Making Decisions under Outcome Performativity

Michael P. Kim
University of California, Berkeley
mpkim@berkeley.edu

Juan C. Perdomo
University of California, Berkeley
jcperdomo@berkeley.edu

Abstract

Decision-makers often act in response to data-driven predictions, with the goal of achieving favorable outcomes. In such settings, predictions don't passively forecast the future; instead, predictions actively shape the distribution of outcomes they are meant to predict. This *performative prediction* setting [PZMH20] raises new challenges for learning "optimal" decision rules. In particular, existing solution concepts do not address the apparent tension between the goals of *forecasting* outcomes accurately and *steering* individuals to achieve desirable outcomes.

To contend with this concern, we introduce a new optimality concept—*performative omniprediction*—adapted from the supervised (non-performative) learning setting [GKR+22]. A performative omnipredictor is a single predictor that simultaneously encodes the optimal decision rule with respect to many possibly-competing objectives. Our main result demonstrates that efficient performative omnipredictors exist, under a natural restriction of performative prediction, which we call *outcome performativity*. On a technical level, our results follow by carefully generalizing the notion of outcome indistinguishability [DKR+21, GHK+23] to the outcome performative setting. From an appropriate notion of Performative OI, we recover many consequences known to hold in the supervised setting, such as omniprediction and universal adaptability [KKG+22].

Authors listed alphabetically.

1 Introduction

Data-driven predictions inform policy decisions that directly impact individuals. Proponents argue that by understanding patterns from the past, decisions can be optimized to improve future outcomes, to the benefit of individuals and institutions [KLMO15]. In the US educational system, for instance, early warning systems (EWS) have become a key tool used by states to combat low graduation rates [BB19, US 16]. The rationale for using such systems is clear. Given a predictor that, for each student, estimates the likelihood of graduation, school districts can identify high-risk students at a young age, directing resources to improve individuals' outcomes, and in turn, the districts' graduation rates. Despite compelling arguments, reliably predicting life outcomes remains a largely-unsolved problem in machine learning.

A key challenge in utilizing predictions to inform decisions is that, often, predictions influence the outcomes they're meant to forecast. In the education example above, districts consider predictions of graduation with the *intention* of effecting graduation outcomes. In this situation—where predictions determine interventions, which influence outcomes—accuracy can be a paradoxical notion. If a predictor correctly identifies high risk individuals as likely to suffer negative outcomes, after successful interventions, the individuals' outcomes will be positive and the initial predictions will appear inaccurate. To apply data-driven tools effectively, decision-makers must resolve an apparent tension between the objectives of *forecasting* individuals' outcomes reliably and *steering* individuals to achieve better outcomes.

Recent work of [PZMH20] introduced *performative prediction* to contend with the fact that predictions not only forecast, but also shape the world. Informally, a prediction problem is performative if the act of prediction influences the distribution on individual-outcome pairs. From early warning systems, to online content recommendations, to public health advisories: across many contexts, individuals respond to predictions in a manner that changes the likelihood of possible outcomes (successful graduation, increased click rate, or decreased disease caseload).

In their original work on the subject, [PZMH20] frame the goal of performative prediction through loss minimization. In this framing, the ultimate goal is to learn a *performatively optimal* decision rule. A decision rule h_{po} is performatively optimal if it achieves the minimal expected loss (within some class of decision rules $\mathcal{H}$) over the distribution that it induces,

$$h_{\text{po}} \in \arg\min_{h \in \mathcal{H}} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}(h)}[\ell(x, h(x), y)]. \tag{1}$$

Here, $\mathcal{D}(h)$ is the distribution over (x, y) pairs observed as response to deploying h.

For generality's sake, performative prediction makes minimal restrictions on how the distribution may respond to a chosen decision rule. In particular, the choice to deploy a hypothesis h, may change the joint distribution $(x, y) \sim \mathcal{D}(h)$ over individual-outcome pairs, essentially arbitrarily.[1] This generality enables us to write a broad range of prediction problems—including supervised learning [SB14], strategic classification [HMPW16], and causal inference [MMH20]—as special cases of performative prediction. In all, [PZMH20] establishes a powerful framework for reasoning about settings where the distribution of examples responds to the predictions.

While powerful, the framework has two noticeable limitations. First, achieving performative optimality is hard. Without any assumptions on the distributional response $\mathcal{D}(\cdot)$, achieving performative optimality requires exhaustive search over the hypothesis class $\mathcal{H}$. Furthermore, even under strong structural assumptions on the distributional response and choice of loss ℓ, it

[1] [PZMH20] assume only a Lipschitzness condition, where similar hypotheses h and h' give rise to similar distributions $\mathcal{D}(h)$ and $\mathcal{D}(h')$, measured in Wasserstein (earth mover's) distance.

is known that convex optimization does not suffice to achieve optimality [PZMH20, MPZ21]. Stated another way: the generality of performative prediction does not come for free. To date, all existing methods for performative optimality require strong specification assumptions on the outcome distribution and distributional response.

The second limitation arises due to formulating performative prediction as a loss minimization problem: the loss ℓ is fixed, once and for all. In performative prediction, different losses can encode drastically different objectives: losses are used not only to promote accuracy of predictions, but also to encourage favorable outcome distributions. Consider a loss designed for accurate forecasting, e.g., the squared error $(\widehat{y} - y)^2$. In this case, the optimal decision rule will prioritize accuracy without regard for the "quality" of the outcome distribution. On the other hand, consider a loss designed to steer towards positive outcomes, $1 - y$. Here, there is no notion of accuracy (the loss ignores the prediction $\widehat{y}$), but instead, the objective is to nudge the distribution of outcomes towards $y = 1$.

Encoding the decision-making objective through a single loss function forces the learner to choose the "correct" objective at train time. Downstream decision-makers, however, may reasonably want to explore different objectives according to their own sense of "optimality". In the existing formulations for performative prediction, exploring different losses requires re-training from scratch. In this work, we investigate an alternative formulation that enables decision-makers to efficiently explore optimal decision rules under many different objectives.

1.1 Decision-Making under Outcome Performativity

To begin, we introduce a special case of the performative prediction setting, which we call *outcome performativity*. Outcome performativity focuses on the effects of local decisions on individuals' outcomes, rather than the effect of broader policy on the distribution of individuals. For instance, our example of graduation prediction is modeled well by outcome performativity. For a given a student, the EWS prediction they receive affects their future graduation outcome, but does not influence their demographic features or historical test scores. In other words, we narrow our attention to the performative effects of decisions $h(x)$ on the conditional distribution over outcomes y, rather than the effects of the decision rule h on the distribution as a whole $\mathcal{D}(h)$. This reframing of performativity still captures many important decision-making problems, but gives us additional structure to address some of the limitations in the original formulation.

On a technical level, outcome performativity imagines a data generating process over triples $(x, \widehat{y}, y^*)$ where $x \sim \mathcal{D}$ is sampled from a *static* distribution over inputs, then a prediction or decision $\widehat{y} \in \widehat{\mathcal{Y}}$ is selected (possibly as a function of x), and finally the true outcome $y^* \in \mathcal{Y}$ is sampled conditioned on x and $\widehat{y}$. We focus on binary outcomes $\mathcal{Y} = \{0, 1\}$.[2] In this setting, the outcome performativity assumption posits the existence of an underlying probability function,

$$p^* : \mathcal{X} \times \widehat{\mathcal{Y}} \to [0, 1],$$

where for a given individual $x \in \mathcal{X}$ and decision $\widehat{y} \in \widehat{\mathcal{Y}}$, the true outcome y^* is sampled as a Bernoulli with parameter $p^*(x, \widehat{y})$. We refer to the true outcome distribution p^* as *Nature*.

By asserting a fixed "ground truth" probability function, the outcome performativity framework does not allow for arbitrary distributional responses and limits the generality of the approach. For instance, outcome performativity does not capture strategic classification. But

[2] In general, outcome performativity could be defined for larger outcome domains. Handling such domains is possible, but technical. We restrict our attention to binary outcomes to focus on the novel conceptual issues.

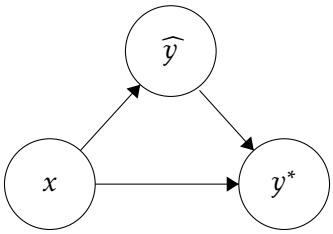

Figure 1: Causal graphical representation of the outcome performativity data generating process.

importantly, by refining the model of performativity, there is hope that we may sidestep the hardness results for learning optimal performative predictors.

Performative Omniprediction. We begin by observing that under outcome performativity, the true probability function p^* suggests an optimal decision rule $f_\ell^* : \mathcal{X} \to \widehat{\mathcal{Y}}$ for any loss ℓ. In our setting, p^* governs the outcome distribution, so given an input $x \in \mathcal{X}$, the optimal decision $f_\ell^*(x)$ is determined by a simple, univariate optimization procedure over a discrete set $\widehat{\mathcal{Y}}$:

$$f_\ell^*(x) \in \arg\min_{\widehat{y} \in \widehat{\mathcal{Y}}} \mathop{\mathbb{E}}_{y^* \sim p^*(x,\widehat{y})} [\ell(x, \widehat{y}, y^*)]. \tag{2}$$

Note that the decision rule $f_\ell^*(x)$ minimizes the loss pointwise for $x \in \mathcal{X}$. Consequently, averaging over any static, feature distribution $\mathcal{D}$, the decision rule f_ℓ^* is performative optimal for *any* hypothesis class $\mathcal{H}$, loss ℓ, and marginal distribution $\mathcal{D}$:

$$\mathop{\mathbb{E}}_{\substack{x \sim \mathcal{D} \\ y^* \sim p^*(x, f_\ell^*(x))}} [\ell(x, f_\ell^*(x), y^*)] \leqslant \min_{h \in \mathcal{H}} \mathop{\mathbb{E}}_{\substack{x \sim \mathcal{D} \\ y^* \sim p^*(x, h(x))}} [\ell(x, h(x), y^*)].$$

While the existence of p^* implies the existence of optimal decision rules under outcome performativity, we make no assumptions about the learnability of p^*. In general, the function p^* may be arbitrarily complex, so learning (or even representing!) p^* may be infeasible, both computationally and statistically. Still, the above analysis reveals the power of modeling the probability function $p^* : \mathcal{X} \times \widehat{\mathcal{Y}} \to [0, 1]$. The optimal probability function p^* encodes the optimal decision rule f_ℓ^* *for every loss function* ℓ. This perspective raises a concrete technical question: short of learning p^*, can we learn a probability function $\tilde{p} : \mathcal{X} \times \widehat{\mathcal{Y}} \to [0, 1]$ that suggests an optimal decision rule, via simple post-processing, for many different objectives?

Recent work of [GKR+22] studied the analogous question in the context of supervised learning (without performativity), formalizing a solution concept which they call *omniprediction*. Intuitively, an omnipredictor is a single probability function $\tilde{p}$ that suggests an optimal decision rule for many different loss functions $\mathcal{L}$. The work of [GKR+22] and follow-up work of [GHK+23] demonstrate—rather surprisingly—that omniprediction in supervised learning is broadly a feasible concept. For a variety of choices of loss classes $\mathcal{L}$ (e.g., Lipschitz losses or convex losses), it is possible to learn an efficient predictor $\tilde{p}$ that gives optimal decisions for any loss $\ell \in \mathcal{L}$.

In this work, we generalize omniprediction to the outcome performative setting. As a solution concept, *performative omniprediction* directly addresses the limiting assumption in performative prediction that the loss ℓ is known and fixed. Given a performative omnipredictor, a decision-maker can explore the consequences of optimizing for different losses, balancing the

desire for forecasting and steering, as they see fit. Technically, given a predictor $\tilde{p}$, we define $\tilde{f}_\ell : \mathcal{X} \to \widehat{\mathcal{Y}}$ to be the optimal decision rule, that acts as if outcomes are governed by $\tilde{p}$.

$$\tilde{f}_\ell(x) \in \arg\min_{\widehat{y} \in \widehat{\mathcal{Y}}} \mathop{\mathbb{E}}_{\tilde{y} \sim \tilde{p}(x, \widehat{y})} [\ell(x, \widehat{y}, \tilde{y})]$$

We emphasize that, for any loss ℓ, the decision rule $\tilde{f}_\ell(x)$ is an efficient post-processing of the predictions given by $\tilde{p}(x, \widehat{y})$ for $\widehat{y} \in \widehat{\mathcal{Y}}$. A performative omnipredictor is a model of nature $\tilde{p} : \mathcal{X} \times \widehat{\mathcal{Y}} \to [0, 1]$ that induces a corresponding decision rule $\tilde{f}_\ell$ that is performatively optimal over a collection of losses $\ell \in \mathcal{L}$.

Definition (Performative Omnipredictor)**.** *For a collection of loss functions $\mathcal{L}$, hypothesis class $\mathcal{H}$, and $\varepsilon \geqslant 0$, a predictor $\tilde{p} : \mathcal{X} \times \widehat{\mathcal{Y}} \to [0, 1]$ is an $(\mathcal{L}, \mathcal{H}, \varepsilon)$-performative omnipredictor for an input distribution $\mathcal{D}$ if for every $\ell \in \mathcal{L}$, the decision rule $\tilde{f}_\ell$ is ε-performative optimal over $\mathcal{H}$.*

$$\mathop{\mathbb{E}}_{\substack{x \sim \mathcal{D} \\ y^* \sim p^*(x, \tilde{f}_\ell(x)))}} [\ell(x, \tilde{f}_\ell(x)), y^*)] \leqslant \arg\min_{h \in \mathcal{H}} \mathop{\mathbb{E}}_{\substack{x \sim \mathcal{D} \\ y^* \sim p^*(x, h(x))}} [\ell(x, h(x), y^*)] + \varepsilon \tag{3}$$

While an intriguing prospect, omniprediction is particularly ambitious in the performative world. Whereas most supervised learning losses have the same moral goal (to accurately forecast the outcome), losses in the performative world can encode entirely contradictory objectives. For instance, we can define a pair of losses ℓ_0 and ℓ_1 that reward decisions that steer outcomes to be 0 and 1, respectively. A performative omnipredictor must contend with these contradictions, providing optimal decision rules under performative effects.

Concretely, under outcome performativity, there is a certain circularity in naively determining the optimal decision $\tilde{f}(x)$ from a prediction $\tilde{p}(x, \widehat{y})$. Choosing an "optimal" decision $\tilde{f}(x)$ causes a shift in the distribution on the outcome $y^* \sim p^*(x, \tilde{f}(x))$, which may imply a different "optimal" decision, which seems to lead to a continuing cycle of dependency. In this way, any performative omnipredictor $\tilde{p}$ must encode the optimal decision rule $\tilde{f}_\ell$ for each $\ell \in \mathcal{L}$, *anticipating the shift* induced by the choice of $\tilde{f}_\ell$. In this work, we ask whether—despite this key challenge—efficient performative omnipredictors exist, and if so, can we learn them?

1.2 Our Contributions

Our first contributions are conceptual, introducing the outcome performativity setting and the notion of performative omnipredictors. As an abstraction, outcome performativity strikes a balance with enough generality to model many real-world phenomena and enough structure to give effective solutions. For settings where the distributional response occurs predominantly as outcome performativity, the framework is well-scoped to contend with the challenges of performative prediction. In particular, performative omnipredictors provide an effective solution concept to address the tension between different objectives under performativity.

With these conceptual contributions in place, we turn to the feasibility of omniprediction under outcome performativity. While outcome performativity introduces a number of new challenges, we show how to apply many techniques established for omniprediction in the supervised learning setting to recover analogous guarantees under performativity. On a technical level, we follow the *Loss Outcome Indistinguishability* approach of [GHK+23], demonstrating how—with the right conceptual framing—arguments for the existence of supervised omnipredictors can be translated into guarantees for the performative setting.

Efficient Performative Omnipredictors Exist. Our first technical contribution demonstrates existence of efficient performative omnipredictors. We prove that for any class of losses $\mathcal{L}$ and any hypothesis class $\mathcal{H}$, there exists a performative omnipredictor $\tilde{p}$ of complexity that scales polynomially with the complexity of computing the losses and hypotheses.

Theorem 1. *Suppose $\mathcal{D}$ is a fixed distribution over $\mathcal{X}$. Let $\mathcal{L} \subseteq \{\ell : \mathcal{X} \times \widehat{\mathcal{Y}} \times \mathcal{Y} \to [0,1]\}$ be a set of bounded loss functions, and let $\mathcal{H} \subseteq \{h : \mathcal{X} \to \widehat{\mathcal{Y}}\}$ be a hypothesis class of decision rules. If the functions in $\mathcal{L}$ and $\mathcal{H}$ can be computed by circuits of size s, then there exists a $(\mathcal{L}, \mathcal{H}, \varepsilon)$-performative omnipredictor of circuit complexity* $\mathrm{poly}(s, |\widehat{\mathcal{Y}}|)/\varepsilon^2$.

Importantly, this result holds *for any* class of bounded losses. The collection $\mathcal{L}$ may include losses for forecasting and steering, or may include losses that steer towards different outcomes. Still, the predictor $\tilde{p}$ will encode a performative optimal decision rule for each such loss $\ell \in \mathcal{L}$. Furthermore, the complexity of this predictor scales gracefully with the complexity of the losses and hypotheses and the available decisions, *independent of the complexity of Nature p^**. Even if the true probability function p^* is intractably-complex, there exists a simple function $\tilde{p}$ that mimics the omniprediction behavior, provided the losses and hypotheses are sufficiently simple.

Learning Performative Omnipredictors Reduces to Supervised Learning. In fact, the proof of existence is constructive. We establish the feasibility of performative omnipredictions by devising a boosting-style learning algorithm, inspired by the original algorithm for learning (non-performative) omnipredictors [HKRR18, DKR+21, GKR+22, GHK+23]. As in the supervised case, we show that learning omnipredictors reduces to an *auditing* task. Despite the fact that in performative prediction, different decision rules induce different distributions, we show that given appropriately randomized data, this auditing task can be solved using only *supervised learning* primitives implementable in finite samples. That is, under outcome performativity, there is a surprising reduction from the task of learning optimal performative predictors to the task of non-performative supervised learning.

