in the arguments.

[5] Throughout, we assume that the marginal distribution of individuals $\mathcal{D}^*_X$ is supported on a discrete domain $\mathcal{X}$.

Theorem 20 *There exists an algorithm for learning an OI Generative Outcome Model over a discrete outcome domain $\mathcal{Y}$. For any given distribution $\mathcal{D}^*$ over $\mathcal{X} \times \mathcal{Y}$, class $\mathcal{A}$ of distinguishers and indistinguishability parameter ε, the sample complexity is bounded by:*

$$\tilde{O}\left(\frac{\log|\mathcal{Y}| \cdot \log|\mathcal{A}|}{\varepsilon^4}\right).$$

Assuming that the distinguishers in $\mathcal{A}$ can be evaluated in time $time_{\mathcal{A}}$, the running time is:

$$time \leq \tilde{O}\left(\frac{|\mathcal{Y}| \cdot time_{\mathcal{A}} \cdot \log|\mathcal{A}|}{\varepsilon^6} + \frac{|\mathcal{A}| \cdot time_{\mathcal{A}} \cdot \log|\mathcal{Y}|}{\varepsilon^4}\right)$$

The algorithm produces a generative model $\tilde{\mathcal{M}} : \mathcal{X} \to \mathcal{Y}$. Given an input $x \in \mathcal{X}$, the model $\tilde{\mathcal{M}}$ can be evaluated (i.e. used to generate a sample) in time:

$$\tilde{O}\left(\frac{|\mathcal{Y}| \cdot time_{\mathcal{A}} \cdot \log|\mathcal{A}|}{\varepsilon^2}\right).$$

Improved Efficiency for No-Access Distinguishers. Finally, we remark that the complexities we derive here can vary quantitatively depending on the qualitative aspects of the distinguisher model. For instance, in the original work Dwork et al. (2021) define a variant of OI, which they call "No-Access," where distinguishers do not look at the value of the predictions $\tilde{\mathcal{M}}(X)$. If we restrict our attention to No-Access Distinguishers for generative outcome models, then the complexities can be bounded more tightly. In particular, the sample complexity and evaluation time complexity avoid costs associated with processing the predictions $\tilde{\mathcal{M}}(X)$. Specifically, for the sample complexity, we only need to obtain valid estimates of the acceptance probability of the distinguishers once (as there is no dependence on $\tilde{\mathcal{M}}$), so we can avoid the factor T blow-up and only take $m = O(\log|\mathcal{A}|/\varepsilon^2)$ labeled samples from $\mathcal{D}^*$. For the evaluation time complexity, we need not Reweight $\tilde{\mathcal{M}}$ at every step, and can instead process all of the updates, then Reweight; again, this saves a factor T on the Reweighting cost.

6. Multicalibrated Statistic Fool Oblivious Distinguishers

Here, we explore the expressiveness of Oblivious Outcome Indistinguishability. Intuitively, oblivious distinguishers focus their attention on the statistics being modeled, and nothing else. In this sense, to capture the constraints of Oblivious OI, it may be sufficient to reason directly about the statistics of interest, rather than a generative model for outcomes. Here, as in the original work on Bernoulli OI, we show a tight connection between Oblivious OI and Multicalibration. Multicalibration, defined by Hébert-Johnson et al. (2018) in the context of fair binary prediction, requires that predictions be well-calibrated not simply on the population as a whole, but even when we restrict our attention to structured subpopulations.

To begin, we generalize the idea of multicalibration to predictions of arbitrary statistics. With the appropriate generalization in hand, we show an equivalence between multicalibrated statistics and oblivious outcome indistinguishable statistics. Finally, we conclude with a discussion of the work of Jung et al. (2021) on moment multicalibration and how it relates to multicalibrated statistics and OI. We show how the connection between oblivious OI and multicalibration, and specifically the technical requirement of linearization, sheds light on approach of mean-conditioning used by Jung et al. (2021) to achieve moment multicalibration.

6.1. Multicalibrated Statistics

Before we can define multicalibration in the context of general statistics, we must first fix a definition of calibration for general statistics. We begin with a technical notion of calibration, that requires the predicted statistics be accurate in expectation, even after conditioning on the predictions themselves.

