Modern Paradigms in Learning and Decision-Making
Lecture #2: Outcome Indistinguishability: A New Lens on Prediction*

Michael P. Kim & Juan C. Perdomo†

Fall 2026

1 Loss Minimization and Outcome Indistinguishability

Traditionally, supervised learning has been framed in terms of loss minimization. We assume that the population we are trying to make predictions over is well modeled by a fixed probability distribution $\mathcal{D}$ over features $x \in \mathcal{X}$ and outcomes $y \in \mathcal{Y}$.

Given a loss function $\ell : \hat{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}^+$ where $\hat{\mathcal{Y}}$ is the prediction space, a hypothesis class $\mathcal{H} \subseteq \{h : \mathcal{X} \to \hat{Y}\}$, and an error parameter $\varepsilon \geq 0$, the goal is to find a predictor $p : \mathcal{X} \to \hat{\mathcal{Y}}$ that achieves loss that competes (up to ε) with the loss-minimizing hypothesis $h \in \mathcal{H}$. We assume throughout the lecture that $\hat{\mathcal{Y}} = [0, 1]$.

$$\mathbf{E}[\ell(p(x), y)] \leq \min_{h \in \mathcal{H}} \ \mathbf{E}[\ell(h(x), y)] + \varepsilon$$

Loss minimization is ubiquitous as a notion of supervised learning. It captures the Agnostic[1] PAC model of learning [Val84, KSS94, Hau92] if we fix ℓ to be the 01-loss:

$$\ell_{01}(h(x), y) = \mathbf{1}[h(x) \neq y].$$

Loss minimization can also encode regression problems, like least squares if we set ℓ to be the squared error:

$$\ell^2(h(x), y) = (h(x) - y)^2.$$

While loss minimization has proven to be very powerful as a paradigm, there are several common critiques that we should bear in mind. First and foremost, loss minimization is an on-average guarantee that holds over the entire population or distribution $\mathcal{D}$. It makes no guarantees with respect to specific groups of people or subpopulations. Second, it is not robust to changes in distribution or in the loss. If the population drifts and the distribution $\mathcal{D}$ changes, the loss minimization guarantee over $\mathcal{D}$ becomes vacuous. Furthermore, if we train under one objective ℓ and later decide our goals were better reflected by a different loss ℓ', we need to retrain from scratch.

*These notes are a work in progress and have not been subjected to the usual scrutiny reserved for formal publications.

[1]The term agnostic refers to the fact that there is no realizability assumption that any specific $h \in \mathcal{H}$ can fit the data perfectly, i.e. has 0-1 loss exactly 0.

Here, we introduce a different paradigm for learning, called Outcome Indistinguishability (OI). Rather than specifying the goal of learning through the minimization of a loss function, OI defines the goal as *indistinguishability*. That is, OI will require that our predictor $\tilde{p} : \mathcal{X} \to [0,1]$ "looks like" the Bayes predictor $p^*(x) = \mathbf{Pr}[Y = 1|X = x]$, from the perspective of a set of tests.

This notion of learning as indistinguishability is directly inspired by the notion of computational indistinguishability from pseudorandomness and cryptography. It is both conceptually fascinating in its own right and technically valuable as a solution concept that addresses several of the shortcomings in loss minimization.

Two Worlds: Building a Model of Nature. Intuitively, OI can be described through the following game: you are presented with a triple $(x, y, \tilde{p}(x))$ of an individual $x \in \mathcal{X}$, an outcome $y \in \{0,1\}$, and a probability $\tilde{p}(x) \in [0,1]$. In addition, you're promised that y is sampled from one of two worlds.

- Nature: in Nature's world, the outcome $y^* \sim \text{Ber}(p^*(x))$ is sampled according to the true conditional distribution of outcomes given the individual x.
- Model: in the modeled world, the outcome $\tilde{y} \sim \text{Ber}(\tilde{p}(x))$ is sampled *according to the prediction presented to you in the triple*. That is, given the individual x, we resample the outcome according to the outcome distribution implied by the predictor $\tilde{p}$.

To win the game, you must guess whether the triple—and specifically the *outcome*—was sampled from Nature or the Model. We say that the predictor $\tilde{p}$ is outcome indistinguishable, if you *cannot* win the game; that is, if the two worlds of outcomes are indistinguishable.

More formally, we use a collection of distinguishers to test our triples. Every function $A : \mathcal{X} \times [0,1] \times \mathcal{Y} \to \mathbb{R}$ takes in an input-outcome-prediction triple and outputs a number. We can think of each A as defining a test, where the function "guesses" whether the outcome was drawn according to Nature or the model (it's helpful to think of the output as 1 or 0, pass or fail).

The goal of OI is to find a model that "fools" all of the distinguishers in a broad class $A \in \mathcal{A}$. Specifically, a predictive model satisfies OI if for every distinguisher, the probability of distinguisher acceptance in the modeled world is nearly identical to the probability in Nature's world.

Definition 1.1 (Outcome Indistinguishability [DKR+21])**.** *Fix a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$, a class of distinguishers $\mathcal{A} \subseteq \{A : \mathcal{X} \times [0,1] \times \{0,1\} \to \mathbb{R}$, and an error tolerance $\varepsilon \geq 0$.*

A predictor $\tilde{p} : \mathcal{X} \to [0,1]$ is $(\mathcal{A}, \varepsilon)$-outcome indistinguishable if for all $A \in \mathcal{A}$:

$$\left| \mathop{\mathbf{E}}_{\substack{x \sim \mathcal{D}_x \\ y^* \sim \text{Ber}(p^*(x))}} [A(x, \tilde{p}(x), y^*)] - \mathop{\mathbf{E}}_{\substack{x \sim \mathcal{D}_{\mathcal{X}} \\ \tilde{y} \sim \text{Ber}(\tilde{p}(x))}} [A(x, \tilde{p}(x), \tilde{y})] \right| \leq \varepsilon.$$

