Predicting from Predictions

Anonymous Author(s)
Affiliation
Address
email

Abstract

Predictions about people, such as their expected educational achievement or their credit risk, can shape the outcome that they aim to predict. Estimating the causal effect of these predictions is important for deciding which predictive models to deploy. However, this estimation poses unique challenges because model predictions are usually deterministic functions of model inputs (making the effects of predictions difficult to disentangle from the effects of other inputs), and predictions are also highly correlated with outcomes. In this work we show that any of the three following conditions are sufficient to identify the causal effect of predictions: overparameterization of the predictive model, randomization in released predictions, and measurement noise. Under these conditions, standard supervised learning implicitly estimates the causal effect of predictions and finds transferable functional relationships to anticipate outcomes from a new model deployment, *as long as model predictions are provided as a feature*. Empirically, we find that supervised learning succeeds even in the presence of mild violations of our assumptions, such as finite data and model misspecification. Lastly, we discuss how this approach, unlike standard individualistic models, is able to handle certain types of interference (people affecting the outcomes of others). These findings emphasize the importance of recording model predictions during deployment, in order to understand social outcomes and feedback loops.

1 Introduction

Predictions can impact sentiments, alter expectations, inform actions, and thus change the course of events. Through their influence on people, predictions have the potential to change the regularities in the population they seek to describe and understand. This insight underlies the theory of performativity [26] and reflexivity [41] that play an important role in modern economics and finance.

Recently, Perdomo et al. [35] pointed out that the social theory of performativity has important implications for machine learning practice. Classical supervised learning assumes that features X and labels Y are sampled jointly from a fixed underlying data distribution that is unaffected by attempts to predict Y from X. Performativity questions this assumption and suggests that the deployment of a predictive model f_θ can disrupt the relationship between X and Y, so that changes to the model f_θ induce distribution shifts. For example, consider a lender with a predictive model for risk of default – performativity could arise if individuals who are predicted as likely to default are given higher interest loans, which make default even more likely [29]. Changing the predictive model and offering more low interest loans could cause some individuals who previously looked risky to be able to pay the loans back, which would appear as a shift in the relationship between features X and loan repayment outcomes Y. This shift poses an important challenge for using historical data to predict the outcomes that will arise under the deployment of a future model.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

1.1 Our work

In this work we study prediction in performative environments under the assumption that in addition to data samples (X, Y) we are also given access to the predictive model f_θ that was deployed when collecting the training data. Using this data, our goal is to learn about the causal relationship between X, Y and the predictions $\hat{Y} = f_\theta(X)$ in order to anticipate the effects of a model update. More precisely, our goal is to build a meta machine learning model that accurately predicts the following: given a new predictive model f_ϕ to be deployed on users, what will be the outcome Y for an individual with features X, subjected to the new prediction $f_\phi(X)$? We term this data-driven strategy *predicting from predictions* because it treats the predictions as an input to the meta machine learning model.

The meta model seeks to answer "what would the outcome be if we were to deploy a different prediction model." Crucially, this "what if" question is causal in nature; it aims to understand the potential outcome from intervening on the world, where the intervention is deploying a prediction model different from the one in the historical data. Thus, our goal is to learn a causal model that can draw conclusions about the impact of unseen models, not merely fit previously seen data. Whether such a transferable model is learnable depends on whether the training data provides causal *identifiability* [34], i.e. the causal relationship can be estimated from training data. Once causal identifiability is established, it gives us the green light to estimate causal relationships from observational training data (under the corresponding causal identifiability conditions); these causal conclusions enable downstream selection of optimal prediction models under performativity.

For our main technical results, we start by showing the hardness of identifying the causal effect of an individual's prediction on their outcome from training data collected under the deployment of a deterministic prediction function f_θ. Then, we establish three natural conditions that are each sufficient to render these effects identifiable. The first condition exploits the property that predictive models are often over-parameterized, which leads to incongruence between the different causal paths. The second condition exploits the presence of randomization in the downstream decision based on the prediction. The third condition exploits side-information about measurement errors in the features that lead to noise in the predictions. Together these three conditions show that natural inaccuracies and particularities of the performative prediction problem are sufficient for identifiability such that standard supervised learning implicitly finds transferable functional relationships, *as long as model predictions are provided as a feature*. Empirically, we find that supervised learning succeeds even in the presence of mild violations of our assumptions, such as finite data and model misspecification.

We conclude with a discussion of interference in the context of prediction and highlight the SUTVA assumption underlying our causal analysis. We argue that spillover effects will implicitly be accounted for in our observation-driven model, whereas typical microfoundation models fail to account for this.