Definition 21 (Calibration, *accuracy in expectation*) *A predictor $\tilde{\rho} : \mathcal{X} \to \Delta(\mathcal{Y})$ for statistic $\boldsymbol{\rho}$ is calibrated over Nature $\mathcal{D}^*$ if for all $\nu \in \operatorname{supp}(\tilde{\rho})$,*

$$\mathbf{E}\left[\, \boldsymbol{\rho}\left\{\mathcal{D}^*_{Y|X}\right\} \,\middle|\, \tilde{\rho}(X) = \nu \,\right] = \nu.$$

This notion of calibration generalizes the Bernoulli formulation, where the constraint is required to hold for the predicted probabilities, $\mathbf{Pr}\left[\, Y = 1 \mid \tilde{p}(X) = v \,\right] = v$. Calibration via accuracy in expectation is a natural desideratum for learning because the true outcome statistics satisfy the constraints. That is, if we take $\rho^*(X) = \boldsymbol{\rho}\left\{\mathcal{D}^*_{Y|X}\right\}$, then

$$\mathbf{E}\left[\, \boldsymbol{\rho}\left\{\mathcal{D}^*_{Y|X}\right\} \,\middle|\, \rho^*(X) = \nu \,\right] = \mathbf{E}\left[\, \nu \mid \rho^*(X) = \nu \,\right] = \nu.$$

As in the Bernoulli case, calibration on its own is a very weak condition for recovery, and does not require that a predictor be informative. For instance, the constant predictor $\tilde{\rho}(X) = \bar{\nu} = \mathbf{E}\left[\, \boldsymbol{\rho}\left\{\mathcal{D}^*_{Y|X}\right\} \,\right]$ that predicts the expected statistic on the marginal distribution of outcomes is calibrated.

$$\mathbf{E}\left[\, \boldsymbol{\rho}\left\{\mathcal{D}^*_{Y|X}\right\} \,\middle|\, \tilde{\rho}(X) = \bar{\nu} \,\right] = \mathbf{E}\left[\, \boldsymbol{\rho}\left\{\mathcal{D}^*_{Y|X}\right\} \,\right] = \bar{\nu}.$$

Of course, this statistic predictor is constant over $\mathcal{X}$, and thus, gives no information about $\mathcal{D}^*_{Y|X}$ beyond that conveyed by $\mathcal{D}^*_Y$. This observation motivates the idea of enforcing multicalibration, a strengthening of calibration, that requires calibration over a rich collection of structured subpopulations. As in Definition 4, we formulate the idea of "conditioning" on members of a subpopulation $S \subseteq \mathcal{X}$ that receive a certain prediction $\tilde{\rho}(X) = \nu$, in terms of a class of functions $\mathcal{C} \subseteq \left\{\mathcal{X} \times \mathbb{R}^d \to \mathcal{B}^d_1\right\}$. That is, we consider vector-valued functions $c \in \mathcal{C}$ that map individual-prediction pairs to vectors of unit ℓ_1-norm; this choice is somewhat arbitrary, but plays nicely with the assumption that the predicted statistics $\tilde{\rho} : \mathcal{X} \to \mathcal{B}^d_\infty$ are ℓ_∞-bounded.

Definition 22 (Statistic Multicalibration) *Fix a class of functions $\mathcal{C} \subseteq \left\{\mathcal{X} \times \mathbb{R}^d \to \mathcal{B}^d_1\right\}$ and an approximation $\alpha \geq 0$. For a bounded d-dimensional statistic $\boldsymbol{\rho} : \Delta(\mathcal{Y}) \to \mathcal{B}^d_\infty$, a predictor $\tilde{\rho} : \mathcal{X} \to \mathbb{R}^d$ is $(\mathcal{C}, \alpha)$-multicalibrated if for all $c \in \mathcal{C}$*

$$\mathbf{E}\left[\, \left\langle c(X, \tilde{\rho}(X)), \left(\boldsymbol{\rho}\left\{\mathcal{D}^*_{Y|X}\right\} - \tilde{\rho}(X)\right)\right\rangle \,\right] \leq \alpha.$$

Definition 22 is yet a further generalization of the Bernoulli variants of multiaccuracy and multicalibration. Note that, as with calibration, Nature's true statistics ρ^* are feasible for multicalibration, for any collection $\mathcal{C}$. Thus, for any Nature $\mathcal{D}^*$, there exists some multicalibrated statistic predictor $\tilde{\rho} = \rho^*$. The strength of the multicalibration guarantee is parameterized by the complexity of $\mathcal{C}$: as the functions $c \in \mathcal{C}$ become more complex, the multicalibration constraints enforce a tighter consistency with Nature.

Linearizing statistics. Despite the fact that multicalibration is definitionally feasible by taking $\tilde{\rho} = \rho^*$, in general, achieving multicalibration from a small set of samples may be practically challenging or impossible. In particular, for arbitrary statistics, it is not even clear how to estimate whether a given $\tilde{\rho}$ satisfies calibration. Without repeated samples from $\mathcal{D}^*_{Y|X=x}$ for various fixed choices of $x \in \mathcal{X}$, it may be impossible to evaluate the expectation of $\boldsymbol{\rho}\left\{\mathcal{D}_{Y|X}\right\}$. Even though the information-theoretic optimal ρ^* may exist, in general, it will be impossible to obtain a close estimate of ρ^* without strong assumptions on $\mathcal{D}^*$.

This shortcoming of the accuracy-in-expectation definition of calibration motivates a different definition, where rather than averaging over the true statistic values on individuals, we average over the mixture outcome distributions, induced by averaging over individuals.