As a notion of learning, OI is parameterized by the collection of distinguishers $\mathcal{A}$, which directly controls the strength of the guarantee. If $\mathcal{A}$ only contains a few simple tests, then $\mathcal{A}$-OI is a relatively weak notion of learning. For instance, assume A takes in $(x, \tilde{p}(x), y)$ and outputs y, then any $\tilde{p}(x)$ that has the right output on average over the entire population passes this test.

$$\mathop{\mathbf{E}}_{(x,y^*)\sim\mathcal{D}}[\tilde{p}(x)] = \mathop{\mathbf{E}}_{\substack{x\sim\mathcal{D}_{\mathcal{X}}\\ \tilde{y}\sim\mathrm{Ber}(\tilde{p}(x))}}[\tilde{y}] = \mathop{\mathbf{E}}_{(x,y^*)\sim\mathcal{D}}[y^*]$$

Taking a richer collection $\mathcal{A}$ of more complicated tests will strengthen the OI guarantee. In particular, you will eventually prove the following facts as an exercise illustrating how OI can recover traditional notions of statistical distance.

Exercise. Let $\mathcal{X}$ a be finite, discrete set, let $p^*(x) = \mathbf{Pr}_{\mathcal{D}}[Y = 1|\mathcal{X} = x]$, and fix a predictor $\tilde{p} : \mathcal{X} \to [0, 1]$. Define the ℓ_1 distance between $\tilde{p}$ and p^*

$$\ell_1(\tilde{p}, p^*) = \frac{1}{2}\sum_{x\in\mathcal{X}} |\tilde{p}(x) - p^*(x)|.$$

Construct a class of tests $\mathcal{A}$ such that if $\tilde{p}$ is ε-OI with respect to $\mathcal{A}$ over the distribution $\mathcal{D}$ then $\ell_1(\tilde{p}, p^*) \leq \varepsilon$.

Exercise. Let $\mathcal{A}$ be the class of all functions $A(x, p, y)$ that are 1-Lipschitz with respect to their second argument. Then $\tilde{p}$ is ε-OI with respect to $\mathcal{A}$ if and only if the earthmover's or Wasserstein 1 distance between p^* and $\tilde{p}$ is at most ε. That is,

$$\sup_{f\in\mathcal{F}} \mathop{\mathbf{E}}_{x\sim\mathcal{D}_x} f(p^*(x)) - \mathop{\mathbf{E}}_{x\sim\mathcal{D}_x} f(\tilde{p}(x)) \leq \varepsilon, \quad \mathcal{F} = \{f : \mathbb{R} \to \mathbb{R}, f \text{ is } 1-\text{Lipschitz}\}$$

We pause to make a few more comments about the definition.

- *Tests vs. Distinguishers.* We refer to $\mathcal{A}$ as a collection of distinguishers or tests interchangeably. Notice, however for binary valued A, the goal of OI is not to "pass" the tests in the sense of finding a $\tilde{p}$ such that $A(x, \tilde{p}(x), \tilde{y}) = 1$ with high probability. Instead, the goal is to pass the test with the same probability as Nature passes. In other words, in the indistinguishability framework, there is nothing inherently good or bad about passing or failing the tests.

- *Computationally-bounded Distinguishers.* In line with pseudorandomness and cryptography, in OI, it is most useful to think of the collection of distinguishers $\mathcal{A}$ as a collection of *efficient* tests. These tests may be defined by a class of algorithms (hence "$\mathcal{A}$"), though are often defined in terms of a (non-uniform) class of computations, relevant to the problem at hand. At the low end of the spectrum, we might take $\mathcal{A}$ to be defined by linear tests or small decision trees, whereas at the upper end of the spectrum, we might take $\mathcal{A}$ to be functions implementable by neural networks or all polynomial-sized circuits.

  As we will see, there is a particular appeal to selecting $\mathcal{A}$ to be a *learnable* class of computations. This is the key conceptual difference relative to statistical notions like TV. As seen in the exercises above, the class of tests in those settings is so complex that we cannot efficiently learn it.

1.1 OI captures Loss Minimization

Above, we have introduced OI as a new notion of learning based on computational indistinguishability. A natural question to ask ourselves, whenever we introduce a new definition, is... *Why? What does this notion give us that prior notions did not?*

We will get to answering these questions, but as a starting point, we begin by showing that—at the very least— we have not lost anything by moving away from loss minimization to OI as our solution concept. In particular, OI is capable of "implementing" loss minimization. Here, we establish how OI can implement loss minimization for any proper scoring rule. For binary y, we say that a loss function is *proper* if for every $p \in [0,1]$:

$$p \in \operatorname*{argmin}_{t \in [0,1]} \mathop{\mathbf{E}}_{\tilde{y} \sim \mathrm{Ber}(p)} \ell(t, \tilde{y}).$$

Proposition 1.2 ([GHK+23]). *Fix any hypothesis class $\mathcal{H} \subseteq \{h : \mathcal{X} \to [0,1]\}$, error parameter $\varepsilon \geq 0$, and proper loss function ℓ. Assume that $\tilde{p}$ is ε-OI with respect to the following class of functions*

$$A(x,p,y) = \ell(p,y) \text{ and } A_h(x,p,y) = \ell(h(x),y)$$

for every $h \in \mathcal{H}$. Then,

$$\mathop{\mathbf{E}}_{\substack{x \sim \mathcal{D}_x \\ y^* \sim \mathrm{Ber}(p^*(x))}} \ell(\tilde{p}(x), y^*) \leq \min_{h \in \mathcal{H}} \mathop{\mathbf{E}}_{\substack{x \sim \mathcal{D}_x \\ y^* \sim \mathrm{Ber}(p^*(x))}} [\ell(h(x), y^*)] + 2\varepsilon.$$

The proposition states that for any hypothesis class $\mathcal{H}$ and proper loss ℓ, we can write down a set of tests such that if $\tilde{p}$ is OI with respect to those tests, then $\tilde{p}$ is also loss minimizing with respect to $\mathcal{H}$.