Formally, we assume that the learner has access to a collection of data triples $(x, \widehat{y}, y) \sim \mathcal{D}_{\mathrm{rct}}$ where inputs are sampled from the data distribution $x \sim \mathcal{D}$, decisions $\widehat{y}$ are assigned uniformly at random, and the outcome $y^* \sim p^*(x, \widehat{y})$ is sampled from Nature, for the given individual and randomly-assigned decision. Given an efficiently bounded number of samples access from this distribution, we show how to learn performative omnipredictors assuming access to a supervised learner for the hypothesis class $\mathcal{H}$. We formalize this learning assumption in terms of cost-sensitive classification [Elk01].

Theorem 2 (Informal)**.** *Assume sample access to $\mathcal{D}_{\mathrm{rct}}$ and suppose that $\mathcal{A}$ is a cost-sensitive learning algorithm for the hypothesis class $\mathcal{H}$. There is a polynomial-time algorithm, that, for any set of bounded losses $\mathcal{L}$, returns a $(\mathcal{L}, \mathcal{H}, \varepsilon)$-performative omnipredictor using at most* $\mathrm{poly}(1/\varepsilon, |\widehat{\mathcal{Y}}|, \log|\mathcal{H}|, \log|\mathcal{L}|)$ *many samples from $\mathcal{D}_{\mathrm{rct}}$ while also making $|\mathcal{L}| \cdot \mathrm{poly}(1/\varepsilon, |\widehat{\mathcal{Y}}|)$ oracle calls to $\mathcal{A}$.*

The guarantees of this algorithm represent a significant point of departure from previous work on performative prediction. Specifically, previous algorithms for learning performatively optimal models (for a single loss) hinged on the condition that predictions had very mild, and highly-structured (e.g. linear) impact on the induced data distributions as in [MPZ21, JZM22]. Conversely, within the outcome performativity restriction, we make no assumptions on the way predictions influence outcomes. Further, the omnipredictor output in the guarantee of Theorem 2 has complexity scaling as stated in Theorem 1. In other words, our learning

algorithm makes no "realizability" assumptions and outputs an efficient predictor, regardless of the complexity of Nature.

Universally-Adaptable Omnipredictors. Outcome performativity focuses attention on performative shifts in the outcome distribution as a function of the chosen decision $\widehat{y} \in \widehat{\mathcal{Y}}$. In particular, it excludes performative effects in the distribution over individuals $\mathcal{X}$. Despite this limitation, our final result shows that we can learn performative omnipredictors that are robust to *exogenous* (non-performative) shifts in the distribution over individuals.

Adapting the notion of universal adaptability, introduced in the context of statistical estimation by [KKG+22], we show how to learn *univerally-adaptable* performative omnipredictors. Whereas performative omnipredictors guarantee optimality on a fixed marginal distribution $\mathcal{D}$ over individuals, universally-adaptable omnipredictors give the same optimality guarantee, simultaneously, over a rich class of input distribution shifts $\mathcal{D}_{\mathcal{W}}$. Each distribution in $\mathcal{D}_\omega \in \mathcal{D}_{\mathcal{W}}$ corresponds to the reweighting of probabilities in $\mathcal{D}$ by some importance weight function ω in some pre-specified class $\mathcal{W}$.

Theorem 3 (Informal)**.** *Let $\mathcal{L}$ be a set of bounded loss functions, $\mathcal{H}$ a hypothesis class of decision rules, and let $\mathcal{W} \subseteq \{\omega : \mathcal{X} \to [0, \omega_{\max}]\}$. If the functions in $\mathcal{L}$, $\mathcal{H}$, and $\mathcal{W}$ can be computed by circuits of size s, then there exists a predictor $\tilde{p}$, computable by a circuit of size at most $\mathrm{poly}(s, |\widehat{\mathcal{Y}}|) \cdot \omega_{\max}^2/\varepsilon^2$, that is a $(\mathcal{L}, \mathcal{H}, \varepsilon)$-performative omnipredictor for every distribution over individuals $\mathcal{D}_\omega \in \mathcal{D}_{\mathcal{W}}$.*

The result follows by augmenting the class of loss functions $\mathcal{L}_{\mathcal{W}}$ to account for shifts under $\mathcal{W}$, and again, applying the constructive learning algorithm from Theorem 2. We emphasize that the learner only needs to account for the class of shifts at *training* time. At evaluation time, the decision-maker need not know anything about the underlying distribution over individuals. Indeed, the decision-maker can simply use $\tilde{p}$ as before, post-processing to decisions $\tilde{f}_\ell(x)$ for any $\ell \in \mathcal{L}$ on an input-by-input basis.

1.3 Our Techniques: Performative Outcome Indistinguishability

We begin our technical overview with a simple motivating example. Consider the following performative prediction problem.

Example. Let individuals and decisions be encoded as signed booleans $\mathcal{X} = \{\pm 1\}$ and $\widehat{\mathcal{Y}} = \{\pm 1\}$, assuming $\mathcal{D}$ is uniform over $\mathcal{X}$. Suppose that Nature's outcome distribution over $\mathcal{Y} = \{0, 1\}$ is governed by the conditional probability function

$$p^*(x, \widehat{y}) = 1/2 + \beta x \widehat{y}$$

for any $0 < \beta < 1/2$. Consider the goal of learning an $(\mathcal{L}, \mathcal{H})$-performative omnipredictor for the following collection of losses and hypotheses:

- $\mathcal{L} = \{\ell_0, \ell_1\}$ contains two opposing steering losses, which steer outcomes towards 0 and 1, respectively.

$$\ell_1(x, \widehat{y}, y^*) = 1 - y^* \qquad\qquad \ell_0(x, \widehat{y}, y^*) = y^*$$

- $\mathcal{H} = \{h_+, h_-\}$ contains two decision rules over $\mathcal{X}$ that either returns x or its negation.

$$h_+(x) = x \qquad\qquad h_-(x) = -x$$

We begin by considering some naive attempts to achieve the goal of performative omniprediction. Note that $\mathcal{H}$ actually contains an optimal decision rule for each loss in $\mathcal{L}$. In particular, for losses that steer to 0 versus 1, the optimal decisions minimize (or maximize) the probability that $y^* = 1$. The decision rule h_+ maximizes the probability $p^*(x, h_+(x)) = 1/2 + \beta$ for all x, whereas h_- minimizes the probability $p^*(x, h_-(x)) = 1/2 - \beta$. To obtain the omniprediction guarantee, then, we must learn a probability function that encodes the best decision under ℓ_0 and ℓ_1.

As such, a natural approach would be to fit a function $p : \mathcal{X} \times \widehat{\mathcal{Y}} \to [0,1]$ that approximates the underlying probability p^*, which we can then post-process for a loss ℓ as in Equation 2 . To fit p, we can simply do supervised learning directly over triples $(x, \widehat{y}, y^*)$, where $\widehat{y}$ is chosen uniformly at random. We argue that without specification (realizability) assumptions, this approach also fails. Consider, for instance, fitting p using logistic regression

$$p(x, \widehat{y}) = \frac{1}{1 + \exp(-(ax + b\widehat{y} + c))},$$

where a, b, c are parameters of the model. In our example, when we select $\widehat{y}$ at random, the outcome y^* is uncorrelated with each of x and $\widehat{y}$ on their own. Thus, the optimal setting of these parameters is $a = b = c = 0$.[3] Consequently, the logistic model is a constant: $p(x, \widehat{y}) = 1/2$ for all $x \in \mathcal{X}$ and $\widehat{y} \in \widehat{\mathcal{Y}}$. Such constant predictions are completely uninformative. Clearly, they cannot suggest the optimal decision rule for any loss, let alone every loss in our collection. The negative result here, follows because the model class for p was misspecified to fit p^*. In this example, of course, we simply need to run regression with quadratic terms to be well-specified. However, without any assumptions about the complexity of p^*, we cannot rely on approaches that require specifying Nature's model exactly, which might have unbounded complexity.

Performative Outcome Indistinguishability. Recently, [DKR+21] introduced the notion of Outcome Indistinguishability (OI) as a new solution concept for supervised learning. In contrast to the traditional framing of learning through loss minimization, OI defines the goal of learning through the lens of indistinguishability. In this view, a predictive model should provide outcomes that cannot be distinguished from true outcomes from Nature. In the world of supervised learning, OI and the closely-related notion of multicalibration [HKRR18] have seen broad application, including in deriving supervised omnipredictors [GKR+22, GHK+23].

Towards our goal of performative omniprediction, we adapt the paradigm of learning via outcome indistinguishability to the outcome performative setting. In particular, we leverage the variant of outcome indistinguishability explored in a recent work of [GHK+23]. In the supervised learning setting, [GHK+23] builds a set of indistinguishability conditions called Loss OI, which they show implies omniprediction. We show that these conditions and the argument are even more general than originally conceived, and can be applied to the outcome performative setting.

Intuitively, we say a predictor $\tilde{p} : \mathcal{X} \times \widehat{\mathcal{Y}} \to [0,1]$ is *performative outcome indistinguishable* if outcomes drawn according to the model $\tilde{y} \sim \tilde{p}(x, h(x))$ are indistinguishable from Nature's outcomes $y^* \sim p^*(x, h(x))$ *under the distribution induced by a decision rule* h. To make this notion precise, we need to specify what we mean by indistinguishability and pin down the decision rules we care to reason about.

[3] More concretely, since $\mathbb{E}[x\widehat{y}] = \mathbb{E}[xy^*] = 0$, one can check that $a = b = c = 0$ solves the first-order optimality conditions for the logistic regression objective $\mathbb{E}_{x,\widehat{y},y^*} -[y \log \sigma(ax + b\widehat{y} + c) + (1 - y) \log(1 - \sigma(ax + b\widehat{y} + c))]$ for $\sigma(z) = 1/(1 + \exp(-z))$.

To encode omniprediction through performative OI, our goal will be to devise a set of tests of a predictor $\tilde{p}$ that—if passed—guarantee for every loss $\ell \in \mathcal{L}$, the decision rule $\tilde{f}_\ell$ is as good as any $h \in \mathcal{H}$. Formally, we start by building a class of tests from a collection of losses $\mathcal{L}$ and a hypothesis class $\mathcal{H}$.[4]

Definition (Performative OI)**.** *For an input distribution $\mathcal{D}$, collection of losses $\mathcal{L}$, hypothesis class $\mathcal{H}$, and $\varepsilon \geqslant 0$, a predictor $\tilde{p} : \mathcal{X} \times \widetilde{\mathcal{Y}} \to [0,1]$ is $(\mathcal{L}, \mathcal{H}, \varepsilon)$-performative outcome indistinguishable (POI) over $\mathcal{D}$ if for all $\ell \in \mathcal{L}$ and all $h \in \mathcal{H}$,*

$$\mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}\\ y^*\sim p^*(x,h(x))}}[\ell(x,h(x),y^*)] \approx_\varepsilon \mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}\\ \tilde{y}\sim\tilde{p}(x,h(x))}}[\ell(x,h(x),\tilde{y})].$$

We emphasize how this performative OI condition is a natural notion of *outcome* indistinguishability. In particular, we note *where* the predictor $\tilde{p}$ occurs in the POI conditions: it is only used to sample outcomes according to $\tilde{p}$ in the modeled world. Importantly, the decisions $h(x)$ are used in the sampling of $\tilde{y} \sim \tilde{p}(x, h(x))$. Using $h(x)$ as the decision associated with x to sample each outcome y (under Nature and the model) ensures that $\tilde{p}$ encodes a reliable estimate of the loss ℓ if the decision rule h is deployed. Performative OI ensures that $\tilde{p}$ "knows" these values for each loss in the collection $\ell \in \mathcal{L}$ and each hypothesis in our class $h \in \mathcal{H}$.

While this OI condition ensures that $\tilde{p}$ captures the behavior of the hypotheses in $\mathcal{H}$, it says nothing about the losses under its own decision rules $\tilde{f}_\ell$. Reasoning about these decision rules is essential for performative omniprediction. As such, we introduce an additional OI condition, which we call performative *decision* OI, to ensure OI under the decision rules suggested by $\tilde{p}$.

Definition (Performative Decision OI)**.** *For an input distribution $\mathcal{D}$, collection of loss functions $\mathcal{L}$, and $\varepsilon \geqslant 0$, a predictor $\tilde{p} : \mathcal{X} \times \widetilde{\mathcal{Y}} \to [0,1]$ is $(\mathcal{L}, \varepsilon)$-performative decision outcome indistinguishable (DOI) over $\mathcal{D}$ if for all $\ell \in \mathcal{L}$,*

$$\mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}\\ y^*\sim p^*(x,\tilde{f}_\ell(x))}}[\ell(x,\tilde{f}_\ell(x),y^*)] \approx_\varepsilon \mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}\\ \tilde{y}\sim\tilde{p}(x,\tilde{f}_\ell(x))}}[\ell(x,\tilde{f}_\ell(x),\tilde{y})].$$

Here, $\tilde{p}$ is still used to sample outcomes $\tilde{y}$, but is also used to determine the decision rule $\tilde{f}_\ell$, for each $\ell \in \mathcal{L}$. Syntactically, changing from $h \in \mathcal{H}$ to $\tilde{f}_\ell$ is a small change, but it has significant impacts on the nature of the performative DOI condition—both in terms of its costs and the strength of its guarantees. Critically, for a given loss ℓ, the decision rule $\tilde{f}_\ell$ is (by definition) optimal for outcomes sampled from $\tilde{y} \sim \tilde{p}(x, \tilde{f}_\ell(x))$. As such, indistinguishability is a powerful tool here: if the losses are indistinguishable on modeled outcomes (where $\tilde{f}_\ell$ is optimal) and on Nature's outcomes, then $\tilde{f}_\ell$ should be optimal for Nature.

We formalize this intuition, demonstrating that, as in the supervised setting, performative OI and decision OI suffice to establish omniprediction. Consider the loss $\ell \in \mathcal{L}$ obtained by any $h \in \mathcal{H}$ on true outcomes. We show that the loss of $\tilde{f}_\ell$ is upper bounded by that of h.

Proposition (Informal)**.** *If a predictor $\tilde{p}$ is $(\mathcal{L}, \mathcal{H})$-POI and $\mathcal{L}$-DOI, then $\tilde{p}$ is an $(\mathcal{L}, \mathcal{H})$-performative omnipredictor.*

[4] Throughout, we use the notational shorthand $A \approx_\varepsilon B$ to denote that $A \in [B - \varepsilon, B + \varepsilon]$.

Proof sketch. Once the appropriate OI conditions are written down, deriving performative omniprediction is almost immediate.

$$\begin{aligned}\mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}\\ y^*\sim p^*(x,\tilde{f}_\ell(x))}}[\ell(x,\tilde{f}_\ell(x),y^*)] &\approx \mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}\\ \tilde{y}\sim\tilde{p}(x,\tilde{f}_\ell(x))}}[\ell(x,\tilde{f}_\ell(x),\tilde{y})]\\ &\leqslant \mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}\\ \tilde{y}\sim\tilde{p}(x,h(x))}}[\ell(x,h(x),\tilde{y})] \approx \mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}\\ y^*\sim p^*(x,h(x))}}[\ell(x,h(x),y^*)]\end{aligned}$$

The first equality follows by $\mathcal{L}$-DOI, the second equality follows by $(\mathcal{L},\mathcal{H})$-POI, and the middle inequality follows by the fact that $\tilde{f}_\ell$ is optimal over modeled outcomes. ∎

In other words, we have managed to reduce the task of learning performative omnipredictors to learning models satisfying performative OI conditions. Clearly, Nature's model p^* is "indistinguishable" from Nature (similarly, it is clear that p^* is a performative omnipredictor), but the question remains whether there exist *efficient* predictors $\tilde{p}$ that satisfy the performative OI conditions. Thus, we turn our attention to learning OI predictors under outcome performativity.

Learning Outcome Performative Predictors. Despite essential differences in the notions of OI in the supervised and performative settings, we show that many of the algorithmic techniques that have become standard in the literature on multicalibration and OI can be adapted to work in the outcome performative setting. In particular, we demonstrate that, quite generically, learning performative OI models reduces to *auditing* for distinguishability. Concretely, if there exists a loss $\ell\in\mathcal{L}$ and hypothesis $h\in\mathcal{H}\cup\{\tilde{f}_\ell\}$, such that the performative OI conditions are violated for a predictor $\tilde{p}$, we can use these "distinguishers" to update the model to address the violation. This observation immediately suggests using a boosting algorithm, in the vein of [HKRR18], to learn performative OI predictors. Provided that updating based on an ε-violation makes significant "progress" towards satisfying performative OI, then the number of auditing steps $T\leqslant O(1/\varepsilon^2)$ will be bounded.

While this learning paradigm of "audit, then update" is intuitive, there are nontrivial challenges in maintaining the efficiency of the learned predictors. Consider, for instance, the performative decision OI constraint. As highlighted above, the DOI constraints require that we reason about the optimal decision rule according to $\tilde{p}$. In particular, to update based on a violation of DOI on loss ℓ, we need to incorporate a copy of the function $\tilde{f}_\ell(\cdot)$. This decision rule, however, is a function of $\tilde{p}(\cdot,\widehat{y})$ for every $\widehat{y}\in\widehat{\mathcal{Y}}$ (as it requires computing the argmin over $\widehat{\mathcal{Y}}$). Naively, then, it would seem that in every iteration where we update the model based on some $\tilde{f}_\ell$, we need to make $|\widehat{\mathcal{Y}}|$ recursive oracle calls to the existing model. Without careful consideration, the [HKRR18]-style learning algorithm will build a performative OI predictor $\tilde{p}$ whose complexity s scales *exponentially* in the number of iterations, $s\geqslant|\widehat{\mathcal{Y}}|^T$.

To avoid this blow-up, we need to choose a more effective representation of the probability function $\tilde{p}(\cdot,\cdot)$. We observe that, in general, the updates required for the algorithm are *sparse* in $\widehat{\mathcal{Y}}$. As a result of this sparsity, we can save on overall computation by implementing $\tilde{p}:\mathcal{X}\times\widehat{\mathcal{Y}}\to[0,1]$ as a map from individuals to vectors of probabilities $\tilde{q}:\mathcal{X}\to[0,1]^{\widehat{\mathcal{Y}}}$. Mathematically, there is a bijection between such functions; computationally, however, the representations behave very differently. By increasing the amount of work per update by a factor of $|\widehat{\mathcal{Y}}|$, we avoid making $|\widehat{\mathcal{Y}}|$ recursive calls. This strategy is reminiscent of an approach [DKR+22] used to learn (supervised) OI predictors for outcomes living in a large domain. With this representation in place, an

appropriate analysis reveals that the resulting predictors can be implemented in complexity that scales only *polynomially* in the number of iterations and decisions.