1.2 Broader Context and Related Work

The work by Perdomo et al. [35], proposing the framework of *performative prediction*, provided inspiration for this project. The authors introduce the concept of performativity to supervised learning and show that optimizing a predictive model should take into account actions and reactions of individuals. Cast in the language of machine learning, performativity gives rise to a non-stationary risk minimization problem where the data distribution depends on the predictive model being optimized. Prior work has studied the convergence of retraining methods in this setting [35, 31, 7] or aimed to find models that are globally optimal ex-post [32, 19, 21]. Our work differs from this scholarship in that we do not focus on achieving a particular solution concept with a prediction function that maps X to Y, such as performative stability or performative optimality. Instead, we are interested in understanding the underlying causal mechanism of the distribution shifts in order to reason about the implications of any potential new model deployment on the users.

In our work we focus on performative effects where predictions lead to changes in the relationship between covariates X and label Y, while the distribution over covariates is fixed. This describes a subclass of problems falling under the umbrella of performative (aka. model-induced or decision-dependent) shifts [35, 25, 7]. It is distinct from strategic classification [14, 2, 16], which focuses on performative effects that impact the distribution over covariates, while $P(Y|X)$ remains stable.

Our work builds on the literature of causal inference [33, 43] and extends classical causal identification results to provide a novel perspective on performative prediction. In particular, our problem has unique properties we take advantage of to design practical assumptions. We also connect to the literature on causal inference with interference to discuss spill-over effects of predictions [39, 43], a phenomenon

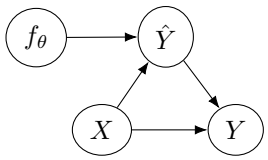
that has so far been ignored in model-based approaches to performative prediction [13, 20, 32, 38, 17]. A notable exception is [36] that discusses peer effects in the context of strategic classification.

2 The causal force of prediction

Predictive models and their predictions can impact the population of individuals they aim to predict. Through the lens of causal inference [33], the deployment of a predictive model poses an intervention on a causal diagram. We will build on this perspective to study the performative prediction problem, its particularities, and unique challenges form a distinct angle.

2.1 Prediction as a partial mediator

Generally, the goal of machine learning is to predict an outcome y from features x, using a predictive model $f_\theta \in \mathcal{F}$. This model is trained to fit patterns in historical data. The insight of performativity is that when f_θ is deployed, the prediction $\hat{y} = f_\theta(x)$ can cause a shift in the outcome variable y. This is not accounted for in traditional supervised learning, which assumes that the joint distribution of (x, y) is fixed. A causal diagram illustrating the setting we consider is below:



$$X = \xi_X \qquad \xi_X \sim \mathcal{D}_X \tag{1}$$
$$\hat{Y} = f_\theta(X) \tag{2}$$
$$Y = g(X, \hat{Y}) + \xi_Y \qquad \xi_Y \sim \mathcal{D}_Y \tag{3}$$

Figure 1: Performative effects through prediction under SUTVA.

The features $X \in \mathcal{X} \subseteq \mathbb{R}^d$ are drawn i.i.d. from a fixed underlying distribution over covariates $\mathcal{D}_X$ that we assume is continuous. The prediction $\hat{Y} = f_\theta(X)$ is determined by the deployed model f_θ. The outcome $Y \in \mathcal{Y} \subseteq \mathbb{R}$ is a function of X, partially mediated by the prediction $\hat{Y}$. We assume ξ_Y is a zero mean noise variable. This setup differs from the traditional supervised learning setting by including the causal arrow between $\hat{Y}$ and Y in the causal graph. Note that this setup is not meant to describe performativity in its full generality (which includes other ways f_θ may affect X and/or Y), but highlights an important class of performativity that we focus on in this work.

Application examples. The phenomenon that predictions or expectations about an event impact the outcome of the event has been documented in various contexts. A bank's prediction of the creditworthiness of a client applying for a loan determine the interest rate assigned to them, which in turn changes a client's financial situation [29]. Mathematical models that predict stock prices inform actions of traders and thus heavily shape financial markets [26]. Zillow's housing price predictions directly impact sales prices [27]. Prediction about the severity of an illness play an important role in treatment decisions and hence the very chance of survival of the patient [24]. Another prominent example from psychology is the Pygmalion effect [37]. It refers to the phenomenon that high expectations lead to improved performance, which is widely documented in the context of education [3], sports [40], organizations [10], and social structures. Examples of such performativity abound, and we hope to have convinced the reader that the effects described by the causal graph in Figure 1 are widely documented and should be a concern for algorithmic predictions.