The proof is very simple. The OI requirements are that

$$\left| \mathop{\mathbf{E}}_{\substack{x \sim \mathcal{D}_x \\ y^* \sim \mathrm{Ber}(p^*(x))}} [\ell(\tilde{p}(x), y^*)] - \mathop{\mathbf{E}}_{\substack{x \sim \mathcal{D}_\mathcal{X} \\ \tilde{y} \sim \mathrm{Ber}(\tilde{p}(x))}} [\ell(\tilde{p}(x), \tilde{y})] \right| \leq \varepsilon. \tag{1}$$

and that for every $h \in \mathcal{H}$,

$$\left| \mathop{\mathbf{E}}_{\substack{x \sim \mathcal{D}_x \\ y^* \sim \mathrm{Ber}(p^*(x))}} [\ell(h(x), y^*)] - \mathop{\mathbf{E}}_{\substack{x \sim \mathcal{D}_\mathcal{X} \\ \tilde{y} \sim \mathrm{Ber}(\tilde{p}(x))}} [\ell(h(x), \tilde{y})] \right| \leq \varepsilon. \tag{2}$$

Therefore, for any $h \in \mathcal{H}$

$$\begin{aligned} \mathop{\mathbf{E}}_{\substack{x \sim \mathcal{D}_x \\ y^* \sim \mathrm{Ber}(p^*(x))}} \ell(\tilde{p}(x), y^*) &\leq \mathop{\mathbf{E}}_{\substack{x \sim \mathcal{D}_x \\ \tilde{y} \sim \mathrm{Ber}(\tilde{p}(x))}} \ell(\tilde{p}(x), \tilde{y}) + \varepsilon \\ &\leq \mathop{\mathbf{E}}_{\substack{x \sim \mathcal{D}_x \\ \tilde{y} \sim \mathrm{Ber}(\tilde{p}(x))}} \ell(h(x), \tilde{y}) + \varepsilon \\ &\leq \mathop{\mathbf{E}}_{\substack{x \sim \mathcal{D}_x \\ y^* \sim \mathrm{Ber}(p^*(x))}} \ell(h(x), y^*) + 2\varepsilon \end{aligned}$$

The first inequality follows from (1), the second inequality follows from the fact that ℓ is a proper scoring rule and $\tilde{y} \sim \text{Ber}(\tilde{p}(x))$, and the last line follows from (2). Since these 3 inequalities are true for every $h \in \mathcal{H}$, they in particular hold for the best $h \in \mathcal{H}$.

It is worth noting a few aspects of the proof of Proposition 1.2 that will come up over and over again in OI proofs.

- *Tailored distinguisher class.* In our proof that OI generalizes loss minimization, we hand-craft a class of distinguishers $\mathcal{A}_{\mathcal{H}}$, based on the hypothesis class $\mathcal{H}$ (and the loss function at hand). Note that because OI universally quantifies over all distinguishers within $\mathcal{A}_{\mathcal{H}}$, OI allowed us to reason about how $\tilde{p}$ compares to the loss minimizing $h \in \mathcal{H}$. A key aspect to working with OI is figuring out what tests are necessary for your application.
- *Working in the Modeled World.* A key aspect of many OI proofs is figuring out how to "transport" the argument from the real world (where Nature controls the outcomes) to the modeled world (where our predictor controls the outcomes). As soon as we're in the modeled world, we have the immense advantage that *we know the distribution of outcomes given inputs.* Intuitively, if we can run the argument with knowledge of $\tilde{p}(x)$, and real outcomes $y^* \sim \text{Ber}(p^*(x))$ look like modeled outcomes $\tilde{y} \sim \text{Ber}(\tilde{p}(x))$, then we can run the argument in the real world (importantly, still, with knowledge of only $\tilde{p}(x)$, not $p^*(x)$).

At this point, it's hopefully clear we haven't lost anything in moving from loss minimization to OI. However, it isn't yet clear what we have gained by doing so. We will figure this out together over the next few lectures. However, before we discuss the many benefits of OI, let's first understand how one might learn an OI predictor.

2 Learning Outcome Indistinguishable Predictors

When thinking about learning a predictor $\tilde{p}$ that is OI with respect to a class of tests $\mathcal{A}$, there are two main sources of hardness:

- *How much data do we need?*

  OI is defined with respect to a distribution $\mathcal{D}$. In practice, we do not know $\mathcal{D}$ but rather have access to a small dataset $S = \{(x_i, y_i)\}_{i=1}^n \sim \mathcal{D}^{(n)}$. How large of a random sample do we need to achieve ε-OI with respect to a class of tests $\mathcal{A}$?

- *How much computation do we need?*

  At first glance, it is not immediately obvious that OI predictors even *exist* if we do not make any restrictions on the class $\mathcal{A}$ or the distribution $\mathcal{D}$. This is a challenge we will need to address. Furthermore, even if they did exist, how do we guarantee that there is a small circuit to compute them if the class of tests $\mathcal{A}$ is exponentially large?

To start to address these questions, we will introduce a different analytical perspective on outcome indistinguishability using discrete derivatives:

Definition 2.1 (Discrete Derivative)**.** *For binary outcomes $y \in \{0,1\}$, the discrete derivative of a distinguisher function $A : \mathcal{X} \times [0,1] \times \{0,1\} \to \mathbb{R}$ is*

$$\Delta A(x, p(x)) = A(x, p(x), 1) - A(x, p(x), 0).$$

Note that, if the distinguisher does not access the context or prediction, we can still define a discrete derivative. The point is that ΔA considers the difference in the distinguisher's behavior on a $y = 1$ outcome and $y = 0$ outcome. This difference in behavior on $y = 1$ and $y = 0$ is in fact the *only* signal that the distinguishers A capture.