Definition 23 (Calibration, *consistency over mixtures*) *A predictor $\tilde{\rho} : \mathcal{X} \to \mathbb{R}^d$ for statistic $\boldsymbol{\rho}$ is calibrated over Nature $\mathcal{D}^*$ if for all $\nu \in \operatorname{supp}(\tilde{\rho})$*

$$\boldsymbol{\rho}\left\{\mathcal{D}_{Y|\tilde{\rho}(X)=\nu}\right\} = \nu.$$

Indeed, one practical appeal of the mixture definition of calibration is that it allows us to use "repeated" outcomes from within a subpopulation (e.g., the individuals that receive prediction $\tilde{\rho}(X) = \nu$) to reason about whether the predicted statistics are accurate in expectation. This framing of calibration suggests the following interpretation of the constraint: over the mixture of individuals that receive predicted statistic $\tilde{\rho}(x) = \nu$, the true statistic is actually ν.

For calibration of Bernoulli predictions, the two framings of calibration coincide. When moving to calibration of more general statistics, however, the mixture definition encounters issues. To begin, the mixture constraints are not generally satisfied—even by Nature's statistics—as discussed by Jung et al. (2021). For instance, the variance of outcomes amongst individuals who have true variance σ^2 is not necessarily σ^2. To see this, consider a collection of individuals who all have deterministic outcomes (i.e., $\sigma^2 = 0$), but half have $Y = 0$ and half have $Y = 1$. The variance of the outcome Y in the uniform mixture over individuals will be $1/4$, not 0. In this sense, without further restrictions, the semantics of calibration do not extend to arbitrary statistics.

To deal with these issues of feasibility, we restrict our attention to statistics where we can use the mixture and accuracy-in-expectation defintions interchangeably. In such cases, it will be possible to derive algorithms that work from a small sample of training data and achieve Definition 22. The key property that we exploit is linearization.

Definition 24 (Linearization) *A statistic $\boldsymbol{\rho}$ linearizes if for any distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$,*

$$\boldsymbol{\rho}\left\{\mathcal{D}_Y\right\} = \mathop{\mathbf{E}}_{\mathcal{D}_X}\left[\ \boldsymbol{\rho}\left\{\mathcal{D}_{Y|X}\right\}\ \right].$$

As defined, linearization immediately resolves the discrepancy in the two notions of calibration: for any linearizing statistic $\boldsymbol{\rho}$, the statistic on the mixture over X must equal the expectation of the individual statistics. A useful technical observation is that linearization is equivalent to requiring that $\boldsymbol{\rho}$ is a linear function of the probability density (mass) function, with coefficients that may depend on $y \in \mathcal{Y}$; that is, $\boldsymbol{\rho}$ is an expectation.

Lemma 25 *A statistic $\boldsymbol{\rho}$ linearizes if and only if $\boldsymbol{\rho}$ is an expectation of some function $r : \mathcal{Y} \to \mathbb{R}^d$ over outcomes,*

$$\boldsymbol{\rho}\left\{F_Y\right\} = \mathop{\mathbf{E}}_{F_Y}\left[\ r(Y)\ \right].$$

Proof For any input-outcome distribution $\mathcal{D}$, the marginal distribution on outcomes $\mathcal{D}_Y$ can be viewed as a convex combination of outcome distributions $\mathcal{D}_{Y|X}$, where the combination weights over X are determined by $\mathcal{D}_X$. By linearization, we know that the statistic of this convex combination $\boldsymbol{\rho}\left\{\mathcal{D}_Y\right\}$ is equal to the convex combination of the individual statistics $\mathbf{E}_{\mathcal{D}_X}\left[\ \boldsymbol{\rho}\left\{\mathcal{D}_{Y|X}\right\}\ \right]$. This property holds if and only if $\boldsymbol{\rho}$ is a linear functional of its input probability distribution, where the coefficients may depend on the outcome value y. Thus, there is some coefficient function $r : \mathcal{Y} \to \mathbb{R}$, such that for any $F_Y \in \Delta(\mathcal{Y})$, we can write the statistic

$$\boldsymbol{\rho}\{F_Y\} = \int r(y) \cdot f_Y(y) dy = \mathop{\mathbf{E}}_{F_Y}\left[\ r(Y)\ \right]$$

as the expectation of $r(Y)$ over $Y \sim F_Y$. ■

6.2. Equivalence of Multicalibration and Oblivious OI

We show a tight connection between multicalibrated statistics and statistics that satisfy Oblivious OI. In particular, we show an equivalence between the two notions for linearizing statistics.