After addressing representation issues, we study sufficient conditions to implement the auditing task efficiently, from a polynomially bounded number of samples. Achieving performative optimality, in general, requires exploration of the consequence of using different decision rules $h \in \mathcal{H}$, as these decision rules effect the distribution on outcomes. Nevertheless, we show that it suffices for this exploration to be done "offline" via randomized assignment of decisions $\widehat{y} \in \widehat{\mathcal{Y}}$. In particular, if we collect triples $\{(x, \widehat{y}, y^*)\}$ through a randomized control trial, assigning $\widehat{y}$ uniformly at random for each $x \in \mathcal{X}$, and observing $y^* \sim p^*(x, \widehat{y})$, we avoid the need to deploy each $h \in \mathcal{H}$.

Given access to such RCT data, we give a reduction from the task of auditing for performative OI to the task of *supervised learning* for the hypothesis class $\mathcal{H}$. This reduction from auditing to learning is familiar in the OI framework [HKRR18, DKR+21], but critically, we go from a performative prediction task to a non-performative task. In all, our reductions show that if we can learn the best decision rule from $\mathcal{H}$ in a supervised learning setting, then we can learn performative omnipredictors with respect to $\mathcal{H}$, assuming access to appropriately sampled data.

Universal Adaptability under Outcome Performativity. The OI viewpoint enables a similarly straightforward analysis of distributional robustness, via universal adaptability. Universal adaptability is a notion introduced by [KKG+22] in the context of statistical estimation. In the original context, a predictor is universally adaptable if it provides an efficient way to estimate statistics across many underlying input distributions.

We translate the notion of universal adaptability to the outcome performative prediction setting. In our context, we parameterize universal adaptability by a class of importance weight functions $\mathcal{W} \subseteq \{\mathcal{X} \to \mathbb{R}_{\geqslant 0}\}$. For a base input distribution $\mathcal{D}$, we define a corresponding collection of shifted distributions $\mathcal{D}_{\mathcal{W}}$ to be the set of distributions reachable after reweighting by some $\omega \in \mathcal{W}$.

$$\mathcal{D}_{\mathcal{W}} = \{\mathcal{D}_\omega : \omega \in \mathcal{W}, \mathrm{supp}(\mathcal{D}_\omega) \subseteq \mathrm{supp}(\mathcal{D})\} \text{ and } \forall x \in \mathrm{supp}(\mathcal{D}_\omega) : \quad \mathcal{D}_\omega(x) = \omega(x) \cdot \mathcal{D}(x)$$

The key observation is that for any hypothesis h, loss function ℓ, and importance weight function ω, the expected loss over $\mathcal{D}_\omega$ is equal to an expected loss over $\mathcal{D}$, for a loss defined in terms of ℓ and ω.

$$\mathop{\mathbb{E}}_{\substack{x \sim \mathcal{D}_\omega \\ \tilde{y} \sim \tilde{p}(x, h(x))}} [\ell(x, h(x), \tilde{y})] = \mathop{\mathbb{E}}_{\substack{x \sim \mathcal{D} \\ \tilde{y} \sim \tilde{p}(x, h(x))}} [\ell(x, h(x), \tilde{y}) \cdot \omega(x)]$$

Importantly, this equality relies on the fact that the outcome probability functions p^* and $\tilde{p}$ are defined conditional on x and $\widehat{y}$, and thus are invariant across shifts in the input distribution.

Using the result that indistinguishability implies omniprediction, by simply enforcing that $\tilde{p}$ satisfy the POI and DOI conditions relative to p^* under this enriched class of loss functions $\ell \cdot \omega$, we can neatly ensure that $\tilde{p}$ is again POI and DOI, not just over $\mathcal{D}$, but over every marginal distribution $\mathcal{D}_\omega$. Consequently, $\tilde{p}$ must a be an omnipredictor under all of these $\mathcal{D}_\omega$. The simplicity of this analysis attests to the versatility of the OI perspective and the value it provides in domains beyond supervised learning.

1.4 Related Work and Discussion

Our work lies at the intersection of several areas including performative prediction, the outcome indistinguishability and multicalibration literature, as well as other fields studying algorithmic decision-making such as contextual bandits. We briefly discuss how our results relate to previous work within these areas and conclude with some speculation regarding broader implications of our conclusions.

Performative Prediction. The performative prediction framework was introduced by [PZMH20] who defined the main solution concepts and analyzed the convergence of repeated risk minimization to *performatively stable* points. A decision rule h_{ps} is performatively stable if it is a fixed point of risk minimization,

$$h_{\text{ps}} \in \arg\min_{h \in \mathcal{H}} \mathop{\mathbb{E}}_{(x,y^*)\sim\mathcal{D}(h_{\text{ps}})}[\ell(x, h(x), y^*)].$$

Subsequent work by [MPZH20, DX22, BHK22, CDH21] studied stochastic optimization algorithms for finding stable points in a variety of settings. However, these stable solutions need not be performatively *optimal* as per the definition outlined in Equation 1. In fact, [MPZ21] proved that performatively stable models can achieve *arbitrarily worse* loss than performative optimal points. This observation motivated the design of algorithms for findings performatively optimal decision rules for a fixed loss ℓ [MPZ21, IYZ21, IZY22, NFD+22, JZM22]. These algorithms work in the general performative prediction setup where the model h can affect the *joint* distribution over pairs (x, y), but make make very strong specification assumptions on how each h influences the distribution, and restrict to loss functions satisfying smoothness and strong convexity.

In contrast, our results rely only on the outcome performativity assumption and mild boundedness assumptions. The recent work of [MDW22] has also considered the outcome performativity setting, aiming to understand when performative effects are identifiable from observational data. Short of identifiability, it remains an interesting direction for future research to give learning algorithms for performative omnipredictors from observational data.

Beyond these optimization results, previous work in performative prediction has acknowledged the tension in performative prediction between accurate forecasting and steering. Concretely, [MPZ21] discuss how the choice of loss function in performative prediction should balance predictive accuracy with any externalities that arise from the impacts of prediction on the observed distribution. In a different direction, [HJM22] uses performativity as a lens with which to study notions of market power in economics. As part of their analysis, they provide a decomposition of the performative risk of a classifier into terms that represent forecasting and steering. While we consider how the choice of loss function determines the high-level objective, [HJM22] considers how, even for a fixed loss function, the performative risk can be decomposed into terms associated with forecasting and steering.

Reinforcement Learning and Contextual Bandits. As discussed in [PZMH20], performative prediction, and in particular outcome performativity, can be cast as reinforcement learning (RL) or contextual bandits problems. Individuals x correspond to the contexts, decisions $\widehat{y}$ correspond to actions, and the loss $\ell(x, \widehat{y}, y^*)$ is captured by the reward $r(x, \widehat{y})$. Due to the breadth of their definitions, most ML problems can be written as RL problems.

Still, important issues that arise in outcome performativity—like the tension between forecasting and steering and the desire for omnipredictors—are best seen by focusing on the

specific interactions between predictions $\widehat{y}$ and outcomes y. The variety of losses that can exist for a given outcome are obscured by encapsulating all feedback within an abstract reward function $r(x, \widehat{y})$. Moreover, on a technical level, performativity has a richer feedback structure that can be used to design more efficient algorithms as illustrated by [JZM22].

Multicalibration and Outcome Indistinguishability. Originally developed by [HKRR18] as a notion of fairness in prediction, multicalibration has seen considerable interest and application in the broader context of supervised learning. At a high-level, multicalibration requires predictions to be *calibrated*, not just overall, but even when restricting our attention to structured subpopulations. The goal of multicalibration and other related notions of "multi-group" fairness [KNRW18, KRR18, KGZ19, JLP+21] is to ensure that learning occurs within important subpopulations that might otherwise be ignored.

Intuitively, the requirements of multicalibration represent a kind of indistinguishability: calibration requires that the predicted probabilities "look like" real probabilities. [DKR+21] formalizes this intuition, introducing the notion of Outcome Indistinguishability, which generalizes multicalibration. They show tight computational equivalences between multi-group fairness notions and variants of OI. Subsequently, OI and multicalibration have been applied in diverse contexts beyond fairness, such as distributional robustness through universal adaptability [KKG+22] and omniprediction [GKR+22].

Omniprediction. Our work draws directly from the work on omnipredictors [GKR+22, GHK+23]. Even in the supervised learning setting, the existence of efficient omnipredictors is not at all obvious. The main result of [GKR+22] demonstrates the feasibility of omnipredictors over any hypothesis class $\mathcal{H}$, for the class $\mathcal{L}_{\text{cvx}}$ of all convex and Lipschitz loss functions. This sweeping result follows by showing that a $\mathcal{H}$-multicalibrated predictor is a $(\mathcal{L}_{\text{cvx}}, \mathcal{H})$-omnipredictor.

A key follow-up work of [GHK+23] shows that omniprediction is even more general than initially thought. This work initiates the study of omniprediction through the lens of outcome indistinguishability. By the equivalence of OI with multicalibration, it has been clear since the work of [GKR+22] that OI captures loss minimization and omniprediction in the context of supervised learning, albeit indirectly. [GHK+23] revisits the question of omnipredictors, directly through the lens of OI, studying a refined notion, which they call Loss OI. They derive a general recipe for omnipredictors for any class of losses $\mathcal{L}$, from calibration and multiaccuracy over a class derived from $\mathcal{L}$ and $\mathcal{H}$.

On a technical level, our analysis follows the OI-based approach of [GHK+23]. Indeed, our proof that Performative OI and Performative Decision OI imply Performative Omniprediction follows the strategy laid out to obtain supervised omnipredictors from Loss OI. Despite the fact that outcome performativity requires us to reason about distributional shifts induced by the predictions, the same indistinguishability framework proves effective for learning omnipredictors in our setting.

While syntactically similar to prior formulations of OI, our notion of performative OI is the first to consider outcome indistinguishability for non-supervised learning distributions. Our objective, in this work, was to derive a notion of performative OI sufficient to imply performative omniprediction. No doubt, further generalizations of the original OI hierarchy [DKR+21] to the (outcome) performative setting—and beyond—may prove useful.

Organization

The remainder of the manuscript is organized as follows. In Section 2, we define performative omniprediction and performative outcome indistinguishability. Here, we show how appropriate performative OI conditions suffice to obtain omniprediction. In Section 3, we establish the universal adaptability properties of performative omnipredictors. In Section 4, we give a generic learning algorithm for performative omnipredictors, demonstrating concrete instantiations of the algorithm using randomized control trial data. Finally, in Section 5, we discuss notions of multicalibration in the context of performative prediction. We speculate that some notions translate to the performative setting naturally, yielding efficient approaches to performative OI, while other notions seem to resist efficient translation.

Notation Overview

We denote individual's features by $x \in \mathcal{X}$ and the available decisions by $\widehat{y} \in \widehat{\mathcal{Y}}$. We assume that $\mathcal{X}$ and $\widehat{\mathcal{Y}}$ are discrete sets, and assume $|\widehat{\mathcal{Y}}|$ is finite. Throughout our presentation, we assume that outcomes $\mathcal{Y} = \{0, 1\}$ are binary. The true conditional expectation over Nature's outcomes is given by

$$p^* : \mathcal{X} \times \widehat{\mathcal{Y}} \to [0, 1],$$

where for every individual $x \in \mathcal{X}$ and decision $\widehat{y} \in \widehat{\mathcal{Y}}$, $p^*(x, \widehat{y})$ gives the true probability of positive outcome.

$$p^*(x, \widehat{y}) = \mathbb{P}\mathrm{r}[y^* = 1 \mid x, \widehat{y}]$$

In analogy to p^*, we denote the learner's predictor (i.e., the "model of Nature") by

$$\tilde{p} : \mathcal{X} \times \widehat{\mathcal{Y}} \to [0, 1].$$

We use the shorthand $y^* \sim p^*(x, \widehat{y})$ to denote true outcomes drawn from the Bernoulli distribution with parameter $p^*(x, \widehat{y})$, and $\tilde{y} \sim \tilde{p}(x, \widehat{y})$ to represent modeled outcomes drawn from the Bernoulli distribution with parameter $\tilde{p}(x, \widehat{y})$. Throughout, we use $\mathcal{D}$ to denote a marginal distribution over individual features $x \in \mathcal{X}$. To denote the approximate equality of expectations, we use the notational shorthand $A \approx_\varepsilon B$ to denote that $A \in [B - \varepsilon, B + \varepsilon]$.

We take a loss function to be a map from individual-decision-outcome triples to a nonnegative value,

$$\ell : \mathcal{X} \times \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}_{\geqslant 0}.$$

For a collection of losses $\mathcal{L}$, we let $\ell_{\max} = \sup_{x, \widehat{y}, y} \ell(x, \widehat{y}, y)$. For a fixed loss ℓ, we define $\tilde{f}_\ell : \mathcal{X} \to \widehat{\mathcal{Y}}$ to be the optimal post-processing of $\tilde{p}$ according to ℓ. More specifically, for every $x \in \mathcal{X}$,

$$\tilde{f}_\ell(x) = \arg\min_{\widehat{y} \in \widehat{\mathcal{Y}}} \mathop{\mathbb{E}}_{\tilde{y} \sim p(x, \widehat{y})} [\ell(x, \widehat{y}, \tilde{y})].$$

Note that $\tilde{f}_\ell$ is defined pointwise, without regard to the marginal distribution over $\mathcal{X}$. We use $\mathcal{H} \subseteq \{h : \mathcal{X} \to \widehat{\mathcal{Y}}\}$ to represent a set of decision rules (i.e., a hypothesis class) which map individuals to decisions. Typically, we assume that hypotheses $h \in \mathcal{H}$ have an efficient representation.

2 Performative Omniprediction

Introduced in the supervised learning setting by [GKR+22], an omnipredictor is an outcome probability model that suggests an optimal decision rule for any loss within a class of loss functions. To discuss this concept more formally, we first review the definition optimality with respect to a fixed loss function, under our setting of outcome performativity.

Definition 2.1 (Performative Optimality)**.** *For input distribution $\mathcal{D}$, loss function ℓ, hypothesis class $\mathcal{H}$, and $\varepsilon \geqslant 0$, a decision rule $f : \mathcal{X} \to \widehat{\mathcal{Y}}$ is $(\ell, \mathcal{H}, \varepsilon)$-performatively optimal over $\mathcal{D}$ if*

$$\mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}\\ y^*\sim p^*(x,f(x))}}[\ell(x, f(x), y^*)] \leqslant \min_{h\in\mathcal{H}} \mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}\\ y^*\sim p^*(x,h(x))}}[\ell(x, h(x), y^*)] + \varepsilon.$$

That is, a decision rule f is performative optimal if f obtains expected loss under the induced outcome distribution (i.e. performative risk) competitive with the best hypothesis in some reference class $h \in \mathcal{H}$. [5] A simple, but key observation is that if we knew the model that generates outcomes for each individual perfectly, then performative optimality is straightforward to achieve. Given an outcome model p^* that specifies, for a given individual $x \in \mathcal{X}$ and decision $\widehat{y} \in \widehat{\mathcal{Y}}$, the probability of outcome $y \sim p^*(x, \widehat{y})$, the optimal decision rule for any loss function can be determined by a simple optimization over the choice of decisions $\widehat{y}$, as follows.

Fact 2.2 (Optimal Decision Rule)**.** *Fix an outcome model $p^* : \mathcal{X}\times\widehat{\mathcal{Y}} \to [0, 1]$. For a loss $\ell : \mathcal{X}\times\widehat{\mathcal{Y}}\times\mathcal{Y} \to \mathbb{R}_{\geqslant 0}$, there exists a globally optimal decision rule $f^*_\ell : \mathcal{X} \to \widehat{\mathcal{Y}}$ for ℓ under p^*, given by the pointwise loss-minimizer,*

$$f^*_\ell(x) = \arg\min_{\widehat{y}\in\widehat{\mathcal{Y}}} \mathop{\mathbb{E}}_{y^*\sim p^*(x,\widehat{y})}[\ell(x, \widehat{y}, y^*)]. \tag{4}$$

For instance, when $\widehat{\mathcal{Y}}$ is discrete and finite, then the optimal prediction can be computed by linearly enumerating over $\widehat{\mathcal{Y}}$, and computing a finite sum of two terms each time. We will often consider the optimal decision rule $\tilde{f}_\ell$ after post-processing an approximate outcome model, or predictor, denoted as $\tilde{p}$.

With these definitions of performative optimality and optimal decision rule in place, we are ready to define performative omniprediction. An omnipredictor is an outcome model $\tilde{p}$ where for *every* loss $\ell \in \mathcal{L}$ within some collection, the optimal decision rule derived from $\tilde{p}$ is performatively optimal with respect to some class of hypothesis $\mathcal{H}$.

Definition 2.3 (Performative Omniprediction)**.** *For input distribution $\mathcal{D}$, collection of loss functions $\mathcal{L}$, hypothesis class $\mathcal{H}$, and $\varepsilon \geqslant 0$, a predictor $\tilde{p} : \mathcal{X} \times \widehat{\mathcal{Y}} \to [0, 1]$ is an $(\mathcal{L}, \mathcal{H}, \varepsilon)$-performative omnipredictor over $\mathcal{D}$ if for all $\ell \in \mathcal{L}$, the optimal decision rule $\tilde{f}_\ell$ is $(\ell, \mathcal{H}, \varepsilon)$-performative optimal.*

Omniprediction is a very strong solution concept. Whereas the optimal decision rule typically depends intimately on the chosen loss, an omnipredictor needs to encode the optimal decision rule for every loss in $\mathcal{L}$, even if these losses encode very different preferences over predictions. It is not hard to see that the optimal predictor is an omnipredictor for any hypothesis and loss class.