Let $A(x, t, y)$ be any test. Since $y \in \{0, 1\}$, we can rewrite $A(x, t, y)$ using indicator functions as,

$$\begin{aligned} A(x,t,y) &= A(x,t,0)\mathbf{1}\{y=0\} + A(x,t,1)\mathbf{1}\{y=1\} \\ &= A(x,t,0)(1 - 1\{y=1\}) + A(x,t,1)\mathbf{1}\{y=1\} \\ &= [A(x,t,1) - A(x,t,0)] \cdot \mathbf{1}\{y=1\} + A(x,t,0). \end{aligned}$$

Letting $t = p(x)$, given any distribution $\mathcal{D}$ over (x, y^*) where $p^*(x) = \mathbf{Pr}_{\mathcal{D}}[y = 1|\mathcal{X} = x]$, we can write

$$\begin{aligned} \mathop{\mathbf{E}}_{(x,y)\sim\mathcal{D}}[A(x, p(x), y)] &= \mathop{\mathbf{E}}_{\mathcal{D}}[[A(x, p(x), 1) - A(x, p(x), 0)] \cdot \mathbf{1}\{y = 1\} + A(x, p(x), 0)] \\ &= \mathop{\mathbf{E}}_{x}[\Delta A(x, p(x)) \cdot p^*(x) + A(x, p(x), 0)] \end{aligned}$$

Consequently, we can (re)write the indistinguishability error of a predictor $\tilde{p}(x)$ with respect to any test A as

$$\mathop{\mathbf{E}}_{(x,y)\sim\mathcal{D}}[A(x, \tilde{p}(x), y)] - \mathop{\mathbf{E}}_{\substack{x\sim\mathcal{D}_x \\ \tilde{y}\sim\mathrm{Ber}(\tilde{p}(x))}}[A(x, \tilde{p}(x), \tilde{y})] = \mathop{\mathbf{E}}_{x\sim\mathcal{D}_x}[\Delta A(x, \tilde{p}(x)) \cdot (p^*(x) - \tilde{p}(x))].$$

Note that the "bias" term $A(x, \tilde{p}(x), 0)$ cancels out when we take the difference.

A predictor $\tilde{p}(x)$ is hence outcome indistinguishable with respect to a test A if, on average over the distribution $\mathcal{D}$, the residual $\tilde{p}(x) - p^*(x)$ is uncorrelated with the discrete derivative ΔA. Said otherwise, the only thing a distinguisher can examine is if there is any correlation between $A(x, \tilde{p}(x), 1) - A(x, \tilde{p}(x), 0)$ and $\tilde{p}(x) - p^*(x)$. We summarize this insight in the next proposition:

Proposition 2.2. *Assume outcomes y are binary. A predictor $\tilde{p} : \mathcal{X} \to [0,1]$ is ε-OI with respect to a class of tests $\mathcal{A}$ if and only if*

$$\sup_{A\in\mathcal{A}} \left| \mathop{\mathbf{E}}_{x\sim\mathcal{D}_x}[\Delta A(x, \tilde{p}(x)) \cdot (\tilde{p}(x) - p^*(x))] \right| \leq \varepsilon.$$

This rewriting hopefully reminds us of the BOOST procedure from the last lecture and inspires a simple algorithmic approach to achieving OI. To see the connection a bit more clearly, recall the audit and boosting procedures from the prior lecture, now slightly adapted to the new OI definition:

```
AUDIT(p, calA, eps)
    IF there is a test A in the class calA such that p is not eps-OI w.r.t. A
        return A
    ELSE
        return PASS
```

```
BOOST(calA, eps):
Initialize p_0(x) = 0 for all x
For t=0,1,2,...
    Call AUDIT(calA, p_t, eps)
    If AUDIT returns a function A_t, update p_{t+1} = p_t - eta_t Delta A_t
    If AUDIT returns PASS, return p_t
```

Using the same potential style argument as before, we can prove that the BOOST algorithm returns a predictor $\tilde{p}$ that is ε-OI with respect to any class of tests $\mathcal{A}$. The following lemma assumes that we have access to a subroutine that can implement the AUDIT function with respect to $\mathcal{D}$ which we will soon figure out how to actually implement.

Lemma 2.3. *Let $\mathcal{A}$ be any class of distinguishers such that $\sup_{A\in\mathcal{A}} |\Delta A(x)| \leq 1$. The BOOST procedure returns a predictor $\tilde{p}$ that is ε-OI with respect to $\mathcal{A}$ in at most $\lceil 1/\varepsilon^2 \rceil$ many rounds.*

Proof. The proof follows the same template as before. The only difference is that we need to slightly change the potential function. Define

$$\Phi_t = \mathop{\mathbf{E}}_{\mathcal{D}}[(p^*(x) - p_t(x))^2]$$

By the previous proposition, if p_t is not ε-OI with respect to a test A, then

$$\left| \mathop{\mathbf{E}}_{x\sim\mathcal{D}_x} [\Delta A(x, p_t(x)) \cdot (p^*(x) - p_t(x))] \right| > \varepsilon.$$

Therefore, whenever we make an update, we have that

$$\begin{aligned}
\Phi_{t+1} &= \mathop{\mathbf{E}}_{\mathcal{D}}[(p^*(x) - p_{t+1}(x))^2] \\
&= \mathop{\mathbf{E}}_{\mathcal{D}}[(p^*(x) - p_t(x) + \eta_t \Delta A(x, p_t(x)))^2] \\
&= \Phi_t + 2\eta_t \mathop{\mathbf{E}}_{x\sim\mathcal{D}_x} [\Delta A(x, p_t(x)) \cdot (p^*(x) - p_t(x))] + \eta_t^2 \mathop{\mathbf{E}}_{x\sim\mathcal{D}_x} [\Delta A(x, p_t(x))^2]
\end{aligned}$$

Again setting $\eta_t = -\varepsilon \cdot \text{sign}(\mathbf{E}_{x\sim\mathcal{D}_x}[\Delta A(x, p_t(x)) \cdot (p^*(x) - p_t(x))])$, and using the assumption that $\Delta A(x, t)$ is uniformly bounded by 1, we get that the last expression is at most

$$\Phi_t - 2\varepsilon^2 + \varepsilon^2$$

Since $p_0 = 0$ and $p^*(x) \in [0, 1]$, we have that $\Phi_0 \leq 1$ and $\Phi_t \geq 0$ for any t since the potential is nonnegative. Therefore, the procedure must halt after $\lceil 1/\varepsilon^2 \rceil$ many iterations since every iteration decreases the potential by (at least) $\varepsilon^2 > 0$. □

2.1 Implementing the Audit Function in Finite Samples

Given that the BOOSTING procedure works just as well in this new setting, the last thing we need to address is how to implement the AUDIT function. Notice that while indistinguishability is defined with respect to the true distribution $\mathcal{D}$, any real algorithm must work in finite samples.