Restatement of Theorem 1 *Suppose $\boldsymbol{\rho}$ linearizes. For any class of functions $\mathcal{C}$, there exists a family of $\boldsymbol{\rho}$-oblivious distinguishers $\mathcal{A}$ such that:*

- *For each $c \in \mathcal{C}$, there exists some $A^c \in \mathcal{A}$ that makes a single oracle call to c, and*
- *$(\mathcal{A}, \varepsilon)$-Oblivious OI implies $(\mathcal{C}, \alpha)$-multicalibration for $\alpha \leq 2\varepsilon$.*

For any family of $\boldsymbol{\rho}$-oblivious distinguishers $\mathcal{A}$, there exists a class of functions $\mathcal{C}$, such that $(\mathcal{C}, \alpha)$-multicalibration implies $(\mathcal{A}, \varepsilon)$-OI for $\varepsilon \leq \alpha$.

In the remainder of the section, we prove the theorem, explaining the nuances in the reductions as we go.

6.2.1. OBLIVIOUS OI CAPTURES MULTICALIBRATION

In the first direction, we show that for any class of functions $\mathcal{C}$, there is an efficient reduction to a family of oblivious distinguishers $\mathcal{A}$, such that $\mathcal{A}$-OI implies $\mathcal{C}$-multicalibration. For each $c \in \mathcal{C}$, there is an associated distinguisher $A^c \in \mathcal{A}$, such that evaluating A^c requires a single call to c and a constant amount of additional logic.

Proposition 26 (Oblivious OI captures Multicalibration) *Suppose that $\boldsymbol{\rho} : \Delta(\mathcal{Y}) \to \mathcal{B}^d_\infty$ is a bounded linearizing statistic. Then, for any class of functions $\mathcal{C} \subseteq \left\{\mathcal{X} \times \mathbb{R}^d \to \mathcal{B}^d_1\right\}$, there is an efficient black-box reduction to a family of distinguishers $\mathcal{A} \subseteq \left\{\mathcal{X} \times \mathcal{Y} \times \mathbb{R}^d \to \{0,1\}\right\}$ such that:*

- *For each $c \in \mathcal{C}$, there is some $A^c \in \mathcal{A}$ that makes a single call to c.*
- *Every $A^c \in \mathcal{A}$ is $(\boldsymbol{\rho}, 1/2)$-Lipschitz, and thus, $\boldsymbol{\rho}$-oblivious.*
- *For any Nature $\mathcal{D}^*$, if a predictor $\tilde{\rho} : \mathcal{X} \to \mathbb{R}^d$ is $(\mathcal{A}, \varepsilon)$-OI from Nature, then $\tilde{\rho}$ is $(\mathcal{C}, \alpha)$-multicalibrated on $\mathcal{D}^*$ for $\alpha \leq 2\varepsilon$.*

Proof Consider $\boldsymbol{\rho}$. By linearization and Lemma 25, we know that there exists some $r : \mathcal{Y} \to \mathcal{B}^d_\infty$ such that for any $F_Y \in \Delta(\mathcal{Y})$, the statistic is the expectation of $r(Y)$,

$$\boldsymbol{\rho}\left\{F_Y\right\} = \mathop{\mathbf{E}}_{F_Y}\left[\, r(Y) \,\right] \in \mathcal{B}^d_\infty$$

Starting with a class of functions $\mathcal{C}$, we show how to construct a family of distinguishers that accept with probability proportional to the expectations in the multicalibration constraints. Specifically, for each $c \in \mathcal{C}$, we build a randomized oblivious distinguisher A^c as follows.

$$A^c(x, y; \nu) = \begin{cases} 1 & \text{w.p. } \dfrac{\langle c(x,\nu), r(y)\rangle + 1}{2} \\ 0 & \text{o.w.} \end{cases}$$

First, we note that the stated acceptance rate for a fixed $y \in \mathcal{Y}$ is a valid probability, because by Hölder's inequality for $c(x,\nu) \in \mathcal{B}^d_1$ and $r(y) \in \mathcal{B}^d_\infty$, $\langle c(x,\nu), r(y)\rangle \in [-1, 1]$. Next, we consider the acceptance probability of A^c for any fixed $x \in \mathcal{X}$ and $\nu \in \mathcal{B}^d_\infty$, on an outcome distribution $F_Y \in \Delta(\mathcal{Y})$.

$$\begin{aligned} \mathop{\mathbf{Pr}}_{F_Y}\left[\, A^c(x, Y; \nu) = 1 \,\right] &= \int \frac{\langle c(x,\nu), r(y)\rangle + 1}{2} \cdot f_Y(y) dy \\ &= \frac{\langle c(x,\nu), \boldsymbol{\rho}_i\left\{F_Y\right\}\rangle}{2} + \frac{1}{2} \end{aligned}$$