Corollary 2.4. *For any input distribution $\mathcal{D}$, collection of loss functions $\mathcal{L}$, and hypothesis class $\mathcal{H}$, the optimal predictor $p^* : \mathcal{X} \times \widehat{\mathcal{Y}} \to [0, 1]$ is an $(\mathcal{L}, \mathcal{H}, 0)$-performative omnipredictor over $\mathcal{D}$.*

[5] To review, as in [PZMH20] we refer to $\mathbb{E}_{x\sim\mathcal{D}, y^*\sim p^*(x,h(x))}[\ell(x, h(x), y^*)]$ as the performative risk of h on ℓ.

This corollary follows directly from Fact 2.2, because the optimal predictor p^* gives the true probability law governing the performative outcome distribution. Still, the optimal predictor may be of arbitrary complexity and is generally inaccessible. The question remains whether *efficient* performative omnipredictors exist, and if so, how to learn them. To attack this question, we introduce a generalization of the outcome indistinguishability framework to the outcome performativity setting.

2.1 Performative Outcome Indistinguishability

Outcome Indistinguishability (OI) was introduced by [DKR+21] as an alternative paradigm for supervised learning. Rather than focusing on loss minimization, OI formalizes learning as a computational indistinguishability condition. In this view, a predictor should produce outcomes that are indistinguishable from Nature's outcome distribution. While OI can encode classic learning goals like loss minimization, the abstraction is quite generic and amenable to modern supervised learning desiderata, like fairness [HKRR18] and distributional robustness [KKG+22].

Here, we propose an indistinguishability definition for the performative world.[6] This definition extends what [GHK+23] refer to as Hypothesis OI in the supervised setting.

Definition 2.5 (Performative OI)**.** *For input distribution $\mathcal{D}$, collection of losses $\mathcal{L}$, hypothesis class $\mathcal{H}$, and $\varepsilon \geqslant 0$, a predictor $\tilde{p} : \mathcal{X} \times \widehat{\mathcal{Y}} \to [0,1]$ is $(\mathcal{L}, \mathcal{H}, \varepsilon)$-performative outcome indistinguishable (POI) over $\mathcal{D}$ if for all $\ell \in \mathcal{L}$ and all $h \in \mathcal{H}$,*

$$\Big| \mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D} \\ y^*\sim p^*(x,h(x))}} [\ell(x,h(x),y^*)] - \mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D} \\ \tilde{y}\sim \tilde{p}(x,h(x))}} [\ell(x,h(x),\tilde{y})] \Big| \leqslant \varepsilon$$

In this definition, we fix our collection of distinguishers to be parameterized by a collection of loss functions and a hypothesis class. The POI condition states that, even when the outcome distribution can depend nontrivially on the hypothesis value $h(x)$, the outcomes y^* and $\tilde{y}$ are indistinguishable, as measured by the expected loss of each hypothesis. Note that the distinguishers take as input the individual x, the decision $h(x)$, and either Nature's outcome y^* or the modeled outcome $\tilde{y}$. In particular, these distinguishers do not receive access to the predictions $\tilde{p}(x,h(x))$ themselves.[7]

As a step towards obtaining omniprediction, we require indistinguishability between $\tilde{p}$ and p^* not just under the reference decision rules h, but also under the optimal decision rules $\tilde{f}_\ell$ derived from $\tilde{p}$. This motivates the notion of Performative Decision OI, which extends the idea of decision calibration, introduced in [ZKS+21], and decision OI, introduced in [GHK+23], to the performative setting.

Definition 2.6 (Performative Decision OI)**.** *For input distribution $\mathcal{D}$, collection of loss functions $\mathcal{L}$, and $\varepsilon \geqslant 0$, a predictor $\tilde{p} : \mathcal{X} \times \widehat{\mathcal{Y}} \to [0,1]$ is $(\mathcal{L}, \varepsilon)$-performative decision outcome indistinguishable (DOI) over $\mathcal{D}$ if for all $\ell \in \mathcal{L}$,*

$$\Big| \mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D} \\ y^*\sim p^*(x,\tilde{f}_\ell(x))}} [\ell(x,\tilde{f}_\ell(x),y^*)] - \mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D} \\ \tilde{y}\sim \tilde{p}(x,\tilde{f}_\ell(x))}} [\ell(x,\tilde{f}_\ell(x),\tilde{y})] \Big| \leqslant \varepsilon.$$

[6] In its original formulation, OI is a hierarchy of related notions. We generalize the framework to our setting, focusing on notions of performative OI that will imply performative omniprediction. Understanding a full generalization of the OI framework to the performative setting is an interesting question for future investigations.

[7] In the language of [DKR+21], this notion corresponds to the "No-Access" level of the OI hierarchy. In principle, we could also extend the upper levels to the performative setting as well. We comment this issue further within our discussion of performative calibration in Section 5.

Operationally, DOI allows us to sample outcomes $\tilde{y} \sim \tilde{p}(x, \tilde{f}_\ell(x))$ from our model of Nature, evaluate the expected loss of $\ell(x, \tilde{f}_\ell(x), \tilde{y})$, and be confident that it is close to the loss on outcomes sampled from Nature $y^* \sim p^*(x, \tilde{f}_\ell(x))$.

Note that, technically, the indistinguishability conditions in Performative OI and Performative Decision OI look the same, but just refer to different hypothesis classes; that is, $(\mathcal{L}, \varepsilon)$-Performative Decision OI can be phrased as $(\mathcal{L}, \{\tilde{f}_\ell : \ell \in \mathcal{L}\}, \varepsilon)$-Performative OI. We make a distinction between these notions because, semantically, the hypothesis class $\{\tilde{f}_\ell : \ell \in \mathcal{L}\}$ is derived from the predictor $\tilde{p}$, whereas $h \in \mathcal{H}$ is independent of $\tilde{p}$. As we will see later, this semantic difference manifests as a concrete difference in the computational complexity of achieving each notion of indistinguishability.

2.2 Performative Omniprediction via OI

With these definitions in place, we can prove our first main result: performative omniprediction from performative outcome indistinguishability. One of the main benefits of studying the problem from the indistinguishability lens is that it enables an especially clean and simple analysis. The proof strategy we employ here follows the proof of omniprediction in the supervised learning world by [GHK+23]. Curiously, the proof only needs one direction of the indistinguishability inequalities.

Theorem 2.7. *Fix an input distribution $\mathcal{D}$, collection of losses $\mathcal{L}$, hypothesis class $\mathcal{H}$, and $\varepsilon \geqslant 0$. Suppose that $\tilde{p} : \mathcal{X} \times \widehat{\mathcal{Y}} \to [0,1]$ is $(\mathcal{L}, \varepsilon)$-performative decision OI and $(\mathcal{L}, \mathcal{H}, \varepsilon)$-performative OI. Then, $\tilde{p}$ is a $(\mathcal{L}, \mathcal{H}, 2\varepsilon)$-performative omnipredictor.*

Proof. The proof exploits the fact that for each loss $\ell \in \mathcal{L}$, $\tilde{f}_\ell$ is the optimal decision rule for ℓ under $\tilde{p}$. Fix a loss $\ell \in \mathcal{L}$. First, we upper bound the loss achieved by $\tilde{f}_\ell$ on real outcomes y^* in terms of the loss on modeled outcomes $\tilde{y}$. Under $(\mathcal{L}, \varepsilon)$-performative decision OI,

$$\mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}\\ y^*\sim p^*(x,\tilde{f}_\ell(x))}}[\ell(x, \tilde{f}_\ell(x), y^*)] \leqslant \mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}\\ \tilde{y}\sim\tilde{p}(x,\tilde{f}_\ell(x))}}[\ell(x, \tilde{f}_\ell(x), \tilde{y})] + \varepsilon.$$

Next, we relate the expected loss achieved by $\tilde{f}_\ell$ on modeled outcomes $\tilde{y} \sim \tilde{p}(x, \tilde{f}_\ell(x))$ versus that of other decision rules h. By its definition, $\tilde{f}_\ell(x)$ is the optimal decision over any $\widehat{y} \in \widehat{\mathcal{Y}}$ for the loss $\ell(x, \widehat{y}, \tilde{y})$ under $\tilde{y} \sim \tilde{p}(x, \widehat{y})$. So, averaging over the distribution on inputs $x \sim \mathcal{D}$, the loss of $\tilde{f}_\ell$ is upper bounded by the loss of any other decision rule h, and in particular those in $\mathcal{H}$:

$$\mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}\\ \tilde{y}\sim\tilde{p}(x,\tilde{f}_\ell(x))}}[\ell(x, \tilde{f}_\ell(x), \tilde{y})] \leqslant \mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}\\ \tilde{y}\sim\tilde{p}(x,h(x))}}[\ell(x, h(x), \tilde{y})].$$

Finally, by $(\mathcal{L}, \mathcal{H}, \varepsilon)$-POI, we upper bound the loss achieved by h on real outcomes y^* by that achieved on modeled outcomes $\tilde{y}$.

$$\mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}\\ \tilde{y}\sim\tilde{p}(x,h(x))}}[\ell(x, h(x), \tilde{y})] \leqslant \mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}\\ y^*\sim p^*(x,h(x))}}[\ell(x, h(x), y^*)] + \varepsilon.$$

Combining these three inequalities,

$$\mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}\\ y^*\sim p^*(x,\tilde{f}_\ell(x))}}[\ell(x, \tilde{f}_\ell(x), y^*)] \leqslant \mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}\\ y^*\sim p^*(x,h(x))}}[\ell(x, h(x), y^*)] + 2\varepsilon,$$

so $\tilde{p}$ is a $(\mathcal{L}, \mathcal{H}, 2\varepsilon)$-performative omnipredictor. ∎

3 Universal Adaptability

In addition to minimizing expected risk, performative omniprediction can also be viewed as a guarantee of robustness. So far, we've seen how a performative omnipredictor induces optimal predictions $\widehat{y}$ even if these predictions lead to endogenous shifts in the distribution over outcomes y^*. In this section, we argue that with little additional work, the OI framework can be adapted to yield performative omnipredictors that are robust to exogenous shifts in the marginal distribution over individuals x.[8]

The results here build on the recent work of [KKG+22], who introduced a notion of *universal adaptability* in the context of statistical inference problems. In our context, universal adaptability may be interpreted as a guarantee that the performative omnipredictor properties hold, not only on the original input distribution $\mathcal{D}$, but also on a broad family of shifts of this input distribution $\mathcal{D}$. In particular, we show that by augmenting the class of loss functions, we can learn an outcome prediction model $\tilde{p}$ that can handle exogenous shifts in the input distribution, while still maintaining performative optimality.

We parameterize universal adaptability by a class of importance weight functions $\mathcal{W} \subseteq \{\mathcal{X} \to \mathbb{R}_{\geqslant 0}\}$. For a base input distribution $\mathcal{D}$, we define a corresponding collection of shifted distributions $\mathcal{D}_{\mathcal{W}}$ to be the set of distributions reachable after reweighting the probabilities in $\mathcal{D}$ by some $\omega \in \mathcal{W}$.

$$\mathcal{D}_{\mathcal{W}} = \{\mathcal{D}_\omega : \omega \in \mathcal{W},\ \mathrm{supp}(\mathcal{D}_\omega) \subseteq \mathrm{supp}(\mathcal{D})\},\ \text{where } \forall x \in \mathrm{supp}(\mathcal{D}_\omega),\ \ \mathcal{D}_\omega(x) = \omega(x) \cdot \mathcal{D}(x)$$

Note that to yield a valid probability distribution $\mathcal{D}_\omega$ it is necessary and sufficient that the importance weight function ω have unit weight over $\mathcal{D}$; that is, for any $\omega \in \mathcal{W}$, $\mathbb{E}_{x\sim\mathcal{D}}[\omega(x)] = 1$.[9] Given an importance weight class $\mathcal{W}$, we say that an omnipredictor is universally adaptable if it is an omnipredictor over any $\mathcal{D}_\omega \in \mathcal{D}_{\mathcal{W}}$.

Definition 3.1 (Universal Adaptability)**.** *For input distribution $\mathcal{D}$, weight class $\mathcal{W}$, collection of losses $\mathcal{L}$, hypothesis class $\mathcal{H}$, and $\varepsilon \geqslant 0$, a performative omnipredictor is $\mathcal{W}$-universally adaptable over $\mathcal{D}$ if $\tilde{p}$ is an $(\mathcal{L}, \mathcal{H}, \varepsilon)$-performative omnipredictor over every $\mathcal{D}_\omega \in \mathcal{D}_{\mathcal{W}}$.*

Note that universal adaptability guarantees robustness under *exogeneous* shifts in the marginal distribution over $\mathcal{X}$, not under *endogeneous* shifts in the input distribution induced by the act of prediction. The distributional robustness is with respect to shifts that are defined in advance, independent of the chosen decision rule. Under the guarantees of universal adaptability, the prevalence of various individuals may vary, but the response of any specific individual x to a prediction $\widehat{y}$, as measured by the distribution p^* governing the outcome y^*, remains the same. Intuitively, this type of robustness is the best that we can hope for without explicitly modeling how the predictions $\widehat{y} \in \widehat{\mathcal{Y}}$ change the distribution over individuals $x \in \mathcal{X}$, which $\tilde{p}$ does not model. If predictions $\widehat{y}$ affect both x and y^*, it is not at all obvious to us what invariant property of Nature we should choose to model. We believe these are important questions for future work.

One consequence of this adaptability definition is that any model $\tilde{p}$ that is an omnipredictor for a class of distributions $\mathcal{D}_{\mathcal{W}}$ must also be an omnipredictor for any mixture distribution with

[8] By endogenous we mean that the distribution shift is caused by the act of prediction itself, which is considered in the outcome performativity framework. Exogeneous shifts are not influence by predictions. They refer to changes in the data distribution caused by factors like a change in external environment, or the passage of time.

[9] Other properties of $\mathcal{W}$ will affect whether universal adaptability is feasible, but not its definition. We discuss these issues further in Section 4.

components drawn from this class. We say that a distribution $\mathcal{D}_m$ is a mixture distribution if for all $x \in \mathcal{X}$,

$$\Pr_{\mathcal{D}_m}[X = x] = \sum_{\omega} \lambda_\omega \Pr_{\mathcal{D}_\omega}[X = x] \text{ where } \mathcal{D}_\omega \in \mathcal{D}_{\mathcal{W}},\ \lambda_\omega \geqslant 0 \text{ for all } \omega \text{ and } \sum_{\omega} \lambda_\omega = 1.$$

We denote by $\mathsf{mixt}(\mathcal{D}_{\mathcal{W}})$ the set of all such mixture distributions $\mathcal{D}_m$.

Proposition 3.2. *Let $\mathcal{D}_{\mathcal{W}}$ be a set of distributions over $\mathcal{X}$. If $\tilde{p}$ is a $(\mathcal{L}, \mathcal{H}, \varepsilon)$-performative omnipredictor over every $\mathcal{D}_\omega \in \mathcal{D}_{\mathcal{W}}$, then it is also a $(\mathcal{L}, \mathcal{H}, \varepsilon)$-performative omnipredictor over every $\mathcal{D}_m$ in $\mathsf{mixt}(\mathcal{D}_{\mathcal{W}})$.*

Proof. Fix a loss $\ell \in \mathcal{L}$, a hypothesis $h \in \mathcal{H}$ and a distribution $\mathcal{D}_m$ in $\mathsf{mixt}(\mathcal{D}_{\mathcal{W}})$. Then,

$$\begin{aligned}
\mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}_m \\ y^*\sim p^*(x,\tilde{f}_\ell(x))}} [\ell(x, \tilde{f}_\ell(x), y^*)] &= \sum_{\omega} \lambda_\omega \cdot \mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}_i \\ y^*\sim p^*(x,\tilde{f}_\ell(x))}} [\ell(x, \tilde{f}_\ell(x), y^*)] \\
&\leqslant \sum_{\omega} \lambda_\omega \cdot \mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}_\omega \\ y^*\sim p^*(x,h(x))}} [\ell(x, h(x), y^*)] + \sum_{\omega} \lambda_\omega \varepsilon \\
&= \mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}_m \\ y^*\sim p^*(x,h(x))}} [\ell(x, h(x), y^*)] + \varepsilon
\end{aligned}$$

The first line follows by expanding the definition of the mixture distribution and the second by the omniprediction guarantee on mixture components. In the last line we again applied the definition of a mixture and the fact that the λ_ω sum to 1. Because the inequalities hold for every $h \in \mathcal{H}$, it must be the case that $\tilde{p}$ is an $(\mathcal{L}, \mathcal{H}, \varepsilon)$-omnipredictor for every mixture distribution . ■

We establish universal adaptability for performative omnipredictors by augmenting the loss class $\mathcal{L}$ using the weight class $\mathcal{W}$. Specifically, we define the augmented loss class $\mathcal{L}_{\mathcal{W}}$ as the class of losses $\ell \in \mathcal{L}$ reweighted by importance weight functions $\omega \in \mathcal{W}$.

$$\mathcal{L}_{\mathcal{W}} = \{\ell_\omega : \ell \in \mathcal{L},\ \omega \in \mathcal{W}\}$$
$$\text{where } \forall x \in \mathcal{X}, \widehat{y} \in \widehat{\mathcal{Y}}, y \in \mathcal{Y}: \quad \ell_\omega(x, \widehat{y}, y) = \omega(x) \cdot \ell(x, \widehat{y}, y)$$

With this class of losses in place, we argue that universally-adaptable performative omniprediction is, again, a consequence of performative outcome indistinguishability.