To deal with this gap, we use the following standard fact relating the mean and sample average of bounded random variables.

Lemma 2.4 (Hoeffding's Inequality)**.** *Let $f : \mathcal{Z} \to [-1, 1]$ be any function, and let $\{z_i\}_{i=1}^n \sim \mathcal{D}^n$ be n i.i.d samples drawn from the distribution $\mathcal{D}$, then*

$$\mathop{\mathbf{Pr}}_{\mathcal{D}}\left[\left|\frac{1}{n}\sum_{i=1}^{n} f(z_i) - \mathop{\mathbf{E}}_{z\sim\mathcal{D}}[f(z)]\right| \geq t\right] \leq 2\exp(-t^2 n).$$

Consequently,

$$\mathop{\mathbf{Pr}}_{\mathcal{D}}\left[\left|\frac{1}{n}\sum_{i=1}^{n} f(z_i) - \mathop{\mathbf{E}}_{z\sim\mathcal{D}}[f(z)]\right| \leq \sqrt{\frac{\log(2/\delta)}{n}}\right] \geq 1 - \delta.$$

By using Hoeffding along with a standard union bound, we can prove the following proposition. It states that given a sufficiently large dataset, any distinguisher whose empirical OI error is large, must be in violation of the indistinguishability condition at the population level:

Proposition 2.5. *Fix a predictor $\tilde{p} : \mathcal{X} \to [0, 1]$. Let $S = \{(x_i, y_i)\}_{i=1}^n$ be a dataset of n i.i.d. samples from a distribution $\mathcal{D}$ and let $\mathcal{A}$ be a finite class of tests such that $\sup_{A\in\mathcal{A}} |\Delta A| \leq 1$. For $A \in \mathcal{A}$, define the sample average. That is,*

$$\widehat{A} = \frac{1}{n}\sum_{i=1}^{n} \Delta A(x_i, \tilde{p}(x_i))(y_i - \tilde{p}(x_i)).$$

Fix parameters $\varepsilon, \delta > 0$ and assume $n \geq \log(|\mathcal{A}|/\delta)/\varepsilon^2$. For any $A \in \mathcal{A}$, if the predictor $\tilde{p}$ is not 2ε-OI with respect to A,

$$\left|\mathop{\mathbf{E}}_{\substack{x\sim\mathcal{D}_x \\ y^*\sim\mathrm{Ber}(p^*(x))}}[A(x, \tilde{p}(x), y^*)] - \mathop{\mathbf{E}}_{\substack{x\sim\mathcal{D}_x \\ \tilde{y}\sim\mathrm{Ber}(\tilde{p}(x))}}[A(x, \tilde{p}(x), \tilde{y})]\right| > 2\varepsilon,$$

then $|\widehat{A}| > \varepsilon$.

Proof. Define

$$\mathsf{OIErr}_A = \mathop{\mathbf{E}}_{\substack{x\sim\mathcal{D}_x \\ y^*\sim\mathrm{Ber}(p^*(x))}}[A(x, \tilde{p}(x), y^*)] - \mathop{\mathbf{E}}_{\substack{x\sim\mathcal{D}_{\mathcal{X}} \\ \tilde{y}\sim\mathrm{Ber}(\tilde{p}(x))}}[A(x, \tilde{p}(x), \tilde{y})]$$

By definition of $\widehat{A}$ and the previous derivation using the discrete derivative, for any $A \in \mathcal{A}$, we have that $\mathbf{E}_{\mathcal{D}}[\widehat{A}] = \mathsf{OIErr}_A$. Therefore, for any fixed $A \in \mathcal{A}$, by Hoeffding's inequality,

$$\mathop{\mathbf{Pr}}_{\mathcal{D}}\left[\left|\widehat{A} - \mathsf{OIErr}_A\right| \geq \varepsilon\right] \leq 2\exp(-2\varepsilon^2 n).$$

Using a union bound, we can bound the deviations uniformly over all tests in $\mathcal{A}$. That is,

$$\mathop{\mathbf{Pr}}_{\mathcal{D}}\left[\exists A \in \mathcal{A} \text{ s.t } \left|\widehat{A} - \mathsf{OIErr}_A\right| \geq \varepsilon\right] \leq 2|\mathcal{A}|\exp(-2\varepsilon^2 n)$$

Setting n as in the theorem statement, we get that $2|\mathcal{A}|\exp(-2\varepsilon^2 n) \leq \delta$. Therefore, with probability $1-\delta$,

$$\sup_{A\in\mathcal{A}} \left| \widehat{A} - \mathsf{OIErr}_A \right| \leq \varepsilon.$$

Since $|\mathsf{OIErr}_A| > 2\varepsilon$ and $|\widehat{A} - \mathsf{OIErr}_A| \leq \varepsilon$ it must be the case that $|\widehat{A}| \geq \varepsilon$. □

This in particular means that we can implement the audit function in finite samples by enumerating through all the tests $A \in \mathcal{A}$ at every iteration and computing

$$A_t = \operatorname*{argmax}_{A\in\mathcal{A}} |\frac{1}{n}\sum_{i=1}^{n} \Delta A(x_i, p_t(x_i))(y_i - p_t(x_i))| \tag{3}$$

If $\widehat{A}_t > \varepsilon$, then we return A_t otherwise we return PASS.