Note that this equality implies that every A^c is $\boldsymbol{\rho}$-oblivious: the acceptance probability is a function of $\boldsymbol{\rho}\left\{F_Y\right\}$. In fact, it is also $(\boldsymbol{\rho}, 1/2)$-Lipschitz. Using the equality, we can rewrite the multicalibration violation in terms of the acceptance probabilities of the worst A^c. That is, for all $c \in \mathcal{C}$, we have the following equality

$$\begin{aligned} &\mathop{\mathbf{E}}_{\mathcal{D}^*_X}\left[\, \left\langle c(X, \tilde{\rho}(X)), \left(\boldsymbol{\rho}_i\left\{\mathcal{D}^*_{Y|X}\right\} - \tilde{\rho}_i(X)\right)\right\rangle \,\right] \\ &\qquad = 2 \cdot \left(\mathop{\mathbf{Pr}}_{\mathcal{D}^*}\left[\, A^c(X, Y^*; \tilde{\rho}(X)) = 1 \,\right] - \mathop{\mathbf{Pr}}_{\mathcal{D}^*_X}\left[\, A^c(X, \tilde{Y}; \tilde{\rho}(X)) = 1 \,\right]\right) \end{aligned}$$

where $\tilde{Y} = \mathcal{G}(\mathcal{M}(X; \tilde{\rho}(X)))$ for any individually-consistent model $\mathcal{M}$. Importantly, because A^c_i is $\boldsymbol{\rho}$-oblivious, the equality holds for all individually-consistent models. In all, taking $\mathcal{A} = \left\{A^c : c \in \mathcal{C}\right\}$, if a predictor $\tilde{\rho}$ is $(\mathcal{A}, \varepsilon)$-OI, then $\tilde{\rho}$ is $(\mathcal{C}, \alpha)$-multicalibrated for $\alpha \leq 2\varepsilon$. ■

6.2.2. MULTICALIBRATION CAPTURES OBLIVIOUS OI

We show that for a linearizing statistic $\boldsymbol{\rho}$, the acceptance probability of an oblivious distinguisher factorizes into a term that depends on the individual and prediction $(x, \tilde{\rho}(x))$ and a term that depends on the outcome y. This factorization can be leveraged for each $A \in \mathcal{A}$ to derive a function c_A, such that multicalibration over the class of $\mathcal{C} = \left\{c_A\right\}$ implies OI over $\mathcal{A}$. In this direction, the computational complexity of computing c_A given an oracle for A is less direct.

Proposition 27 (Multicalibration captures Oblivious OI) *Suppose that $\boldsymbol{\rho} : \Delta(\mathcal{Y}) \to \mathcal{B}^d_\infty$ is a bounded linearizing statistic. Then, for any family of $\boldsymbol{\rho}$-oblivious distinguisher algorithms $\mathcal{A} \subseteq \left\{\mathcal{X} \times \mathcal{Y} \times \mathbb{R}^d \to \{0,1\}\right\}$, there is a class of functions $\mathcal{C} \subseteq \left\{\mathcal{X} \times \mathbb{R}^d \to \mathcal{B}^d_1\right\}$ such that for any Nature $\mathcal{D}^*$, if a statistic predictor $\tilde{\rho} : \mathcal{X} \to \mathcal{B}^d_\infty$ is $(\mathcal{C}, \alpha)$-multicalibrated on $\mathcal{D}^*$, then $\tilde{\rho}$ is $(\mathcal{A}, \varepsilon)$-OI from Nature for $\varepsilon \leq \alpha$.*

Proof We begin by showing that for a linearizing $\boldsymbol{\rho}$, the acceptance probability of any $\boldsymbol{\rho}$-oblivious distinguisher satisfies a certain separability condition, into a term that depends on X and $\tilde{\rho}(X)$ and another term that depends on Y. First, we observe that because distinguishers output a binary decision, the behavior on $y \sim F_Y$ is a Bernoulli random variable. Specifically, for $A \in \mathcal{A}$, we can write the acceptance probability directly as an expectation.

$$\Pr_{F_Y}[\, A(x, Y, \nu) = 1 \,] = \mathop{\mathbf{E}}_{F_Y}[\, A(x, Y, \nu) \,] = \int A(x, y, \nu) f_Y(y) dy$$

Next, by $\boldsymbol{\rho}$-obliviousness, we can also express the acceptance probability directly as a function of $\boldsymbol{\rho}\{F_Y\}$, $h_A : \mathcal{X} \times \mathcal{B}^d_\infty \times \mathcal{B}^d_\infty \to [0, 1]$.

$$\Pr_{F_Y}[\, A(x, Y; \nu) = 1 \,] = h_A(x, \boldsymbol{\rho}\{F_Y\}; \nu)$$

Finally, by linearization, $\boldsymbol{\rho}\{F_Y\} = \int r(y) f_Y(y) dy$ can be expressed as the expected value of $r(Y)$, as in Lemma 25. Combining the equalities, we have two expressions for the acceptance probability of A.

$$\int A(x, y, \nu) f_Y(y) dy = h_A\left(x, \boldsymbol{\rho}\{F_Y\}; \nu\right) = h_A\left(x, \int r(y) f_Y(y) dy; \nu\right)$$