Proposition 3.3. *For a base input distribution $\mathcal{D}$, weight class $\mathcal{W}$, collection of losses $\mathcal{L}$, hypothesis class $\mathcal{H}$, and $\varepsilon \geqslant 0$, if a predictor $\tilde{p}$ is $(\mathcal{L}_{\mathcal{W}}, \mathcal{H}, \varepsilon)$-performative OI and $(\mathcal{L}_{\mathcal{W}}, \varepsilon)$-performative decision OI over $\mathcal{D}$, then $\tilde{p}$ is an $(\mathcal{L}, \mathcal{H}, 2\varepsilon)$-performative omnipredictor that is $\mathcal{W}$-universally adaptable over $\mathcal{D}$.*

Proof. The proposition follows as a corollary of Theorem 2.7. The key observation is that multiplying by the importance weight $\omega(x)$ allows us to switch from an expectation over $\mathcal{D}$ to an expectation over $\mathcal{D}_\omega$. By the definition of $\mathcal{D}_\omega$, we have that for supported $x \in \mathcal{X}$, ω is the odds ratio,

$$\omega(x) = \frac{\mathcal{D}_\omega(x)}{\mathcal{D}(x)}.$$

Further, by the definition of ℓ_ω, for any $h : \mathcal{X} \to \widehat{\mathcal{Y}}$ and any outcome probability model p, the following equality of expectations holds

$$\mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D} \\ y\sim p(x,h(x))}} [\ell_\omega(x, h(x), y)] = \mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D} \\ y\sim p(x,h(x))}} [\omega(x) \cdot \ell(x, h(x), y)] = \mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}_\omega \\ y\sim p(x,h(x))}} [\ell(x, h(x), y)], \tag{5}$$

where we rely on the identity that for any function $g : \mathcal{X} \to \mathbb{R}$,

$$\mathop{\mathbb{E}}_{\mathcal{D}}[g(x) \cdot \omega(x)] = \mathop{\mathbb{E}}_{\mathcal{D}}[g(x) \cdot \mathcal{D}_\omega(x)/\mathcal{D}(x)] = \mathop{\mathbb{E}}_{\mathcal{D}_\omega}[g(x)].$$

The equality in Equation 5 immediately implies that if $\tilde{p}$ is $(\mathcal{L}_\mathcal{W}, \mathcal{H}, \varepsilon)$-POI over $\mathcal{D}$, then $\tilde{p}$ is $(\mathcal{L}, \mathcal{H}, \varepsilon)$-POI over every $\mathcal{D}_\omega \in \mathcal{D}_\mathcal{W}$. That is, by applying the identity to the expectation under Nature's outcomes $y^* \sim p^*(x, h(x))$ and separately to the expectation under the modeled outcomes $\tilde{y} \sim \tilde{p}(x, h(x))$, $(\mathcal{L}_\mathcal{W}, \mathcal{H}, \varepsilon)$-performative OI imples that we obtain indistinguishability for all $\ell \in \mathcal{L}, h \in \mathcal{H}$ and $\mathcal{D}_\omega \in \mathcal{D}_\mathcal{W}$:

$$\left| \mathop{\mathbb{E}}_{\substack{x \sim \mathcal{D}_\omega \\ y^* \sim p^*(x, h(x))}} [\ell(x, h(x), y^*)] - \mathop{\mathbb{E}}_{\substack{x \sim \mathcal{D}_\omega \\ \tilde{y} \sim \tilde{p}(x, h(x))}} [\ell(x, h(x), \tilde{y})] \right| \leqslant \varepsilon.$$

The corresponding statement for performative decision OI is a bit more subtle. Whereas above, the decision rules $h \in \mathcal{H}$ do not depend in any way on ω, the optimal decision rule $\tilde{f}_{\ell_\omega}$ based on $\tilde{p}$, is allowed to depend on the specified loss and, thus, on ω. Still, we argue that for any ω, $\tilde{f}_{\ell_\omega} = \tilde{f}_\ell$. This equality follows by the fact that the optimal decision rule is chosen pointwise, for each $x \in \mathcal{X}$. In particular, for all $x \in \mathcal{X}$, scaling the loss by $\omega(x)$ changes the scale of the optimization, but not the minimizer:

$$\tilde{f}_{\ell_\omega}(x) = \mathop{\arg\min}_{\widehat{y} \in \widehat{\mathcal{Y}}} \mathop{\mathbb{E}}_{y \sim \tilde{p}(x, \widehat{y})} [\omega(x) \cdot \ell(\widehat{y}, y)] = \mathop{\arg\min}_{\widehat{y} \in \widehat{\mathcal{Y}}} \mathop{\mathbb{E}}_{y \sim \tilde{p}(x, \widehat{y})} [\ell(\widehat{y}, y)] = \tilde{f}_\ell(x).$$

Thus, the same identities from above can be applied to prove that if $\tilde{p}$ is $(\mathcal{L}_\omega, \varepsilon)$-DOI for a fixed distribution $\mathcal{D}$, then it is also $(\mathcal{L}, \varepsilon)$-DOI for every $\mathcal{D}_\omega \in \mathcal{D}_\mathcal{W}$. The proposition follows by applying Theorem 2.7 separately over each $\mathcal{D}_\omega \in \mathcal{D}_\mathcal{W}$. ■

Before moving on, we highlight that designing omnipredictors requires the learner to account for possible shifts in the distribution at *training* time, not at test time. At test time, the learner simply chooses predictions $\widehat{y}$ according to the function $\tilde{f}_\ell$, without needing to first infer what the underlying distribution $\mathcal{D}$ may be. The decision rule $\tilde{f}_\ell$ is simultaneously optimal for all of them. This design choice shifts the burden of technical sophistication and expertise from the user of the system to its designer. The user is free to focus on the choice of loss function ℓ to balance between forecasting and steering knowing that naive usage of $\tilde{f}_\ell$ is guaranteed to work.

4 Learning Algorithms for Performative Omniprediction

In this section, we introduce a general purpose algorithm, POI-Boost, which provably returns a performative omnipredictor $\tilde{p}$ for any class of hypothesis $\mathcal{H}$ and collection of losses $\mathcal{L}$. Our algorithmic approach is centered on establishing two reductions. First, we prove that, similar to previous work in the OI literature, learning Performative OI predictors reduces to the problem of *auditing* for outcome indistinguishability.

The auditing problems we reduce to involve determining whether the losses under the decision rules in $\mathcal{H}$ and $\{\tilde{f}_\ell\}$ are the same for Nature's outcomes and our modeled outcomes *under outcome performativity*. While in the supervised learning setting we only need to reason about a single outcome distribution, in our setting, different different decision rules induce different distributions over outcomes, and we want to audit for indistinguishability over each of these induced distributions. Despite this challenge, we show that, given access to appropriately

Performative OI Boost (POI-Boost)

Input: Set of losses $\mathcal{L}$, hypotheses $\mathcal{H}$, distribution $\mathcal{D}$, tolerance $\varepsilon > 0$

Initialize: $q^{(1)}(\cdot) \leftarrow [1/2, \ldots, 1/2] \in [0,1]^{|\widehat{\mathcal{Y}}|}$

For $t = 1, 2, \ldots$

– **If:** there exists **a**) $(h, \ell) \in \mathcal{H} \times \mathcal{L}$ or, **b**) $(h, \ell) \in \{(f_{\ell,t}, \ell) : \ell \in \mathcal{L}\}$

$$\underbrace{\left| \mathop{\mathbb{E}}_{\substack{x \sim \mathcal{D} \\ \tilde{y} \sim \tilde{p}(x, h(x))}} [\ell(x, h(x), \tilde{y})] - \mathop{\mathbb{E}}_{\substack{x \sim \mathcal{D} \\ y^* \sim p^*(x, h(x))}} [\ell(x, h(x), y^*)] \right|}_{\mathsf{err}_t} \geqslant \varepsilon \tag{6}$$

Then: Update the representation $q^{(t)}$:

$$q^{(t+1)}(\cdot) \leftarrow \Pi\Big(q^{(t)}(\cdot) - \eta^{(t)} \, v_{\ell,h}(\cdot) \Big) \tag{7}$$

where $\eta_t = -\varepsilon \cdot \mathsf{sign}(\mathsf{err}_t) \,/\, \ell_{\max}$ and

$$v_{\ell,h}(\cdot) = \begin{bmatrix} (\ell(\cdot, \widehat{y}_1, 1) - \ell(\cdot, \widehat{y}_1, 0)) \, \mathbf{1}\{h(\cdot) = \widehat{y}_1\} \\ \ldots \\ (\ell(\cdot, \widehat{y}_k, 1) - \ell(\cdot, \widehat{y}_k, 0)) \, \mathbf{1}\{h(\cdot) = \widehat{y}_k\} \end{bmatrix} \in \mathbb{R}^{|\widehat{\mathcal{Y}}|}, \quad \widehat{\mathcal{Y}} = \{\widehat{y}_1, \ldots, \widehat{y}_k\} \tag{8}$$

– **Else:** terminate and return the function $\tilde{p}(x, \widehat{y}) = q^{(t)}(x)[\,\widehat{y}\,]$

Figure 2: Algorithm for generating performative omnipredictors. The algorithm proceeds by repeatedly verifying whether the intermediate predictors $p^{(t)}$ satisfy the POI definition, outlined in **a**), as well as the DOI definition, outlined in **b**). If neither is violated, the procedure terminates. Otherwise, the algorithm implicitly updates the representation $q^{(t)}$ of the predictor $p^{(t)}$. Given an input x, $q^{(t)}(x)$ is a vector of length $|\widehat{\mathcal{Y}}|$ whose $\widehat{y}$ entry, $q^{(t)}(x)[\,\widehat{y}\,]$, represents $p^{(t)}(x, \widehat{y})$. The operator Π clips entries of its input vector to lie in $[0,1]$. For the sake of clarity, here we present the simplest version of the algorithm where the search outlined in Equation 6 is proper, however this condition can be easily relaxed as discussed in Section 4.2.

randomized data, we can reduce this performative auditing problem to standard supervised learning primitives. In this second reduction, we make use of computational and statistical assumptions: access to an appropriate supervised learner (computational) and access to randomized control data (statistical).

While we instantiate our algorithm with specific computational and statistical assumptions, the framework for learning is completely generic and modular. In particular, any solution to the auditing problem can be used to implement the algorithm. It stands to reason that our assumptions could be relaxed in the future, or that in certain settings, incomparable assumptions lead to more effective auditing, which in turn would lead to more efficient learning of performative omnipredictors.

4.1 Reducing Indistinguishability to Auditing

We start by establishing our first reduction. We prove that the POI-Boost algorithm (Figure 2) returns a performative omnipredictor $\tilde{p}$ after a small, polynomial number of calls to an auditing subroutine (described in Equation 6 & Figure 3), without yet describing the runtime or sample complexity of the auditing step itself. We address these questions in the next subsection. The algorithm works for any outcome performative problem where the number of predicted labels $\widehat{\mathcal{Y}}$ is finite and the loss functions are bounded.

Representing Predictors. For the sake of our analysis, it is helpful to distinguish between the predictor $\tilde{p} : \mathcal{X} \times \widehat{\mathcal{Y}} \to [0,1]$ as a function, and the implementation of $\tilde{p}$ in code. In our learning algorithm, we represent the function $\tilde{p}$ in terms of vector-valued functions $\tilde{q} : \mathcal{X} \to [0,1]^{|\widehat{\mathcal{Y}}|}$. Given $x \in \mathcal{X}$, $\tilde{q}(x)$ is a vector of length $|\widehat{\mathcal{Y}}|$ whose $\widehat{y}$ entry, $q^{(t)}(x)[\,\widehat{y}\,]$, represents $\tilde{p}(x, \widehat{y})$.

Of course, there is a correspondence between these functions $\tilde{p}$ and $\tilde{q}$ where each $\tilde{p}$ leads to a unique $\tilde{q}$ and vice versa. The key difference is that $\tilde{q}(x)$ returns $\tilde{p}(x, \widehat{y})$ for all $\widehat{y} \in \widehat{\mathcal{Y}}$ in a single function call, while computing the same information using the direct $\tilde{p}$ representation would require $|\widehat{\mathcal{Y}}|$ functions calls. While this might seem like a minor detail, these representations have meaningful differences in terms of the circuit complexity of performative omnipredictors. Crucially, for any loss ℓ, to compute $\tilde{f}_\ell(x)$, we need the value of of $\tilde{p}(x, \widehat{y})$ for all $\widehat{y} \in \widehat{\mathcal{Y}}$, and this computation can be performed using a single call to $\tilde{q}$. In other words, we perform $|\widehat{\mathcal{Y}}|$ times the work per call to $\tilde{q}$ to avoid $|\widehat{\mathcal{Y}}|$ *recursive* calls to the predictor within the construction. Avoiding further calls to these functions avoids branching factors and an exponential blowup in the complexity of the resulting predictors as we illustrate.

Algorithm Description. As outlined in Figure 2, POI-Boost is an iterative algorithm which nonparametrically constructs a predictor $\tilde{p}$, represented in terms of a vector-valued function $\tilde{q}$, by stringing together copies of circuits which compute losses ℓ and decision rules h. At each iteration, the algorithm first appeals to auditing subroutines to check if there: is a) a pair h, l for which the current predictor $p^{(t)}$, fails the performative OI guarantee, or b) a loss function ℓ for which the decision rule $f_{\ell,t}$ fails the decision OI guarantee. If neither condition is violated, then the algorithm terminates since $p^{(t)}$ satisfies both indistinguishability conditions and consequently must be an omnipredictor as per Theorem 2.7.

On the other hand, if one of these conditions is violated, we perform an update to the representation $q^{(t)}$ of the current predictor $p^{(t)}$. These updates nudge the predictor closer to p^* by essentially performing gradient descent in function space [MBBF99]. These updates are done implicitly in the sense that we can update the representation $q^{(t)}$ for all x in $\mathcal{X}$ simultaneously by simply adding a copy of the circuit computing $v_{\ell,h} : \mathcal{X} \to \mathbb{R}^{|\widehat{\mathcal{Y}}|}$ (Equation 8) which is defined in terms of a loss ℓ and decision rule h. By bounding the total number of updates via a potential argument, we can ensure that we don't add too many copies of these functions so that the final predictor is computationally efficient.

Proposition 4.1. *The POI-Boost algorithm described in Figure 1 terminates in at most $|\widehat{\mathcal{Y}}|\ell_{\max}^2/\varepsilon^2$ many iterations and returns a predictor $\tilde{p}$ that is $(\mathcal{L}, \mathcal{H}, \varepsilon)$-performative OI and $(\mathcal{L}, \varepsilon)$-performative decision OI. Consequently, $\tilde{p}$ is a $(\mathcal{L}, \mathcal{H}, 2\varepsilon)$-performative omnipredictor.*

Proof. The guarantee that $\tilde{p}$ is performative decision OI and performative OI follow directly from the termination criterion. Therefore, the proposition follows from proving that this termination criteria is met within the stated number of iterations.

The key insight is that if the indistinguishability constraint in Equation 6 is violated for any ℓ or h, then updating the representation $q^{(t)}$ ensures that we will have made nontrivial progress on a common potential function. Since this potential is bounded from above and below, and we make nontrivial progress with every update, the total number of updates must be bounded. In more detail, first, note that for any model p, loss ℓ, hypothesis h, and $\mathcal{Y} = \{0, 1\}$:

$$\begin{aligned}\mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}\\ y\sim p(x,h(x))}}[\ell(x,h(x),y)] &= \mathop{\mathbb{E}}_{x}\mathop{\mathbb{E}}_{y|x}[\ell(x,h(x),y)]\\ &= \mathop{\mathbb{E}}_{x}[\ell(x,h(x),0) + (\ell(x,h(x),1) - \ell(x,h(x),0))\cdot p(x,h(x))].\end{aligned}$$

From this rewriting, and the definition of $v_{\ell,h}$ in Equation 8, the difference in performative risks between the predictors $p^{(t)}$ and p^* for a pair ℓ, h can be expanded as,

$$\mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}\\ y_t\sim p^{(t)}(x,h(x))}}[\ell(x,h(x),y_t)] - \mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}\\ y^*\sim p^*(x,h(x))}}[\ell(x,h(x),y^*)] = \mathop{\mathbb{E}}_{x}\langle q^{(t)}(x) - q^*(x),\ v_{\ell,h}(x)\rangle. \tag{9}$$

Now consider the potential, written in terms of the representations $q^{(t)}$, $\mathbb{E}_x\|q^{(t+1)}(x) - q^*(x)\|_2^2$. By definition of the update rule in the algorithm, this potential is equal to:

$$\mathop{\mathbb{E}}_{x}\|\Pi\left(q^{(t)}(x) - \eta^{(t)}v_{\ell,h}(x)\right) - q^*(x)\|_2^2.$$

Because the projection (or clipping) operator Π can only decrease the distance to p^*, if an update is performed, the difference between potentials at adjacent time steps,

$$\mathop{\mathbb{E}}_{x}\|q^{(t+1)}(x) - q^*(x)\|_2^2 - \mathop{\mathbb{E}}_{x}\|q^{(t)}(x) - q^*(x)\|_2^2,$$

is upper bounded by the sum of two terms,

$$-2\eta_t \mathop{\mathbb{E}}_{x}\langle q^{(t)}(x) - q^*(x), v_{\ell,h}(x)\rangle + \eta_t^2 \mathop{\mathbb{E}}_{x}\|v_{\ell,h}(x)\|_2^2.$$

Using the identity from Equation 9 and the definition of $v_{\ell,h}$ from Equation 8, this is equal to:

$$-2\eta_t \mathsf{err}_t + \eta_t^2 \mathop{\mathbb{E}}_{x}[(\ell(x,h(x),1) - \ell(x,h(x),0))^2].$$

Because losses lie in $[0, \ell_{\max}]$, the second term is less than $\ell_{\max}^2\eta_t^2$. Furthermore, from the auditing guarantee, $|\mathsf{err}_t| > \varepsilon$. By setting the step size $\eta^{(t)}$ to be $-\varepsilon\cdot\mathrm{sign}(\mathsf{err}_t)\,/\,\ell_{\max}$, we conclude that the difference in potentials across adjacent time steps satisfies,

$$2\eta_t\mathsf{err}_t + \eta_t^2 \mathop{\mathbb{E}}_{x}[(\ell(x,h(x),1) - \ell(x,h(x),0))^2] \leqslant -2\eta_t\mathsf{err}_t + \eta_t^2\ell_{\max}^2 \leqslant -\varepsilon^2/\ell_{\max}^2.$$

Since the potential is nonnegative and bounded above by $|\widehat{\mathcal{Y}}|$, the maximum number of iterations until the termination criterion is met must be at most $|\widehat{\mathcal{Y}}|\ell_{\max}^2/\varepsilon^2$. ∎

An important consequence of this result is that it reveals the existence of omnipredictors $\tilde{p}$ that admit computationally efficient approximations. This result is subtle, even in light of previous work on OI-style boosting algorithms. Intuitively, the final $\tilde{p}$ is built out by stringing together copies of functions in $\mathcal{H}$, $\mathcal{L}$, and decision rules $f_{\ell,t}$. Because these decision rules $f_{\ell,t}$ are defined in terms of an optimization procedure involving the intermediate constructions $p^{(t)}$,

which themselves depend on previous models $p^{(t-1)}$, a naive implementation of $\tilde{p}$ can result in a recursion that induces an exponentially large circuit. Specifically, the naive implementation would make $|\widehat{\mathcal{Y}}|$ recursive calls to the prior circuit in order to compute $f_{\ell,t}$, resulting in a growth rate of $|\widehat{\mathcal{Y}}|^t$.