This result addresses the statistical complexity of implementing the audit procedure. It is however computationally inefficient for superpolynomially sized $\mathcal{A}$, since without assuming any further structure, we need to iterate through $\mathcal{A}$ to find the argmax. In the next exercise, you will show how this search problem can be solved efficiently for special classes of tests.

Exercise. Let $\Phi : \mathcal{X} \times [0,1] \to \mathbb{R}^d$ be a fixed feature map and let $B > 0$ be a constant. Assume that the set of tests $\mathcal{A}$ satisfies the property that for every $A \in \mathcal{A}$, there exists a vector $\theta \in \mathbb{R}^d$ with $\|\theta\|_2 \leq B$ such that

$$\Delta A(x,t) = \Phi(x,t)^\top \theta.$$

Prove that the optimization problem in (3) is a convex optimization problem.

Remark. In the previous proposition, we assumed that the predictor $\tilde{p}$ was fixed and independent of the dataset S. In the AUDIT procedure, however, the predictor at round t is a function of the choices of tests that we boosted on which are in fact determined by S. This *adaptivity* where the questions we ask depend on the data can lead to issues of overfitting.[2]

To address it, we either need to union bound over all of the possible questions we *could* have asked, or simply collect a fresh sample of $\mathcal{O}(1/\varepsilon^2)$ many samples at every time step. Here, we will opt for the latter. We summarize our analysis so far in the following theorem statement.

Theorem 2.6. *Let $\mathcal{A}$ be a finite class of tests such that $\sup_{A\in\mathcal{A}} |\Delta A| \leq 1$. Assume that the AUDIT procedure is implemented via empirical risk minimization as in Equation (3) where at every time step t, one collects a new dataset $S \sim \mathcal{D}^{(n)}$ for $n \geq \Omega((\log(|\mathcal{A}|) + \log(1/\delta) + \log(1/\varepsilon))/\varepsilon^2)$.*

Then, the following statements are true regarding the BOOST procedure

- *The algorithm runs in time polynomial in $1/\varepsilon, |\mathcal{A}|$, and $\log(1/\delta)$.*
- *In total, it uses $\widetilde{O}(\log(|\mathcal{A}|/\delta)/\varepsilon^4)$ many samples.*

[2]The question of how to draw valid conclusions when the questions asked depend on the data is the subject of an entire field called adaptive data analysis. See https://adaptivedataanalysis.com/ for a class on the topic.

- *With probability $1-\delta$ over the random samples drawn across all rounds, BOOST returns a predictor $\tilde{p}$ that is 2ε-OI with respect to $\mathcal{A}$. Furthermore, if the tests $A \in \mathcal{A}$ can be computed by circuits of size s, the circuit computing $\tilde{p}$ consists of a circuit of size $\mathcal{O}(s/\varepsilon^2)$.*

Proof. The first statement follows from the fact that the runtime of the algorithm is bounded by the number of BOOST iterations which is at most $1/\varepsilon^2$ times the runtime per iteration which requires one pass through S for every $A \in \mathcal{A}$ where S is $\widetilde{\mathcal{O}}(\log(1/\delta)/\varepsilon^2)$ per round.[3]

To show the second statement, we know that at every iteration, the AUDIT procedures succeeds with probability $1-\delta$ given $\mathcal{O}(\log(|\mathcal{A}|/\delta)/\varepsilon^2)$ many samples. If we set $\delta' = \delta/T$ and take a union bound over all $T = 1/\varepsilon^2$ many iterations, we get that with probability $1-\delta'$ it succeeds at all T iterations if we set $n \geq \Omega((\log(|\mathcal{A}|) + \log(1/\delta) + \log(1/\varepsilon))/\varepsilon^2)$. To get the total sample complexity, we multiply this quantity by the total number of rounds which in turn yields the $\widetilde{\mathcal{O}}(1/\varepsilon^4)$ sample complexity.

The last statement follows from the fact that by construction, the final predictor consists of at most $1/\varepsilon^2$ copies of circuits $A \in \mathcal{A}$ stitched together. □

3 Outcome Indistinguishability and Multicalibration

In this section, we provide yet another perspective on outcome indistinguishability showing how it is intimately connected to notions of calibration.

Assume again for the sake of simplicity, that outcomes y are binary. A predictor $\tilde{p} : \mathcal{X} \to [0,1]$ is calibrated if the predictions "mean what they say".

Definition 3.1 (Calibration)**.** *Fix a distribution $\mathcal{D}$. A predictor $p : \mathcal{X} \to [0,1]$ is perfectly calibrated if for all $v \in [0,1]$,*

$$\mathbf{E}[y|p(x) = v] = v.$$

That is, if we look at all of the data points s where $\tilde{p}(x) = v$ the outcome y happened a v-fraction of the time. It's helpful to draw a picture. See Figure 1. Note that if $(x,y) \sim \mathcal{D}$, then the Bayes predictor $p^*(y) = \mathbf{Pr}_{\mathcal{D}}[y = 1|\mathcal{X} = x]$ is perfectly calibrated. However, the constant predictor that states $\tilde{p}(x) = \mathbf{E}[y]$ is *also* calibrated.

While this notion is standard, working with conditional expectations is quite cumbersome. Operationally, it's often better to work with the following notion of weighted calibration.

Definition 3.2 (Weighted Calibration)**.** *Fix a distribution $\mathcal{D}$ and a class of weight functions $\mathcal{W} \subseteq \{w : [0,1] \to [-1,1]\}$. A predictor $p : \mathcal{X} \to [0,1]$ is $(\mathcal{W}, \alpha)$-calibrated if for all $w \in \mathcal{W}$,*

$$\mathbf{E}[w(p(x)) \cdot (y - p(x))] \leq \alpha.$$

Having seen how the constant predictor is calibrated, it's clear that one clump together many different types of data points into coarse predicted buckets while preseriving calibration. In fact,

[3]Here, we are not counting the time it takes to evaluate the individual functions $A \in \mathcal{A}$, or rather, treating the evaluation of each individual function as a constant.