The left-most expression is an expectation, and thus, a linear functional of the probability densities $f_Y(y)$. This quantity is equal to the right-most expression, which takes a linear functional of the probability densities $f_Y(y)$, as the input $\boldsymbol{\rho}\{F_Y\}$. From this equality, we conclude that h_A must act on $\boldsymbol{\rho}\{F_Y\}$ linearly. In other words, for some scalar $\ell_A \in \mathbb{R}$ and constant $\tau_A \in \mathbb{R}$, there exists a function $c_A : \mathcal{X} \times \mathbb{R}^d \to \mathcal{B}^d_1$ such that

$$\begin{aligned} h_A\left(x, \int r(y) f_Y(y) dy; \nu\right) &= \ell_A \cdot \left\langle c_A(x, \nu), \int r(y) f_Y(y) dy \right\rangle + \tau_A \\ &= \ell_A \cdot \langle c_A(x, \nu), \boldsymbol{\rho}\{F_Y\} \rangle + \tau_A. \end{aligned}$$

Note by the assumption that c_A maps to $\mathcal{B}^d_1$, we may assume without loss of generality that $\ell_A \geq 1$ (as $\mathcal{B}^d_1$ is closed under multiplicative contractions). By taking an expectation over X, we can bound the distinguishing advantage of A in terms of the expectation of c_A and the difference between the true and predicted statistics. For some individually-consistent model $\mathcal{M}$, let $\tilde{Y} = \mathcal{G}(\mathcal{M}(X, \tilde{\rho}(X)))$ be sampled according to the prediction $\tilde{\rho}$.

$$\begin{aligned} &\Pr_{\mathcal{D}^*}[\, A(X, Y; \tilde{\rho}(X)) = 1 \,] - \Pr_{\mathcal{D}^*_X}\left[\, A(X, \tilde{Y}; \tilde{\rho}(X)) = 1 \,\right] \\ &= \mathop{\mathbf{E}}_{\mathcal{D}^*_X}\left[\, h_A\left(X, \boldsymbol{\rho}\left\{\mathcal{D}^*_{Y|X}\right\}; \tilde{\rho}(X)\right) - h_A\left(X, \tilde{\rho}(X); \tilde{\rho}(X)\right) \,\right] \\ &= \ell_A \cdot \mathop{\mathbf{E}}_{\mathcal{D}^*_X}\left[\, \left\langle c_A(X, \tilde{\rho}(X)) \cdot \left(\boldsymbol{\rho}\left\{\mathcal{D}^*_{Y|X}\right\} - \tilde{\rho}(X)\right)\right\rangle \,\right] \end{aligned}$$

Taking $\mathcal{C} = \{c_A : A \in \mathcal{A}\}$, if $\tilde{\rho}$ is $(\mathcal{C}, \alpha)$-multicalibrated, then the distinguishing advantage for each $A \in \mathcal{A}$ is upper bounded by α/ℓ_A. Because $\ell_A \geq 1$ for all $A \in \mathcal{A}$, then $\tilde{\rho}$ is $(\mathcal{A}, \varepsilon)$-OI for $\varepsilon \leq \alpha$. ■

Note that this reduction is analytical, rather than computational. In particular, c_A does not directly use A as an oracle, but rather must exist due to the properties of linearizing statistics and oblivious distinguishers. Still, we argue informally that in well-motivated settings, the complexity of c_A cannot greatly exceed that of A.

In particular, suppose that $\boldsymbol{\rho}$ is is non-degenerate, in the sense that we can choose a selection of distributions $\left\{F_Y^{(i)}\right\}$, such that $\left\{\boldsymbol{\rho}\left\{F_Y^{(i)}\right\}\right\}$ form a basis for $\mathbb{R}^d$. Then, assuming the basis is sufficiently well-conditioned, then we can "decode" c_A using oracle calls to A. In particular, for any given x, ν, using the equality derived above for each $F_Y^{(i)}$,

$$\Pr_{F_Y^{(i)}}\left[\, A(x, Y; \nu) = 1 \,\right] = h_A\left(x, \boldsymbol{\rho}\left\{F_Y^{(i)}\right\}; \nu\right) = \ell_A \cdot \left\langle c_A(x,\nu), \boldsymbol{\rho}\left\{F_Y^{(i)}\right\}\right\rangle + \tau_A$$

we can set up a linear system to evaluate $c_A(x, \nu)$. Specifically, we can call A on samples from each $F_Y^{(i)}$ to approximate the acceptance probability on the left-hand side, and explicitly calculate (or estimate) the statistic $\boldsymbol{\rho}\left\{F_Y^{(i)}\right\}$ on the right-hand side. Then, we can solve for $c_A(x, \nu)$ (up to the Lipschitz constant) with a linear system solver. The number of calls to A needed to invert the linear system accurately will depend exactly on the choice of basis based on properties of $\boldsymbol{\rho}$. While this method of deriving c_A from A is indirect, it establishes the fact that the complexity of c_A cannot grow unbounded compared to that of A.