However, by carefully ordering the relevant computations and "caching" previous work, we avoid this blow-up. The key insight is the following. By designing a circuit that computes the value of $\tilde{p}(x, \widehat{y})$ for every $\widehat{y} \in \widehat{\mathcal{Y}}$ simultaneously, we can avoid recursive calls to the circuit. By maintaining the intermediate computations of $q^{(t)}$, we can avoid a branching factor in the program and preserve efficiency.

Theorem 4.2. *Assume that the functions in $\mathcal{H}$ and $\mathcal{L}$ are computable by circuits of size at most s, then the predictor $\tilde{p}$ returned by the POI-Boost algorithm has size at most $\ell_{\max}^2/\varepsilon^2 \cdot \mathrm{poly}(s, |\widehat{\mathcal{Y}}|)$.*

Proof. The final predictor consists of a summation of the initial prediction, followed by the update from each iteration. We bound the growth of the circuit computing the predictor by induction. Formally, let S_t be the circuit size for computing $q^{(t)}$. Then, we show that $S_{t+1} \leqslant S_t + \mathrm{poly}(|\widehat{\mathcal{Y}}|, s)$ for all $t \geqslant 1$. Thus, by the overall bound on the iteration complexity, the final predictor can be implemented using a circuit of size $S \leqslant \ell_{\max}^2/\varepsilon^2 \cdot \mathrm{poly}(s, |\widehat{\mathcal{Y}}|)$. To begin, the initial constant predictor $q^{(1)}$ can be implemented using a circuit of size at most $S_1 = |\widehat{\mathcal{Y}}| \leqslant \mathrm{poly}(s, |\widehat{\mathcal{Y}}|)$ by hard-coding the constant vector.

By the update rule, each update incorporates a function of the form,

$$g^{(t)}(x) \stackrel{\text{def}}{=} \eta^{(t)} v_{\ell,h}^{(t)}(x) = \eta^{(t)} \begin{bmatrix} (\ell^{(t)}(x, \widehat{y}_1, 1) - \ell^{(t)}(x, \widehat{y}_1, 0)) \, \mathbf{1}\{h^{(t)}(x) = \widehat{y}_1\} \\ \cdots \\ (\ell^{(t)}(x, \widehat{y}_k, 1) - \ell^{(t)}(x, \widehat{y}_k, 0)) \, \mathbf{1}\{h^{(t)}(x) = \widehat{y}_k\} \end{bmatrix} \in \mathbb{R}^{|\widehat{\mathcal{Y}}|}, \tag{10}$$

where $\ell^{(t)}$ and $h^{(t)}$ define the test function surfaced by the auditing subroutine at time step t. Within these updates, the function $h^{(t)}$ may *a*) come from $\mathcal{H}$ due to a POI violation or *b*) equal $f_{\ell,t}$ for some $\ell \in \mathcal{L}$ due to a DOI violation.

In the first case where $h^{(t+1)} \in \mathcal{H}$, $q^{(t+1)}(x)$ can be computed by evaluating evaluating $q^{(t)}(x)$, and then evaluating $h(x)$ and $\ell(x, \widehat{y}, y)$, for every $\widehat{y}$ and y. By assumption, the latter operations require circuits of size at most $\mathrm{poly}(s, |\widehat{\mathcal{Y}}|)$. Paired with the inductive hypothesis, the resulting circuit size can be bounded as $S_{t+1} \leqslant S_t + \mathrm{poly}(s, |\widehat{\mathcal{Y}}|) \leqslant (t+1) \cdot \mathrm{poly}(s, |\widehat{\mathcal{Y}}|)$.

For the second case, we recall the definition of $f_{\ell,t}(\cdot)$, we can express its computation as a minimization over $\widehat{\mathcal{Y}}$ of expected losses that depend on $q^{(t)}(\cdot)[\widehat{y}]$ for each $\widehat{y} \in \widehat{\mathcal{Y}}$.

$$f_{\ell,t}(x) = \operatorname*{arg\,min}_{\widehat{y} \in \widehat{\mathcal{Y}}} \{\ell(x, \widehat{y}, 0) + (\ell(x, \widehat{y}, 1) - \ell(x, \widehat{y}, 0)) \cdot q^{(t)}(x)[\widehat{y}]\}.$$

Importantly, to compute each term in the minimization, we only need to compute the vector $q^{(t)}(x)$ once. The remaining terms, $\ell(x, \widehat{y}, 0)$ and $\ell(x, \widehat{y}, 1)$ (for every $\widehat{y}$), can again be computed by a circuit of size $\mathrm{poly}(|\widehat{\mathcal{Y}}|, s)$. Since the minimization itself can be done by linearly enumerating over $\widehat{\mathcal{Y}}$, we again preserve the invariant that $S_{t+1} \leqslant S_t + \mathrm{poly}(|\widehat{\mathcal{Y}}|, s)$. ∎

4.2 Reducing Auditing to Supervised Learning

Having shown how omniprediction reduces to an auditing problem, we now complete our analysis of the POI-Boost algorithm by showing that auditing itself reduces to cost-sensitive

Audit$(\mathcal{H}, \mathcal{L}, \tilde{p}, \varepsilon)$

If: there exists $(h, \ell) \in \mathcal{H} \times \mathcal{L}$

$$\left| \mathop{\mathbb{E}}_{\substack{x \sim \mathcal{D} \\ \tilde{y} \sim \tilde{p}(x, h(x))}} \ell(x, h(x), \tilde{y}) - \mathop{\mathbb{E}}_{\substack{x \sim \mathcal{D} \\ y^* \sim p^*(x, h(x))}} \ell(x, h(x), y^*) \right| \geqslant \varepsilon$$

Then: return h, l

Else: return False

Figure 3: The key auditing step in the POI-Boost algorithm. In each iteration of the algorithm, we run two auditing steps: once to check for the POI condition over $\mathcal{H} \times \mathcal{L}$ and once to check for the DOI condition over $\{(f_{\ell,t}, \ell) : \ell \in \mathcal{L}\}$. See the proof of Corollary 4.8 for further discussion.

classification over a single, static distribution. In doing so, we address the statistical and computational complexity of solving this auditing step.

From examining the auditing condition in Figure 3, perhaps the most obvious strategy is to choose a decision rule h, and to collect a dataset of triples $(x, \widehat{y}, y^*)$ where $\widehat{y} = h(x)$ for every x and $y^* \sim p^*(x, h(x))$. If the loss ℓ is bounded, a standard application of Hoeffding's inequality shows that empirical risk of the loss concentrates around its expectation:

$$\frac{1}{n} \sum_{i=1}^{n} \ell(x_i, h(x_i), y_i^*) \approx \mathop{\mathbb{E}}_{\substack{x \sim \mathcal{D} \\ y^* \sim p^*(x, h(x))}} [\ell(x, h(x), y_*)].$$

Therefore, one could implement the auditing step by enumerating over all h, deploying h to collect a new dataset every time, and then nonadaptively computing the empirical performative risk of h on every $\ell \in \mathcal{L}$. This procedure would however require $\tilde{O}(|\mathcal{H}|/\varepsilon^2 \log |\mathcal{L}|)$ many samples.

On the other hand, if we have access to randomized predictions $\widehat{y}$, we can estimate the empirical risk of every pair h, ℓ off of a *single* distribution by using inverse propensity scoring. The following lemma is well-known within various communities, and, in particular, the contextual bandits literature (see e.g. [AHK+14, DHK+11]).[10]

We use the shorthand $(x, \widehat{y}, y) \sim \mathcal{D}_{\text{rct}}$ to denote the sampling process where inputs are sampled from the base distribution $x \sim \mathcal{D}$, decisions $\widehat{y}$ are assigned uniformly at random, $\widehat{y} \sim \text{Unif}(\widehat{\mathcal{Y}})$, and the outcomes are sampled according to Nature's model $y^* \sim p^*(x, \widehat{y})$.

Lemma 4.3. *Assume that $\widehat{\mathcal{Y}}$ is a finite set. Then, for any hypothesis $h : \mathcal{X} \to \widehat{\mathcal{Y}}$,*

$$\mathop{\mathbb{E}}_{\substack{x \sim \mathcal{D}_x \\ y^* \sim p^*(x, h(x))}} [\ell(x, h(x), y^*)] = |\widehat{\mathcal{Y}}| \cdot \mathop{\mathbb{E}}_{(x, \widehat{y}, y^*) \sim \mathcal{D}_{\text{rct}}} [\ell(x, \widehat{y}, y^*) \mathbf{1}\{h(x) = \widehat{y}\}].$$

Proof. We present the proof for the case where $\widehat{\mathcal{Y}} = \{0, 1\}$ is binary, but the general case follows

[10] This result can be generalized to the case where for every x, $\widehat{y} \sim q(x)$ for some known distribution q, where $q \neq \text{Unif}(\widehat{\mathcal{Y}})$ but where q has full support over $\widehat{\mathcal{Y}}$.

the same pattern. We expand out the left hand side as:

$$\begin{aligned}\mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}\\ y^*\sim p^*(x,h(x))}}[\ell(x,h(x),y^*)] &= \mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}\\ y^*_{(1)}\sim p^*(x,1),y^*_{(0)}\sim p^*(x,0)}}\Big[\ell(x,1,y^*_{(1)})\mathbf{1}\{h(x)=1\}+\ell(x,0,y^*_{(0)})\mathbf{1}\{h(x)=0\}\Big]\\ &= \mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}\\ y^*_{(1)}\sim p^*(x,1)}}\Big[\ell(x,1,y^*_{(1)})\mathbf{1}\{h(x)=1\}\Big]+\mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}\\ y^*_{(0)}\sim p^*(x,0)}}\Big[\ell(x,1,y^*_{(0)})\mathbf{1}\{h(x)=0\}\Big].\end{aligned}$$

Reweighting the term on the right hand side, we observe our desired equality:

$$\begin{aligned}\mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}\\ \widehat{y}\sim\mathrm{Ber}(1/2)\\ y^*\sim p^*(x,\widehat{y})}}[\ell(x,\widehat{y},y^*)\mathbf{1}\{h(x)=\widehat{y}\}] &= \frac{1}{2}\mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}\\ y^*_{(1)}\sim p^*(x,1)}}\Big[\ell(x,1,y^*_{(1)})\mathbf{1}\{h(x)=1\}\Big]\\ &\quad+\frac{1}{2}\mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}\\ y^*_{(0)}\sim p^*(x,0)}}\Big[\ell(x,1,y^*_{(0)})\mathbf{1}\{h(x)=0\}\Big].\end{aligned}$$

■

There are two main takeaways from this lemma. First, it shows that the statistical complexity of auditing can be exponentially better than the the naive strategy outlined previously.

Corollary 4.4. *Let $\{(x_i,\widehat{y}_i,\tilde{y}_i)\}_{i=1}^n$ be a dataset of n i.i.d samples from $\mathcal{D}_{\mathrm{rct}}$. If*

$$n \geqslant \frac{2\ell_{\max}^2|\widehat{\mathcal{Y}}|^2\cdot\log(2|\mathcal{H}||\mathcal{L}|/\delta)}{\varepsilon^2},$$

then with probability $1-\delta$,

$$\max_{h\in\mathcal{H},\ell\in\mathcal{L}}\Big|\mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}\\ y^*\sim p^*(x,h(x))}}[\ell(x,h(x),y^*)]-\frac{1}{n}\sum_{i=1}^n|\widehat{\mathcal{Y}}|\cdot\ell(x_i,\widehat{y}_i,y_i^*)\mathbf{1}\{h(x_i)=\widehat{y}_i\}\Big|\leqslant\varepsilon.$$

Proof. From the previous lemma, we have that for any loss ℓ and decision rule h,

$$\mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}\\ y^*\sim p^*(x,h(x))}}[\ell(x,h(x),y^*)]=|\widehat{\mathcal{Y}}|\cdot\mathop{\mathbb{E}}_{(x,\widehat{y},y^*)\sim\mathcal{D}_{\mathrm{rct}}}[\ell(x,\widehat{y},y^*)\mathbf{1}\{h(x)=\widehat{y}\}].$$

Because $|\widehat{\mathcal{Y}}|\cdot\ell(x_i,\widehat{y}_i,y_i^*)\mathbf{1}\{h(x_i)=\widehat{y}_i\}$ is uniformly bounded by $\ell_{\max}|\widehat{\mathcal{Y}}|$, we can apply Hoeffding's inequality to argue that the probability that the empirical estimate is far from the true expectation

$$\Big|\mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}\\ y^*\sim p^*(x,h(x))}}[\ell(x,h(x),y^*)]-\frac{1}{n}\sum_{i=1}^n|\widehat{\mathcal{Y}}|\cdot\ell(x_i,\widehat{y}_i,y_i^*)\mathbf{1}\{h(x_i)=\widehat{y}_i\}\Big|>\varepsilon$$

is bounded as $2\exp\left(\frac{2\varepsilon^2}{n(\ell_{\max}|\widehat{\mathcal{Y}}|)^2}\right)$. The result follows by rearranging for failure probability δ, and taking a union bound over all $h\in\mathcal{H}$ and $\ell\in\mathcal{L}$. ■

Consequently, for a single iteration of the POI-Boost algorithm, the auditing step can be implemented by enumerating over all $\mathcal{L}$ and $\mathcal{H}$ and non-adaptively evaluating their empirical risks on a single dataset of size $\tilde{\mathcal{O}}(\ell_{\max}^2|\widehat{\mathcal{Y}}|^2/\varepsilon^2 \log(|\mathcal{H}||\mathcal{L}|))$.[11] Typically, we think of $|\widehat{\mathcal{Y}}|$ as a small constant and the class of decision rules $\mathcal{H}$ as a rich collection. From this result, we see that at least statistically, we can hope to design omnipredictors that are optimal with respect to an exponential number of losses and decision rules.

Here, we present the simplest possible analysis of this result and state our bounds for finite classes $\mathcal{H}$ and $\mathcal{L}$. It is certainly feasible to achieve sharper results and to state bounds in terms of VC-dimension or other sharper notions of statistical complexity. However, the goal of our initial work on outcome performativity is not to establish the tightest bounds, but to provide a broad overview of what is possible. We hope future work will provide a precise understanding of the sample complexity of omniprediction in outcome performativity.

The following proposition summarizes the sample complexity of omniprediction if the auditing steps for the POI and DOI conditions are implemented via a naive learner that linearly enumerates over all h, ℓ and evaluates their empirical risk on a single dataset of RCT samples.

Proposition 4.5. *Given labeled data $(x, \widehat{y}, y^*) \sim \mathcal{D}_{\text{rct}}$ drawn from Nature and unlabeled samples $x \sim \mathcal{D}$, the POI-boost can be implemented using at most:*

- $\mathcal{O}(\ell_{\max}^2|\widehat{\mathcal{Y}}|^2 \log(\frac{|\mathcal{H}||\mathcal{L}|}{\delta})/\varepsilon^2 + \ell_{\max}^4|\widehat{\mathcal{Y}}|^3 \log(\frac{|\mathcal{L}|\ell_{\max}|\widehat{\mathcal{Y}}|}{\delta\varepsilon})/\varepsilon^4)$ *labeled samples*
- $\mathcal{O}(\ell_{\max}^4|\widehat{\mathcal{Y}}|^3 \log(\frac{|\mathcal{H}||\mathcal{L}|\ell_{\max}|\widehat{\mathcal{Y}}|}{\delta\varepsilon})/\varepsilon^4)$ *unlabeled samples*

Proof. In each iteration of the POI-boost algorithm, we need to audit for the POI and DOI guarantees (conditions a and b). We can implement each of the auditing steps by explicit enumeration.

For POI, at each iteration t we enumerate over $\mathcal{H}$ and $\mathcal{L}$ and evaluate the empirical counterparts of

$$\mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}\\ \tilde{y}\sim p^{(t)}(x,h(x))}}[\ell(x,h(x),\tilde{y})] \quad \text{and} \quad \mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}\\ y^*\sim p^*(x,h(x))}}[\ell(x,h(x),y^*)]. \tag{11}$$

By Corollary 4.4, the empirical versions of these quantities concentrate around their expectations. To get an ε approximation, with probability $1-\delta$, we require at most $\mathcal{O}(\ell_{\max}^2|\widehat{\mathcal{Y}}|^2 \log(|\mathcal{H}||\mathcal{L}|/\delta)/\varepsilon^2)$ many samples. At each iteration t, the expectation on the left changes, since we update $p^{(t)}$. However, to evaluate this expectation we only need *unlabeled* samples, since labels $\tilde{y}$ come from our own model $p^{(t)}$. On the other hand, the expectation on the right in Equation 11 does not depend on t, so we need not recompute it at every iteration. Because the total number of iterations is bounded by $\ell_{\max}^2|\widehat{\mathcal{Y}}|/\varepsilon^2$, applying a union bound on δ, to achieve the POI guarantee we only need a total of $\mathcal{O}(\ell_{\max}^4|\widehat{\mathcal{Y}}|^3 \log(|\mathcal{H}||\mathcal{L}|\ell_{\max}|\widehat{\mathcal{Y}}|\varepsilon^{-1}\delta^{-1})/\varepsilon^4)$ unlabeled samples and $\mathcal{O}(\ell_{\max}^2|\widehat{\mathcal{Y}}| \log(|\mathcal{H}||\mathcal{L}|/\delta)/\varepsilon^2)$ labeled samples.