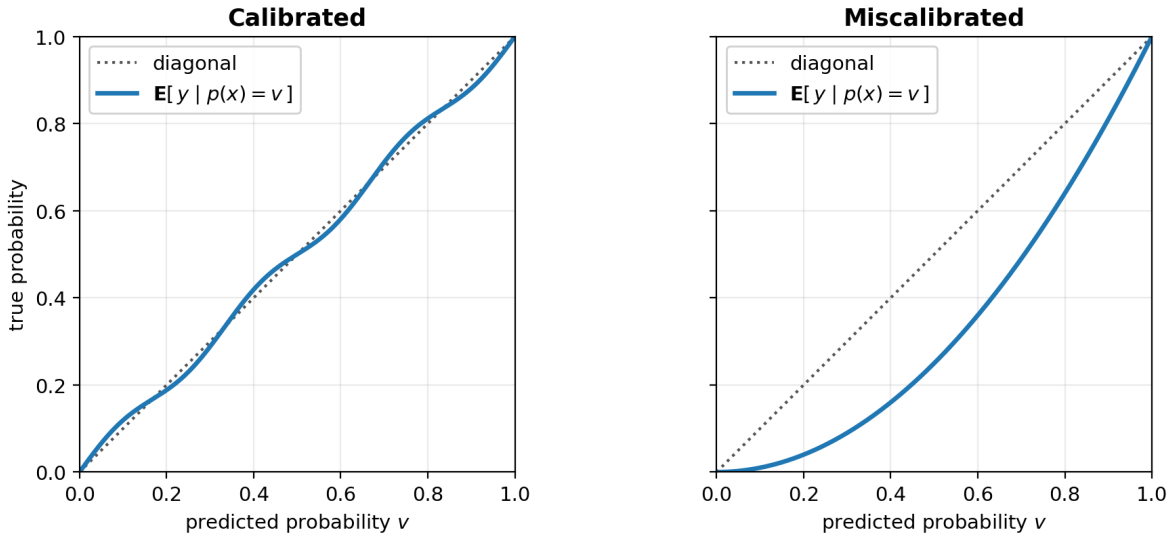

Figure 1: Two calibration curves plotting the true probability $\mathbf{E}[y \mid p(x) = v]$ against the predicted probability v, with the diagonal dotted. Left: a predictor that is close to being calibrated, whose curve lies (approximately) on the diagonal. Right: a miscalibrated predictor, whose curve is far from the diagonal — for instance, among the points receiving prediction $v = 0.6$, the outcome occurs only 36% of the time.

calibration does not rule out *algorithmic stereotyping*, where a significant group of individuals $S \subseteq \mathcal{X}$ is treated as a monolith, receiving the same prediction $p(x) = \mathbf{E}[y|x \in S]$, for all $x \in S$. Even if the individuals in the group have significant variance in $p^*(x)$, by clumping them into a single prediction bucket, a calibrated predictor can ignore the variation. See the example in Figure 2

Still, the ability to stereotype is not a problem that arise due to calibration, but due to the fact that the constraints only bind at the level sets $\{x : \tilde{p}(x) = v\}$. To address this problem, we can consider how to strengthen calibration by considering richer sets.

At the extreme, we might imagine a notion of *individual* calibration, where we condition on the individual's features x.

$$\mathbf{E}[y|p(x) = v, x] \approx v$$

Such a condition, however, is equivalent to requiring the predictor to equal the optimal predictor p^*, which typically is information-theoretically (not to mention computationally) infeasible.

The computationally-identifiable masses. Intuitively, algorithmic stereotyping is possible, when the predictor is not required to distinguish between qualified (i.e., high-risk) individuals within a group and unqualified. If we could reason about groups defined by p^*, for instance $S_{>1/2} = \{x \in \mathcal{X} : p^*(x) > 1/2\}$, then we could hope to prevent stereotyping. Of couse, we don't have access to p^*, which is the problem.

One way to identify groups of "qualified" individuals is through classification, using a concept

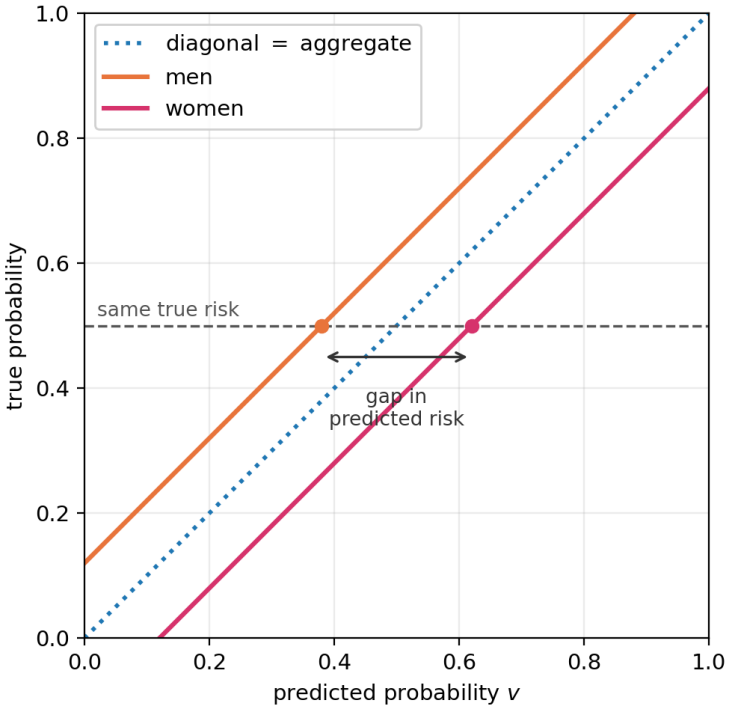

Figure 2: Declumping a calibrated predictor into two subpopulations. Among men, the true probability lies above the diagonal at every predicted value; among women, it lies below; the two group lines average out to the diagonal, so the predictor is calibrated over the population as a whole while being miscalibrated on each group. The horizontal line fixes a single level of true risk: men and women at identical true risk 0.5 receive very different predictions (0.38 versus 0.62). Any ranking or thresholding based on predicted risk therefore places women above equally risky men.

class $\mathcal{C}$. You should think of $\mathcal{C}$ as a collection of binary functions c where each function identifies a particular subgroup of the population. For instance, we might have one such c for every demographic subgroup. These groups can be arbitrarily overlapping or intersecting. See Figure 3.