2.2 Hardness of performativity-agnostic learning

Begin with considering the classical supervised learning task of predicting the label Y from the features X. Under the structural equation model from Figure 1 it follows that

$$P(Y|X) = \int P(Y|\hat{Y}, X) P(\hat{Y}|X) \mathrm{d}\hat{Y}. \tag{4}$$

This illustrates the inherent challenge of prediction under performativity: the causal relationship between X and Y depends on the predictive model governing $P(\hat{Y}|X)$. Updates to the predictive model f_θ can change $P(\hat{Y}|X)$ and hence $P(Y|X)$. Denote the predictive model deployed when collecting the training data as f_θ and the predictive model we aim to make inference about as f_ϕ. If

the predictions of the two models are different, we can have

$$P(Y|X, f_\theta) \neq P(Y|X, f_\phi),$$

which is known as concept shift [11] in the supervised learning literature. Such extrapolation problems are known to be intractable without structural assumptions, implying that we can not expect our model to generalize out of distribution, to data induced by a different predictive model. However, in contrast to the classical concept drift problem, we know that the distribution shift is solely due to a change in the predictive model. Liu et al. [25] have shown that this assumption allows for more fine-grained error bounds for domain adaptation compared to classical results by Mansour et al. [30]. Let us inspect the extrapolation gap in more detail and put existing positive results into perspective.

Extrapolation error. To illustrate the effect of performativity on predictive performance with a concrete example, we consider the following simple structural equation model with a linear performative effect of strength $\alpha > 0$ and a base function $g_1 : \mathcal{X} \to \mathcal{Y}$

$$g(X, \hat{Y}) := g_1(X) + \alpha \hat{Y}. \tag{5}$$

Now, assume we collect data under the deployment of a predictive model f_θ and we want to make inferences about a test distribution under the deployment of f_ϕ. Using our running example of a lender predicting risk of default, the lender may have historical data about individuals who defaulted or not, under a previous predictive model f_θ for risk that determined interest rates. The lender may want to learn a meta-model h that can predict whether similar individuals will default if a different policy for interest rates f_ϕ is deployed.

Adopting the notion of a distribution map from [35], we write $\mathcal{D}_{X,Y}(f)$ for the marginal distribution over (X, Y) surfacing from the deployment of a model f. We assess the quality of our model $h : \mathcal{X} \to \mathcal{Y}$ over a distribution $\mathcal{D}_{X,Y}(f)$ induced by f via the loss function ℓ and write $\mathrm{R}_f(h) := \mathrm{E}_{x,y\sim\mathcal{D}_{X,Y}(f)}\ell(h(x), y)$ for the risk, and $\psi(f)$ for the risk minimizer $\psi(f) := \operatorname{argmin}_h \mathrm{R}_f(h)$. The following result shows that the extrapolation loss of $\psi(f_\theta)$ grows with the strength of performativity and the distance between f_θ and f_ϕ as measured in prediction space.

Proposition 1. *Consider the data generation process in Figure 1 with g given in (5). Let $\psi(f_\theta)$ be the risk minimizer over the training distribution and assume the problem is realizable. Assume the loss function ℓ is γ-smooth and μ-strongly convex. Then, we can bound the extrapolation loss on a test distribution induced by f_ϕ as*

$$\frac{\gamma}{2}\alpha^2 d^2_{\mathcal{D}_X}(f_\theta, f_\phi) \geq R_{f_\phi}(\psi(f_\theta)) \geq \frac{\mu}{2}\alpha^2 d^2_{\mathcal{D}_X}(f_\theta, f_\phi) \tag{6}$$

where $d^2_{\mathcal{D}_X}(f_\theta, f_\phi) := \mathrm{E}_{x\sim\mathcal{D}_X}(f_\theta(x) - f_\phi(x))^2$.

The extrapolation gap is zero if and only if the strength of performativity tends to zero ($\alpha \to 0$), or the predictions of the two predictors f_θ and f_ϕ are identical over the support of $\mathcal{D}_X$. If this is not the case, an extrapolation gap is inevitable.

The special case where $\alpha = 0$ recovers the assumption of classical supervised learning, in which there is no performativity. It may hold if the predictive model is solely used for descriptive purposes, or if the agency making the prediction does not enjoy any economic power [15]. However, in practice, the strength of performative effects is not a parameter we can influence as machine learning practitioners. Thus, our goal is to anticipate the effects of any prediction that is likely to be performative.

The second special case where the extrapolation error is small is when $d^2_{\mathcal{D}_X}(f_\theta, f_\phi) = 0$. This case has been exploited in prior work on performative prediction [35, 31] in order to achieve positive results without strong assumptions on the distribution shift. In particular, the authors studied prespecified retraining strategies for f_θ which allowed them to identify local optima for which $f_\theta \approx f_\phi$, and the extrapolation gap diminishes under appropriate regularity conditions. In this work, instead of taking the perspective of a decision-maker minimizing the extrapolation gap, we aim to build a causal understanding of performative effects and anticipate the impacts of potential predictive model deployments, when extrapolation is required.