6.3. Understanding Moment Multi-Calibration

We discuss how the connection between oblivious OI and multicalibrated statistics sheds light on the work on moment multicalibration of Jung et al. (2021). In this work, the goal is the predict central moments for individual outcomes that satisfy multicalibration. Here, they must be careful because central moments do not linearize; thus, they instead work with a technical notion of mean-conditioned moment multicalibration, which allows the statistics to linearize and thus is feasible.

We start by arguing that we can achieve the guarantees of moment multicalibration directly as a variant of statistic multicalibration using linearizing statistics. In this sense, we can similarly implement it in the oblivious OI framework. The idea is to avoid central moments all together, and instead, simply design distinguishers that accept with probability proportional to the non-central moments. Then, by conditioning on individuals' predicted means, the non-central moments distinguishers will imply central moment multicalibration.

For simplicity, we explain the construction in the case of predicting the mean and variance. Instead of working with the variance, we will work directly with the non-central second moment. For any given subpopulation $S \subseteq \mathcal{X}$, we can define the following functions[6] that effectively "condition" on membership in S, as well as the predicted mean $\tilde{\mu}_1(x)$ and predicted second moment $\tilde{\mu}_2(x)$.

$$c_{S,m_1,m_2}(x; \tilde{\mu}_1(x), \tilde{\mu}_2(x)) = \mathbf{1}\left[\, x \in S \wedge \tilde{\mu}_1(x) \approx m_1 \wedge \tilde{\mu}_2(x) \approx m_2 \,\right]$$

Note that the non-central moments linearize. Thus, by the arguments above, we can define corresponding distinguishers, whose acceptance probabilities are proportional to the statistics of interest. In particular, for any $F_Y \in \Delta(\mathcal{Y})$,

$$\Pr_{F_Y}\left[\, A^1_{S,m_1,m_2}(x, Y; \tilde{\mu}_1(x), \tilde{\mu}_2(x)) = 1 \,\right] \propto c_{S,m_1,m_2}(x; \tilde{\mu}_1(x), \tilde{\mu}_2(x)) \cdot \boldsymbol{\mu_1}\left\{F_Y\right\}$$
$$\Pr_{F_Y}\left[\, A^2_{S,m_1,m_2}(x; \tilde{\mu}_1(x), \tilde{\mu}_2(x)) = 1 \,\right] \propto c_{S,m_1,m_2}(x; \tilde{\mu}_1(x), \tilde{\mu}_2(x)) \cdot \boldsymbol{\mu_2}\left\{F_Y\right\}$$

[6] Note that we describe the indicator functions on predictions informally, using $\approx$ as shorthand for the formal rounding strategy discussed in Jung et al. (2021).

Importantly, by restricting attention to individuals who received a given mean prediction $\tilde{\mu}_1(X) = m_1$ and second moment prediction $\tilde{\mu}_2(X) = m_2$, fooling the distinguishers implies multicalibration for the variance predictor derived as

$$\tilde{\sigma}(x) = \tilde{\mu}_2(x) - (\tilde{\mu}_1(x))^2.$$

In particular, when we condition on both predictions $\tilde{\mu}_1(X)$ and $\tilde{\mu}_2(X)$, the linearizing statistics are still accurate in expectation, and thus, give a valid multicalibrated estimate of the variance.

Mean-conditioning is necessary. Finally, we argue that, in a sense, the techniques of Jung et al. (2021) to obtain central moment multicalibration are necessary. We argue that, without conditioning on the mean, there is no hope for a meaningful guarantee for central moments from multicalibration, or even from Existential OI. In particular, it would be impossible to derive any sorts of concentration inequalities from the predicted statistics, as Jung et al. (2021) show is possible with mean-conditioned moment multicalibration.

This argument follows from the following proposition, which intuitively says that the trivial central moment predictor, that claims all central moments are 0, can always pass any OI tests, provided the distinguishers has no information about the predicted mean. For convenience we state the proposition in terms of a predictor $\tilde{\mu} : \mathcal{X} \to \mathbb{R}^d$ that predicts the first $d+1$ central moments, excluding the mean.

Proposition 28 *Suppose $\tilde{\mu} : \mathcal{X} \to \mathbb{R}^d$ that predicts the first $d+1$ central moments, excluding the mean. For any family of distinguishers $\mathcal{A}$, the trivial predictor $\tilde{\mu}_0(x) = \mathbf{0}$ for all $x \in \mathcal{X}$, is $(\mathcal{A}, \varepsilon)$-existential-OI for any constant $\varepsilon > 0$.*

Proof Consider, for the sake of argument, drawing a random sample $Y_x \sim \mathcal{D}_{Y|X}$ for each $x \in \mathcal{X}$. Further, consider the model $\mathcal{M} : \mathcal{X} \to \Delta(\mathcal{Y})$ that for each $x \in \mathcal{X}$, returns the singleton distribution where $\mathbf{Pr}_{\mathcal{M}(x)}\left[\, Y = Y_x \,\right] = 1$. Note that because each outcome under $\mathcal{M}$ is deterministic, every central moment beyond the mean of each of $\mathcal{M}(x)$ is 0. Thus, $\mathcal{M}$ is individually-consistent with $\tilde{\mu}_0$.