For the DOI guarantee outline in condition b, we instead need to approximate

$$\mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}\\ \tilde{y}\sim p^{(t)}(x,f_{\ell,t}(x))}}[\ell(x,f_{\ell,t}(x),\tilde{y})] \quad \text{and} \quad \mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}\\ y^*\sim p^*(x,f_{\ell,t}(x))}}[\ell(x,f_{\ell,t}(x),y^*)]. \tag{12}$$

Note that both of these expectations now depend on t, because the decision rules $f_{\ell,t}$ can change between iterations. Again, by Corollary 4.4, if we enumerate over all $|\mathcal{L}|$ losses and

[11] An analogous result applies if we replace p^* by $\tilde{p}$, which we assume that the learner can easily sample from.

decision rules $f_{\ell,t}$ at each iteration, the empirical counterparts of these expressions on a dataset of size $\mathcal{O}(\ell_{\max}^2|\widehat{\mathcal{Y}}|^2\log(|\mathcal{L}|/\delta)/\varepsilon^2)$ concentrates. Collecting a new dataset at every iteration, we get that the total number of labeled (and unlabeled) samples is bounded by $\mathcal{O}(\ell_{\max}^4|\widehat{\mathcal{Y}}|^3\log(|\mathcal{L}||\widehat{\mathcal{Y}}|\ell_{\max}\delta^{-1}\varepsilon^{-1})/\varepsilon^4)$. ∎

Cost-Sensitive Classification. The second main takeaway from Lemma 4.3 is that auditing can now be rewritten as as the solution to a cost-sensitive multiclass classification problem over $|\widehat{\mathcal{Y}}|$ many classes. This result completes our analysis showing how omniprediction can be reduced to basic supervised learning problems.

In light of previous results, the main benefit of this reduction is that it enabled the design of oracle-efficient algorithms which can be faster than the naive learner used in Proposition 4.5. We start by first defining what we mean by cost-sensitive classification.

Definition 4.6. *Let $\mathcal{X}$ be a feature space, $\widehat{\mathcal{Y}}$ be a finite set of k classes, and $\mathcal{D}$ be a distribution over $\mathcal{X}\times[-1,1]^k$. For $(x,c)\sim\mathcal{D}$, we say that c is a cost vector whose entries $c(\widehat{y})$ denote the costs of predicting label $\widehat{y}$ on feature x. An algorithm $\mathcal{A}_{\mathrm{csc}}$ is an ρ-cost-sensitive learner for a hypothesis class $\mathcal{H}$ if for any distribution $\mathcal{D}$ over $\mathcal{X}\times[-1,1]^k$, promised that there exists $h\in\mathcal{H}$ such that $\mathbb{E}_{(x,c)\sim\mathcal{D}}\,c(h(x))\leqslant-\rho$, $\mathcal{A}_{\mathrm{csc}}$ returns a hypothesis h' such that $\mathbb{E}_{(x,c)\sim\mathcal{D}}\,c(h'(x))\leqslant-\rho/2$.*

Cost-sensitive classification is a well-studied supervised learning problem for which many, both passive and active learning algorithms, have been designed [BLR09, AZL04, LB05, KAH+17, Elk01]. There are a number of software packages that can be used to solve applied cost-senstive classification problems [PGV+18]. Like many problems in computational learning theory, cost-sensitive classification is known to be hard in the worst-case, but can be solved effectively in practice. As such, our goal is to design *oracle-efficient* learning algorithms, that make an small number of calls to cost-sensitive learner.

Here, we frame a "weak" version of the problem where the learning need not be exact, but where the search is proper, in the sense that $\mathcal{A}_{\mathrm{csc}}$ returns a hypothesis in $\mathcal{H}$. This latter condition can easily also be relaxed without changing the overall results. However, we opt to keep it as is for the sake of simplifying the presentation. The following proposition completes our reduction of auditing to supervised learning.

Proposition 4.7. *Let $\mathcal{A}_{\mathrm{csc}}$ be a cost-sensitive learner as per Definition 4.6. Then, given access to RCT samples $(x,\widehat{y},y^*)\sim\mathcal{D}_{\mathrm{rct}}$, we can solve the auditing problem outlined in Figure 3 using $2|\mathcal{L}|$ many calls to $\mathcal{A}_{\mathrm{csc}}$ with parameters $\rho=\varepsilon/(4\ell_{\max}|\widehat{\mathcal{Y}}|)$.*

Proof. By Lemma 4.3 we have that the difference in performative risk between p^* and $\tilde{p}$,

$$\mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}\\ \tilde{y}\sim\tilde{p}(x,h(x))}}[\ell(x,h(x),\tilde{y})]-\mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}\\ y^*\sim p^*(x,h(x))}}[\ell(x,h(x),y^*)],$$

is equal to:

$$|\widehat{\mathcal{Y}}|\cdot\mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}\\ \widehat{y}\sim\mathrm{Unif}(\widehat{\mathcal{Y}})\\ \tilde{y}\sim\tilde{p}(x,\widehat{y}),y^*\sim p^*(x,\widehat{y})}}[\mathbf{1}\{h(x)=\widehat{y}\}(\ell(x,h(x),\tilde{y})-\ell(x,h(x),y^*))].$$

Now, we note that terms inside the expectation can be written as entries in a cost vector c where for every sample $(x,\widehat{y},y^*,\tilde{y})$ we define the corresponding vector c to be,

$$c_\sigma(h(x))=\begin{cases}\sigma\cdot(\ell(x,h(x),\tilde{y})-\ell(x,h(x),y^*)) & \text{if } h(x)=\widehat{y}\\ 0 & \text{o.w}\end{cases},$$

Here, $\sigma \in \{\pm 1\}$ and we set $\sigma = 1$ to get the desired equality. Hence, for a fixed loss ℓ, we can transform RCT samples, $x \sim \mathcal{D}, \widehat{y} \sim \mathrm{Unif}(\widehat{\mathcal{Y}}), y^* \sim p^*(x, \widehat{y})$ to a cost sensitive classification problem such that for every $h \in \mathcal{H}$,

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}'}[c_{+1}(h(x))] = |\widehat{\mathcal{Y}}| \mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D} \\ \widehat{y}\sim\mathrm{Unif}(\widehat{\mathcal{Y}}) \\ \tilde{y}\sim\tilde{p}(x,\widehat{y}), y^*\sim p^*(x,\widehat{y})}} [\mathbf{1}\{h(x) = \widehat{y}\}(\ell(x, h(x), \tilde{y}) - \ell(x, h(x), y^*))].$$

To solve the auditing problem outlined in Figure 3, we need to check whether the absolute value of the difference is larger than ε. To do this, it therefore to suffices to run $\mathcal{A}_{\mathrm{csc}}$ twice (once with $\sigma = 1$ and once with $\sigma = -1$) for every loss $\ell \in \mathcal{L}$ to check if there exists a decision rule $h \in \mathcal{H}$ such that:

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}'}[c_{+1}(h(x))] \leqslant -\varepsilon \text{ or } \mathop{\mathbb{E}}_{x\sim\mathcal{D}'}[c_{-1}(h(x))] \leqslant -\varepsilon.$$

Because we normalize the cost vectors to have entries in $[-1, 1]$ in Definition 4.6, we can scale the vectors c_σ by $1/(4|\widehat{\mathcal{Y}}|\ell_{\max})$ and divide the tolerance parameter ε by the corresponding amount to match the desired interface. ∎

4.3 End-to-End Analysis

Having now presented these reductions showing how omniprediction can be reduced to cost sensitive classification, we now summarize our results so far and establish end-to-end bounds on the runtime and sample complexities of achieving omniprediction.

Corollary 4.8. *Assume that $h \in \mathcal{H}$ and $\ell \in \mathcal{L}$ can be evaluated in time* $\mathrm{poly}(\log(|\mathcal{H}|))$ *and* $\mathrm{poly}(\log(|\mathcal{L}|))$*, respectively. Let $\mathcal{A}_{\mathrm{csc}}$ be a ρ-cost-sensitive weak learner for $\mathcal{H}$ as per Definition 4.6. Assume that for any distribution $\mathcal{D}_{\mathrm{csc}}$ over pairs $(x, c) \in \mathcal{X} \times [-1, 1]^k$, $\mathcal{A}_{\mathrm{csc}}$ runs in time* $\mathrm{poly}(\log(|\mathcal{H}|), 1/\rho)$ *and uses at most* $\mathrm{poly}(\log(|\mathcal{H}|), 1/\rho)$ *many samples drawn from $\mathcal{D}_{\mathrm{csc}}$.*[12] *If the learner has access to samples drawn according $(x, \widehat{y}, y^*) \sim \mathcal{D}_{\mathrm{rct}}$, then, the POI-Boost algorithm:*

- *runs in time* $\mathcal{O}\left(|\mathcal{L}| \cdot \mathrm{poly}\left(1/\varepsilon,\ \ell_{\max},\ |\widehat{\mathcal{Y}}|,\ \log|\mathcal{L}|,\ \log|\mathcal{H}|\right)\right)$
- *uses at most* $\mathcal{O}\left(\mathrm{poly}\left(1/\varepsilon,\ \ell_{\max},\ |\widehat{\mathcal{Y}}|,\ \log|\mathcal{L}|,\ \log|\mathcal{H}|\right)\right)$ *many samples*

Proof. To bound the runtime, we note that by Proposition 4.1, the maximum number of iterations for POI-Boost is at most $\ell_{\max}^2|\widehat{\mathcal{Y}}|/\varepsilon^2$. In each iteration, we solve the auditing via two subroutines. One to check for the POI guarantee (condition *a* in Figure 2) and another routine to check the DOI guarantee (condition *b* in Figure 2). To audit for the POI guarantee, we call the cost-sensitive learner $2|\mathcal{L}|$ times with parameters $\rho = \varepsilon/(4\ell_{\max}|\widehat{\mathcal{Y}}|)$ as per Proposition 4.7, using labels derived from calculating each $\ell \in \mathcal{L}$ and evaluating $p^{(t)}(x, \widehat{y})$ in at most $\mathrm{poly}(\log(|\mathcal{L}|)) + \mathrm{poly}(\log(|\mathcal{H}|), 1/\varepsilon, \ell_{\max}, |\widehat{\mathcal{Y}}|)$ time. With the labels calculated, each of these calls has run time and sample complexity at most $\mathrm{poly}(\log(|\mathcal{H}|, 1/\varepsilon, \ell_{\max}, |\widehat{\mathcal{Y}}|))$.

To audit for the DOI guarantee over the $(h, \ell) \in \{(f_{\ell,t}, \ell) : \ell \in \mathcal{L}\}$, at each iteration, we use the naive strategy outlined in Section 4.2 where we enumerate over all $|\mathcal{L}|$ losses and evaluate the performative risk of each pair $(f_{\ell,t}, \ell)$ on a single dataset of RCT samples of size

[12] Here, we have avoided discussion on the failure probability parameter δ for the $\mathcal{A}_{\mathrm{csc}}$. However, it is clear that the relevant complexity bounds should depend only on $\log(1/\delta)$ and that applying a simple union bound would not change the nature of the resulting analysis. We therefore assume that the algorithms succeed with probability 1 for the sake of simplicity.

$\mathcal{O}(\ell_{\max}^2/\varepsilon^2|\widehat{\mathcal{Y}}|^2\log(|\mathcal{L}|))$ as per Corollary 4.4. Each auditing step for DOI therefore runs in time $|\mathcal{L}|\cdot\text{poly}(1/\varepsilon,\ell_{\max},|\widehat{\mathcal{Y}}|)$ and uses $\text{poly}(1/\varepsilon,\ell_{\max},|\widehat{\mathcal{Y}}|,\log(|\mathcal{L}|))$ many samples. All calls to the intermediate predictors $p^{(t)}$ also run in polynomial time as per Theorem 4.2. The final guarantees come from multiplying the sample and run time complexity of each iteration of the POI-boost algorithm by the bound on the total number of iterations. ■

The main take away from this result is that if the cost-sensitive classification problem can be solved efficiently, in the sense that the the relevant statistical and computational complexities scale as polylog$|\mathcal{H}|$, then the overall POI-Boost algorithm runs in time linear in $|\mathcal{L}|$, poly-logarithmically in the size of $\mathcal{H}$ and with at most polylog$|\mathcal{H}||\mathcal{L}|$ many samples. Therefore, we can hope to develop efficient omnipredictors that are optimal for exponentially many decision rules, and polynomially many losses.

Note that, because of the result outlined in Proposition 3.3, this theorem also bounds the statistical and computational complexity of achieving universally adaptable omnipredictors. More specifically, the number of samples and the runtime for achieving universally adaptable omnipredictors are also bounded by the quantities in Corollary 4.8 where we now replace the class $\mathcal{L}$ by augment collection $\mathcal{L}_{\mathcal{W}}$ as defined in Proposition 3.3. The main difference is that the relevant runtime and sample complexity bounds replace dependence on $|\mathcal{L}|$ by dependence on $|\mathcal{L}||\mathcal{W}|$ and replace dependence on $\ell_{\max}$ by $\ell_{\max}\omega_{\max}$. Here, $\omega_{\max}$ is the the worst case density ratio for the class $\mathcal{W}$.

$$\max_{\omega\in\mathcal{W}}\max_{x\in\text{supp}(\mathcal{D}_\omega)}\omega(x)=\frac{\mathcal{D}_\omega(x)}{\mathcal{D}(x)}.$$

This complexity measure capture the intuition that if individuals x are poorly represented over the distribution $\mathcal{D}$ we are learning over, then we need more samples (and consequently runtime), to learn universally adaptable omnipredictors. We think of these complexity parameters like $\omega_{\max}$ as a first step. It is an interesting question for future work to provide sharper notions of problem complexity and to find ways of designing omnipredictors for exponentially large collections of importance weights $\mathcal{W}$.

5 Connections to Multicalibration

So far, we have studied how extensions of outcome indistinguishability definitions enable the design of omnipredictors for performative settings. In the world of supervised learning, [DKR+21] established tight connections between outcome indistinguishability and various notions of multicalibration [HKRR18]. Given the complementary relationship between these two concepts in the supervised world, it is natural to speculate that generalizing multicalibration to the outcome performative setting might be fruitful.

In this section, we begin to examine these questions and discuss analogues of multiaccuracy and multicalibration for the performative setting. We start by showing that multiaccuracy naturally, and efficiently, extends to performative contexts, and provides an effective way to achieve performative outcome indistinguishability in a loss-independent fashion. Conversely, we illustrate how naive translations of multicalibration to performative prediction result in definitions whose complexity blows up exponentially in the number of predictions $|\widehat{\mathcal{Y}}|$. We conclude with some discussion of alternatives to calibration-style guarantees that could, in principle, be used to obtain efficient omnipredictors.

On Multiaccuracy. Theorem 2.7 shows how $(\mathcal{L}, \mathcal{H}, \varepsilon)$-performative omniprediction arises as a consequence of $(\mathcal{L}, \mathcal{H}, \varepsilon)$-performative OI and $(\mathcal{L}, \varepsilon)$-decision OI, where the OI distinguishers explicitly account for the collection of loss functions $\mathcal{L}$. Here, we show an efficient approach for obtaining POI for the class of all bounded input-oblivious loss functions.

We say that a loss function is *input-oblivious* if it only depends on the input $x \in \mathcal{X}$ via the decision $h(x)$. That is, for all x and x' and pairs $(\widehat{y}, y)$, $\ell(x, \widehat{y}, y) = \ell(x', \widehat{y}, y)$. Equivalently, these functions have domain $\widehat{\mathcal{Y}} \times \mathcal{Y}$ instead of $\mathcal{X} \times \widehat{\mathcal{Y}} \times \mathcal{Y}$. We use

$$\mathcal{L}_{\text{io}} = \{\ell : \widehat{\mathcal{Y}} \times \{0,1\} \to [0,1]\}$$

to denote the class of all bounded input-oblivious loss functions. Our first result for this section proves that a performative analogue of multiaccuracy [HKRR18, KGZ19] implies POI for $\mathcal{L}_{\text{io}}$.

Definition 5.1 (Multiaccuracy)**.** *For a distribution $\mathcal{D}$, hypothesis class $\mathcal{H}$, and $\varepsilon \geqslant 0$, a predictor $\tilde{p} : \mathcal{X} \times \widehat{\mathcal{Y}} \to [0,1]$ is $(\mathcal{H}, \varepsilon)$-multiaccurate under outcome performativity over $\mathcal{D}$ if for all $h \in \mathcal{H}$ and $\widehat{y} \in \widehat{\mathcal{Y}}$*

$$\left| \mathop{\mathbb{E}}_{\substack{x \sim \mathcal{D} \\ y^* \sim p^*(x, \widehat{y})}} [y^* \cdot \mathbf{1}\{h(x) = \widehat{y}\}] - \mathop{\mathbb{E}}_{\substack{x \sim \mathcal{D} \\ \tilde{y} \sim \tilde{p}(x, \widehat{y})}} [\tilde{y} \cdot \mathbf{1}\{h(x) = \widehat{y}\}] \right| \leqslant \varepsilon.$$

Here, we require that the expectation of our modeled outcome $\tilde{y} \sim \tilde{p}(x, h(x))$ is accurate after deploying each $h \in \mathcal{H}$, even when restricting our attention to the individuals $x \in \mathcal{X}$ such that $h(x) = \widehat{y}$. While seemingly simpler than performative OI, we show that multiaccuracy in fact implies performative OI for all input-oblivious losses.