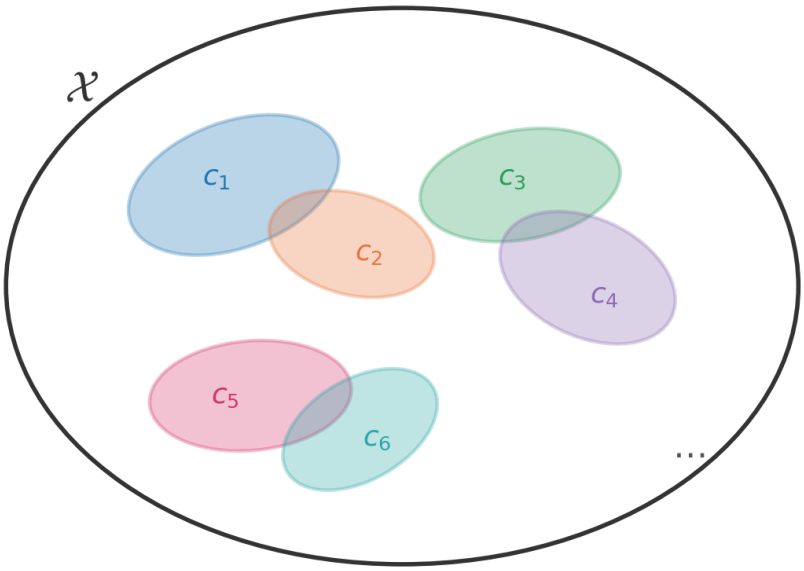

Figure 3: The domain $\mathcal{X}$ together with the subpopulations defined by a concept class $\mathcal{C}$. Each concept $c \in \mathcal{C}$ identifies a subgroup $\{x \in \mathcal{X} : c(x) = 1\}$, and the groups may arbitrarily overlap and intersect.

The benefit of introducing groups is that, as in OI, we can now interpolate between individual notions of calibration where $\tilde{p}(x) = p^*(x)$ for every x and population notions where $\tilde{p}(x) = \mathbf{E}[y]$ by requiring that

$$\mathbf{E}[y|\tilde{p}(x) = v, c(x) = 1] = v \text{ for all} v \in \text{supp}(\tilde{p}) \text{ and } c \in \mathcal{C}$$

This is the exact ($\varepsilon = 0$) version of the following definition, which weights each constraint by the mass of the group and level set.

Definition 3.3 (Multicalibration [HKRR18])**.** *Fix a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$, a concept class $\mathcal{C} \subseteq \{c : \mathcal{X} \to \{0,1\}\}$, and an error tolerance $\varepsilon \geq 0$. A predictor $\tilde{p} : \mathcal{X} \to [0,1]$ is $(\mathcal{C}, \varepsilon)$-multicalibrated if for every $c \in \mathcal{C}$ and $v \in \text{supp}(\tilde{p})$,*

$$\left| \mathop{\mathbf{E}}_{(x,y)\sim\mathcal{D}} [(y - v) \cdot c(x) \cdot \mathbf{1}\{\tilde{p}(x) = v\}] \right| \leq \varepsilon.$$

Dividing through by $\mathbf{Pr}[\tilde{p}(x) = v, c(x) = 1]$ recovers the conditional statement above. Indeed, $c(x) \cdot \mathbf{1}\{\tilde{p}(x) = v\}$ is the indicator of the event $\{\tilde{p}(x) = v, c(x) = 1\}$, so

$$\mathbf{E}\left[(y - v) \cdot c(x) \cdot \mathbf{1}\{\tilde{p}(x) = v\}\right] = (\mathbf{E}[y|\tilde{p}(x) = v, c(x) = 1] - v) \cdot \mathbf{Pr}[\tilde{p}(x) = v, c(x) = 1].$$

The definition therefore asks that $|\mathbf{E}[y|\tilde{p}(x) = v, c(x) = 1] - v| \leq \varepsilon$ / $\mathbf{Pr}[\tilde{p}(x) = v, c(x) = 1]$: groups and level sets that carry more mass receive tighter conditional guarantees.

Multicalibration and OI in the following sense formally equivalent. For $c \in \mathcal{C}$ and $v \in \text{supp}(\tilde{p})$, define the test

$$A_{c,v}(x, p, y) = c(x) \cdot \mathbf{1}\{p = v\} \cdot y.$$

Each test has discrete derivative $\Delta A_{c,v}(x, p) = c(x) \cdot \mathbf{1}\{p = v\}$, so by Proposition 2.2, $\tilde{p}$ is ε-OI with respect to $\mathcal{A}_{\mathcal{C}} = \{A_{c,v}\}$ if and only if for every $c \in \mathcal{C}$ and $v \in \text{supp}(\tilde{p})$,

$$\left| \mathop{\mathbf{E}}_{x \sim \mathcal{D}_x} [c(x) \cdot \mathbf{1}\{\tilde{p}(x) = v\} \cdot (\tilde{p}(x) - p^*(x))] \right| \leq \varepsilon.$$

Therefore, we can use the algorithmic machinery we've developed so far to learn multicalibrated predictors.

References

[DKR+21] 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 (STOC)*, pages 1095–1108, 2021.

[GHK+23] Parikshit Gopalan, Lunjia Hu, Michael P. Kim, Omer Reingold, and Udi Wieder. Loss minimization through the lens of outcome indistinguishability. In *14th Innovations in Theoretical Computer Science Conference (ITCS)*, 2023.

[Hau92] David Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992.

[HKRR18] Úrsula 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 (ICML)*, pages 1939–1948, 2018.

[KSS94] Michael J. Kearns, Robert E. Schapire, and Linda M. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2–3):115–141, 1994.

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