But now, consider the acceptance probability of any distinguisher A on samples from $\mathcal{D}^*$ versus on samples from $\mathcal{D}(\mathcal{M})$. By the choice of $\mathcal{M}$, samples from $\mathcal{D}(\mathcal{M})$ are true empirical samples from $\mathcal{D}^*$. Thus, the acceptance probability on $\mathcal{D}(\mathcal{M})$ can be viewed as an empirical approximation of the acceptance probabilty on $\mathcal{D}^*$. Thus, provided that $\mathcal{D}^*_X$ has sufficient min-entropy to avoid repeatedly sampling the same individual X regularly, the acceptance probability on $\mathcal{D}(\mathcal{M})$ will concentrate around that on $\mathcal{D}^*$. By choosing $\mathcal{D}^*_X$ appropriately, we can obtain ε-closeness in distinguishing advantage for arbitrarily small $\varepsilon > 0$. ■

This construction is a specific example of a general phenomenon: without conditioning on the mean of the outcome distribution, the set of individually-consistent models includes unreasonable, and largely inaccessible generative outcome models. For example, because our choice of $\mathcal{M}$ is random, it will be highly incompressible.

In this sense, some variant of this construction can be made to work with any notion of OI, provided the distinguishers only look at consistency with the predicted central moments. In contrast, the results of Jung et al. (2021) show the power of conditioning on the mean. Once we condition on the mean, then even the weakest notion of oblivious OI is capable of enforcing strong consistency with Nature, allowing us to derive Chebyshev-style concentration inequalities based on the predicted statistics.

Acknowledgments

All authors are involved in the Simons Foundation Collaboration on the Theory of Algorithmic Fairness. **CD** acknowledges the Alfred P. Sloan Foundation, "Pseudo-Randomness and the Crystal Ball," G-2020-13941. **MPK** is supported by the Miller Institute for Basic Research in Science. **OR** is supported by the Simons Foundation Collaboration on the Theory of Algorithmic Fairness, the Sloan Foundation Grant 2020-13941, the Simons Foundation investigators Award 689988 and NSF Award CCF-1763311. **GNR** 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), and from the U.S-Israel Binational Science Foundation (grant No. 2018102). **GY** is supported by the European Research Council (ERC) (grant agreement No. 819702), by the Israel Science Foundation (grant number 5219/17), by the Simons Foundation Collaboration on the Theory of Algorithmic Fairness, by the Israeli Council for Higher Education via the Weizmann Data Science Research Center, by a Google PhD fellowship, and by a grant from the Estate of Tully and Michele Plesser.

References

Noam Barda, Gal Yona, Guy N Rothblum, Philip Greenland, Morton Leibowitz, Ran Balicer, Eitan Bachmat, and Noa Dagan. Addressing bias in prediction models by improving subpopulation calibration. *Journal of the American Medical Informatics Association*, 28(3):549–558, 2021.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

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, 2012.

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.

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.

Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Omnipredictors. *arXiv preprint arXiv:2109.05389*, 2021a.

Parikshit Gopalan, Omer Reingold, Vatsal Sharan, and Udi Wieder. Multicalibrated partitions for importance weights. *arXiv preprint arXiv:2103.05853*, 2021b.

Varun Gupta, Christopher Jung, Georgy Noarov, Mallesh M Pai, and Aaron Roth. Online multivalid learning: Means, moments, and prediction intervals. *arXiv preprint arXiv:2101.01739*, 2021.

Ursula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Multicalibration: Calibration for the (Computationally-identifiable) masses. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1939–1948, 2018.

Christopher Jung, Changhwa Lee, Mallesh Pai, Aaron Roth, and Rakesh Vohra. Moment multicalibration for uncertainty estimation. In *Conference on Learning Theory*, pages 2634–2678. PMLR, 2021.

Adam Kalai. Learning monotonic linear functions. In *International Conference on Computational Learning Theory*, pages 487–501. Springer, 2004.

Adam Tauman Kalai and Rocco A Servedio. Boosting in the presence of noise. *Journal of Computer and System Sciences*, 71(3):266–290, 2005.

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 2564–2572, 2018.

Michael J Kearns, Robert E Schapire, and Linda M Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2-3):115–141, 1994.

Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Fairness through computationally-bounded awareness. *Advances in Neural Information Processing Systems*, 2018.

Michael P. Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 247–254, 2019.

Yishay Mansour and David McAllester. Boosting using branching programs. *Journal of Computer and System Sciences*, 64(1):103–112, 2002.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

Guy N Rothblum and Gal Yona. Multi-group agnostic pac learnability. *ICML*, 2021.

Eliran Shabat, Lee Cohen, and Yishay Mansour. Sample complexity of uniform convergence for multicalibration. *arXiv preprint arXiv:2005.01757*, 2020.

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