Lemma 5.2. *If $\tilde{p} : \mathcal{X} \times \widehat{\mathcal{Y}} \to [0,1]$ is $(\mathcal{H}, \varepsilon)$-multiaccurate, then $\tilde{p}$ is $(\mathcal{L}_{\text{io}}, \mathcal{H}, 2\varepsilon)$-performative OI.*

Proof. The proof shows an approximate equality between the loss $\ell \in \mathcal{L}_{\text{io}}$ of any $h \in \mathcal{H}$ under $\tilde{y} \sim \tilde{p}(x, h(x))$ and $y^* \sim p^*(x, h(x))$. For $\mathcal{Y} = \{0,1\}$,

$$\begin{aligned}
\mathop{\mathbb{E}}_{\substack{x \sim \mathcal{D} \\ \tilde{y} \sim \tilde{p}(x, \widehat{y})}} [\ell(h(x), \tilde{y})] &= \sum_{\widehat{y} \in \widehat{\mathcal{Y}}} \Pr_{\mathcal{D}}[h(x) = \widehat{y}] \cdot \mathop{\mathbb{E}}_{\substack{x \sim \mathcal{D} \\ \tilde{y} \sim \tilde{p}(x, \widehat{y})}} [\ell(\widehat{y}, \tilde{y}) \mid h(x) = \widehat{y}] \\
&= \sum_{\widehat{y} \in \widehat{\mathcal{Y}}} \Pr_{\mathcal{D}}[h(x) = \widehat{y}] \cdot \mathop{\mathbb{E}}_{\substack{x \sim \mathcal{D} \\ \tilde{y} \sim \tilde{p}(x, \widehat{y})}} [\tilde{y} \cdot \ell(\widehat{y}, 1) + (1 - \tilde{y}) \cdot \ell(\widehat{y}, 0) \mid h(x) = \widehat{y}] \\
&\leqslant \sum_{\widehat{y} \in \widehat{\mathcal{Y}}} \Pr_{\mathcal{D}}[h(x) = \widehat{y}] \cdot \mathop{\mathbb{E}}_{\substack{x \sim \mathcal{D} \\ y^* \sim p^*(x, \widehat{y})}} [y^* \cdot \ell(\widehat{y}, 1) + (1 - y^*) \cdot \ell(\widehat{y}, 0) \mid h(x) = \widehat{y}] + 2\varepsilon \\
&= \sum_{\widehat{y} \in \widehat{\mathcal{Y}}} \Pr_{\mathcal{D}}[h(x) = \widehat{y}] \cdot \mathop{\mathbb{E}}_{\substack{x \sim \mathcal{D} \\ y^* \sim p^*(x, \widehat{y})}} [\ell(\widehat{y}, y^*) \mid h(x) = \widehat{y}] + 2\varepsilon \\
&= \mathop{\mathbb{E}}_{\substack{x \sim \mathcal{D} \\ y^* \sim p^*(x, \widehat{y})}} [\ell(h(x), y^*)] + 2\varepsilon.
\end{aligned}$$

The the third line follows under the assumption that $\tilde{p}$ is $(\mathcal{H}, \varepsilon)$-multiaccurate and the bound on the magnitude of $|\ell(\widehat{y}, b)| \leqslant 1$ for all $\widehat{y} \in \widehat{\mathcal{Y}}$ and $b \in \{0,1\}$. Given that an identical argument can be used to show the opposite inequality, we conclude that $\tilde{p}$ is indeed POI. ∎

Inspecting the performative multiaccuracy condition, we can see that it is similarly possible to reduce the problem of auditing for multiaccuracy to supervised learning. This auditing

procedure can be viewed as a special case of the auditing step from Section 4 or as a generalization of previous auditing procedures from work on multiaccuracy in the supervised learning setting [HKRR18, KGZ19]. In more detail, the relevant auditing problem for performative multiaccuracy is to determine whether there exists an $h \in \mathcal{H}$ and $\widehat{y} \in \widehat{\mathcal{Y}}$ such that

$$\Big| \mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D} \\ y^*\sim p^*(x,\widehat{y})}} [(y^* - \tilde{p}(x,\widehat{y})) \cdot \mathbf{1}\{h(x) = \widehat{y}\}] \Big| > \varepsilon.$$

As before, this auditing step reduces to a cost-sensitive classification problem (Definition 4.6). Auditing over a hypothesis class $\mathcal{H}$ can be done with $2|\widehat{\mathcal{Y}}|$ many calls to a cost-sensitive learner $\mathcal{A}_{\text{csc}}$ with tolerance parameter $\mathcal{O}(\varepsilon)$. We omit a formal statement of this result since it follows the exact pattern from Proposition 4.7.

In other words, in order to achieve $(\mathcal{L}, \mathcal{H}, O(\varepsilon))$-POI for any class of input-oblivious losses $\mathcal{L} \subseteq \mathcal{L}_{\text{io}}$, it suffices to audit for and enforce $(\mathcal{H}, \varepsilon)$-multiaccuracy. In this sense, for input-oblivious losses, there is a single auditing procedure that works for all losses, so we can replace $|\mathcal{L}|$-factors by $|\widehat{\mathcal{Y}}|$ factors in the auditing complexity for performative OI.

On Multicalibration. Going beyond multiaccuracy, the original work of [GKR+22] established omniprediction in the supervised setting as a consequence of multicalibration. As such, we might wonder whether there exists an analogous notion of multicalibration for performative prediction that enables a similar result. Defining an efficient notion of calibration under performativity (let alone, multicalibration), turns out to be a subtle task.

In supervised learning, calibration requires that the expectation of $\tilde{p}$ is accurate, even when we partition the inputs $x \in \mathcal{X}$ based on the predicted value $\tilde{p}(x) = v$. Specifically, the constraints quantify over each supported $v \in \text{supp}(\tilde{p})$:

$$\mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D} \\ y^*\sim p^*(x)}} [y^* \cdot \mathbf{1}\{\tilde{p}(x) = v\}] \approx \mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D} \\ \tilde{y}\sim\tilde{p}(x)}} [\tilde{y} \cdot \mathbf{1}\{\tilde{p}(x) = v\}] = v \cdot \Pr[\tilde{p}(x) = v].$$

Such calibration-style constraints suffice to establish omniprediction because for any loss ℓ, the optimal decision $\tilde{f}_\ell(x)$ is completely determined by $\tilde{p}(x)$.

In performative prediction, quantifying over the supported values of $\tilde{p}$ requires considering the decisions $\widehat{y} \in \widehat{\mathcal{Y}}$ as well. In particular, the optimal decision $\tilde{f}_\ell(x)$ is a function of the vector of predictions $\tilde{q}(x) \in [0,1]^{|\widehat{\mathcal{Y}}|}$, which gives the predicted probability $\tilde{y} \sim \tilde{p}(x, \widehat{y})$ for each $\widehat{y} \in \widehat{\mathcal{Y}}$ (recall the $q(\cdot)$ notation from Proposition 4.1). Thus, using the naive translation of the calibration constraints for omniprediction, we must partition $\mathcal{X}$ based on the vector-valued predictions, $\tilde{q}(x) = \overrightarrow{v}$. The cardinality of this calibration partition of $\mathcal{X}$ scales exponentially in the number of decisions $|\widehat{\mathcal{Y}}|$, even for binary outcomes $\mathcal{Y}$.

Still, we may consider more efficient calibration-style conditions that suffice to imply omniprediction in the performative setting. Rather than aiming for full performative calibration, we focus on adapting the notion of decision calibration [ZKS+21] to the performative setting. Decision calibration was introduced to avoid exponential blow-up in the calibration constraints due to multi-class prediction. We show that the notion can equally be adapted to the performative setting to deal with blow-up due to many actions $\widehat{y} \in \widehat{\mathcal{Y}}$. We define decision calibration with respect to the class of input-oblivious losses.

Definition 5.3 (Decision Calibration)**.** *For a distribution $\mathcal{D}$ and $\varepsilon \geqslant 0$, a predictor $\tilde{p} : \mathcal{X} \times \widehat{\mathcal{Y}} \to [0,1]$*

is ε-decision calibrated under outcome performativity over $\mathcal{D}$ if for every loss $\ell \in \mathcal{L}_{\text{io}}$, and for all $\widehat{y} \in \widehat{\mathcal{Y}}$,

$$\Big| \mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D} \\ y^*\sim p^*(x,\tilde{f}_\ell(x))}} [y^* \cdot \mathbf{1}\{\tilde{f}_\ell(x) = \widehat{y}\}] - \mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D} \\ \tilde{y}\sim \tilde{p}(x,\tilde{f}_\ell(x))}} [\tilde{y} \cdot \mathbf{1}\{\tilde{f}_\ell(x) = \widehat{y}\}] \Big| \leqslant \varepsilon.$$

With this definition in place, the proof of Lemma 5.2 can be adapted to show that decision calibration suffices to establish performative decision OI.

Lemma 5.4. *If $\tilde{p} : \mathcal{X} \times \widehat{\mathcal{Y}} \to [0,1]$ is ε-performative decision calibrated, then $\tilde{p}$ is $(\mathcal{L}_{\text{io}}, 2\varepsilon)$-performative decision OI.*

As an immediate corollary of Theorem 2.7 and Lemmas 5.2 & 5.4, we obtain sufficient conditions for omniprediction with respect to all bounded input-oblivious losses.

Corollary 5.5. *Suppose $\tilde{p} : \mathcal{X} \times \widehat{\mathcal{Y}} \to [0,1]$ is $(\mathcal{H}, \varepsilon)$-multiaccurate and ε-decision calibrated under outcome performativity. Then, $\tilde{p}$ is a $(\mathcal{L}_{\text{io}}, \mathcal{H}, 4\varepsilon)$-performative omnipredictor.*

In other words, if we can obtain multiaccuracy and decision calibration under performativity, then we have a direct pathway to obtain omniprediction for all bounded, input-oblivious losses.

On Decision Calibration. Note, however, that unlike the case of multiaccuracy, the decision rules that arise in the decision calibration condition are loss-dependent. That is, the optimal decision rules $\tilde{f}_\ell$ depend on $\ell \in \mathcal{L}_{\text{io}}$.

Motivated by the strong guarantee, we consider the feasibility of auditing for decision calibration. Note that for any $\ell \in \mathcal{L}_{\text{io}}$, the loss is defined by the loss for each outcome $y \in \{0,1\}$ and decision $\widehat{y} \in \widehat{\mathcal{Y}}$. Thus, to audit decision calibration over all input-oblivious losses, it suffices to audit whether there exist a $\widehat{y} \in \widehat{\mathcal{Y}}$ and $w_{a,0}, w_{a,1} \in [-1,1]$ such that:

$$\Big| \mathop{\mathbb{E}}_{x}[(\tilde{p}(x,\widehat{y}) - y^*) \cdot \mathbf{1}\{\arg\min_{a\in\widehat{\mathcal{Y}}}\{w_{a,1} \cdot \tilde{p}(x,a) + w_{a,0} \cdot (1 - \tilde{p}(x,a))\} = \widehat{y}\}] \Big| > \varepsilon$$

where the weights $w_{a,0}$ and $w_{a,1} \in [-1,1]$ represent the choice of $\ell(a,0)$ and $\ell(a,1)$ corresponding to the choice of $a \in \widehat{\mathcal{Y}}$.

Naively, searching for such a violated loss might require time exponential in $|\widehat{\mathcal{Y}}|$. For instance, by explicitly enumerating over some appropriately-fine net of $[-1,1]^{2|\widehat{\mathcal{Y}}|}$, then we can simply consider "every" possible loss. Improving the computational complexity of such a search is an interesting question, which may benefit from the techniques utilized in [ZKS+21].

Even without an improvement in the runtime complexity of learning, note that once an auditor succeeds, and we have a violated loss function, there is an efficient update to the prediction function. In particular, we simply need to record the chosen $2 \cdot |\widehat{\mathcal{Y}}|$ parameters $\{w_{a,0}, w_{a,1}\}$ and the $\widehat{y} \in \widehat{\mathcal{Y}}$, then execute the update from POI-Boost. In all, we can conclude that performative omnipredictors for $\mathcal{L}_{\text{io}}$ exist in complexity independent of the complexity of $\mathcal{L}_{\text{io}}$.

Corollary 5.6. *Suppose $\mathcal{H}$ is a hypothesis class with size-s circuits. Then, for any $\varepsilon > 0$, there exist an $(\mathcal{L}_{\text{io}}, \mathcal{H}, \varepsilon)$-performative omnipredictor implemented by a circuit of size* $\text{poly}(s, |\widehat{\mathcal{Y}}|)/\varepsilon^2$.

This preliminary analysis leaves open the possibility of learning performative omnipredictors via techniques that are independent of the loss class, as in the original work on $(\mathcal{L}_{\text{cvx}}, \mathcal{H})$-omniprediction from $\mathcal{H}$-multicalibration. We leave a more thorough investigation of these ideas to future work.

Acknowledgments

The authors thank Parikshit Gopalan, Moritz Hardt, Celestine Mendler-Dunner, Omer Reingold, and Tijana Zrnic for helpful discussions throughout the development of the project. **MPK** is supported by the Miller Institute for Basic Research in Science and, in part, by the Simons Collaboration on Algorithmic Fairness.

References

[AHK+14] Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pages 1638–1646. PMLR, 2014.

[AZL04] Naoki Abe, Bianca Zadrozny, and John Langford. An iterative method for multi-class cost-sensitive learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 3–11, 2004.

[BB19] Robert Balfanz and Vaughan Byrnes. Early warning indicators and intervention systems: State of the field. *Handbook of student engagement interventions*, pages 45–55, 2019.

[BHK22] Gavin Brown, Shlomi Hod, and Iden Kalemaj. Performative prediction in a stateful world. In *International Conference on Artificial Intelligence and Statistics*, pages 6045–6061. PMLR, 2022.

[BLR09] Alina Beygelzimer, John Langford, and Pradeep Ravikumar. Error-correcting tournaments. In *International Conference on Algorithmic Learning Theory*, pages 247–262. Springer, 2009.

[CDH21] Joshua Cutler, Dmitriy Drusvyatskiy, and Zaid Harchaoui. Stochastic optimization under distributional drift. *arXiv preprint arXiv:2108.07356*, 2021.

[DHK+11] Miroslav Dudik, Daniel Hsu, Satyen Kale, Nikos Karampatziakis, John Langford, Lev Reyzin, and Tong Zhang. Efficient optimal learning for contextual bandits. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, page 169–178, Arlington, Virginia, USA, 2011. AUAI Press.

[DKR+21] Cynthia Dwork, Michael P. Kim, Omer Reingold, Guy N. Rothblum, and Gal Yona. Outcome indistinguishability. In *ACM Symposium on Theory of Computing (STOC'21)*, 2021.

[DKR+22] Cynthia Dwork, Michael P. Kim, Omer Reingold, Guy N. Rothblum, and Gal Yona. Beyond bernoulli: Generating random outcomes that cannot be distinguished from nature. In *International Conference on Algorithmic Learning Theory*, pages 342–380. PMLR, 2022.

[DX22] Dmitriy Drusvyatskiy and Lin Xiao. Stochastic optimization with decision-dependent distributions. *Mathematics of Operations Research*, 2022.

[Elk01] Charles Elkan. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, volume 17, pages 973–978. Lawrence Erlbaum Associates Ltd, 2001.

[GHK+23] Parikshit Gopalan, Lunjia Hu, Michael P. Kim, Omer Reingold, and Udi Wieder. Loss minimization through the lens of outcome indistinguishability. In *ITCS*, 2023.

[GKR+22] Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Omnipredictors. In *ITCS*, 2022.

[HJM22] Moritz Hardt, Meena Jagadeesan, and Celestine Mendler-Dünner. Performative power. *arXiv preprint arXiv:2203.17232*, 2022.

[HKRR18] Ursula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1939–1948. PMLR, 2018.

[HMPW16] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pages 111–122, 2016.

[IYZ21] Zachary Izzo, Lexing Ying, and James Zou. How to learn when data reacts to your model: performative gradient descent. In *International Conference on Machine Learning*, pages 4641–4650. PMLR, 2021.

[IZY22] Zachary Izzo, James Zou, and Lexing Ying. How to learn when data gradually reacts to your model. In *International Conference on Artificial Intelligence and Statistics*, pages 3998–4035. PMLR, 2022.

[JLP+21] 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.

[JZM22] Meena Jagadeesan, Tijana Zrnic, and Celestine Mendler-Dünner. Regret minimization with performative feedback. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 9760–9785. PMLR, 17–23 Jul 2022.

[KAH+17] Akshay Krishnamurthy, Alekh Agarwal, Tzu-Kuo Huang, Hal Daumé III, and John Langford. Active learning for cost-sensitive classification. In *International Conference on Machine Learning*, pages 1915–1924. PMLR, 2017.

[KGZ19] 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.

[KKG+22] Michael P. Kim, Christoph Kern, Shafi Goldwasser, Frauke Kreuter, and Omer Reingold. Universal adaptability: Target-independent inference that competes with propensity scoring. *Proceedings of the National Academy of Sciences*, 119(4):e2108097119, 2022.

[KLMO15] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. Prediction policy problems. *American Economic Review*, 105(5):491–95, 2015.

[KNRW18] 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. PMLR, 2018.

[KRR18] Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Fairness through computationally-bounded awareness. *Advances in Neural Information Processing Systems*, 31, 2018.

[LB05] John Langford and Alina Beygelzimer. Sensitive error correcting output codes. In *International Conference on Computational Learning Theory*, pages 158–172. Springer, 2005.

[MBBF99] Llew Mason, Jonathan Baxter, Peter Bartlett, and Marcus Frean. Boosting algorithms as gradient descent. *Advances in Neural Information Processing Systems*, 12, 1999.

[MDW22] Celestine Mendler-Dünner, Frances Ding, and Yixin Wang. Predicting from predictions. *arXiv preprint arXiv:2208.07331*, 2022.

[MMH20] John Miller, Smitha Milli, and Moritz Hardt. Strategic classification is causal modeling in disguise. In *International Conference on Machine Learning*, pages 6917–6926. PMLR, 2020.

[MPZ21] John P Miller, Juan C Perdomo, and Tijana Zrnic. Outside the echo chamber: Optimizing the performative risk. In *International Conference on Machine Learning*, pages 7710–7720. PMLR, 2021.

[MPZH20] Celestine Mendler-Dünner, Juan Perdomo, Tijana Zrnic, and Moritz Hardt. Stochastic optimization for performative prediction. *Advances in Neural Information Processing Systems*, 33:4929–4939, 2020.

[NFD+22] Adhyyan Narang, Evan Faulkner, Dmitriy Drusvyatskiy, Maryam Fazel, and Lillian J Ratliff. Multiplayer performative prediction: Learning in decision-dependent games. *arXiv preprint arXiv:2201.03398*, 2022.

[PGV+18] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31, 2018.

[PZMH20] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning*, pages 7599–7609. PMLR, 2020.

[SB14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[US 16] US Department of Education. Issue brief: Early warning systems, 2016. Available at https://www2.ed.gov/rschstat/eval/high-school/early-warning-systems-brief.pdf.

[ZKS+21] Shengjia Zhao, Michael P. Kim, Roshni Sahoo, Tengyu Ma, and Stefano Ermon. Calibrating predictions to decisions: A novel approach to multi-class calibration. *Advances in Neural Information Processing Systems*, 34:22313–22324, 2021.