3 Extrapolating from observational data

Realizing the hardness of performativity-agnostic learning, we explore how to incorporate the presence of performative predictions into the learning task. As such, the natural first step is to log

f_θ when collecting data and make $\hat{Y}$ in Figure 1 observable. This allows us to build on valuable structure in the prediction function towards addressing our ultimate goal: *we want to predict the impact of a new model deployment on the population before we actually put it in practice*.

3.1 Problem setup

We are given access to $\mathcal{D}_{X,Y}(f_\theta)$ as well as the prediction function f_θ, meaning we can infer $\hat{Y}$ for every individual. From this observational data we wish to estimate the potential outcome of an individual under the deployment of an unseen (but known) predictive model f_ϕ. More precisely, we are interested in identifying the following causal quantity[1]:

$$h^*(x, \hat{y}) := \mathbb{E}[Y|X = x, \mathrm{do}(\hat{Y} = \hat{y})]. \tag{7}$$

Being able to estimate h^* will enable both policy evaluation (evaluating the effects of a new prediction rule when deployed on a population) and policy optimization (as a decision maker, finding the prediction rule that minimizes the risk under performativity.)

The supervised learning approach. With observable predictions, we can use tools from supervised learning to predict Y from $X, \hat{Y}$ to obtain

$$h_{\mathrm{SL}} = \underset{h \in \mathcal{H}}{\mathrm{argmin}} \ \mathrm{E}_{(x,y)\sim\mathcal{D}_{X,Y}(f_\theta)}[w(x, \hat{y}, y) \ \ (h(x, \hat{y}) - y)^2] \quad \text{with} \quad \hat{y} = f_\theta(x). \tag{8}$$

where $w(x, \hat{y}, y) > 0$ is some reweighing function, and $\mathcal{H}$ denotes our hypothesis class. Note that we transformed the concept shift problem from Section 2.2 into a covariate shift problem. Defining $X' = (X, \hat{Y})$ as the concatenation of an individual's features and their prediction, the conditional probability $P(Y|X')$ is now constant, but $P(X')$ changes with the prediction rule. Thus, without causal identification guarantees, we can not be sure that h_{SL} recovers the causal quantity h^* that allows us to predict from new predictions; there might be other models h' that explain the training data equally well but do not transfer to the distribution induced by the deployment of a new model.

Issue of overlap. The main challenge for identification of h^* is that in general, the prediction rule f_θ is a deterministic function of the covariates. This means that for any X we only get access to one particular $\hat{Y} = f_\theta(X)$ in the training distribution, making it challenging to disentangle the direct and the indirect effects of X on Y. The situation is akin to fitting linear regression against two perfectly colinear covariates; the regression coefficients are not identifiable. This problem is more broadly known as a *lack of overlap* (or lack of positivity) in the literature on causal inference [18], which persists as long as $P(X|\hat{Y} = \hat{y})$ and $P(X|\hat{Y} = \hat{y}')$ do not have common support.

Proposition 2 (Hardness of identification with deterministic predictions)**.** *Consider the structural equation model in Figure 1 where $X, \hat{Y}, Y$ are observed. For a deterministic prediction function f_θ we suffer lack of overlap, and thus the function $h^\star$ is not identifiable from observational data collected under the deployment of f_θ without any structural assumptions.*

3.2 Overlap in practical settings

The overlap issue persists as long as the two variables $X, \hat{Y}$ are deterministically bound. Below we investigate two natural scenarios in the context of prediction that yield sufficient randomness in the relationship between X and $\hat{Y}$ needed for identification of $h^*(x, \hat{y})$.

Releasing randomized decision for the purpose of differential privacy. Assume the prediction is released to the individual through a noisy proxy $T_{\hat{Y}} \in \mathcal{T}$ and performative effects only occur through $T_{\hat{Y}}$. The situation is described by the causal graph in Figure 2. In this case, classical results from causal inference tell us that overlap on $\mathcal{T}$ is sufficient to use front door adjustment to identify $\mathbb{E}[Y|\mathrm{do}(\hat{Y} = \hat{y}), X]$. In practice, overlap is easier to achieve if $|\mathcal{T}|$ is small (e.g., binary decisions).

Assumption 1 (Overlap)**.** *The released prediction $T_{\hat{Y}}$ satisfies $P(T_{\hat{Y},i} \in \mathcal{T}' \,|\, X_i = x) > 0$ for all $x \in \mathcal{X}$ and all sets $\mathcal{T}' \subseteq \mathcal{T}$ with positive measure, i.e. $P(\mathcal{T}') > 0$.*

[1] Instead of aiming for the conditional expectation of Y in (7) we could similarly aim for any function $\kappa(X, Y, \hat{Y})$ that measures social welfare, individual improvement, and other goals for socially beneficial machine learning. The criteria for identifiability would remain the same. See Appendix A for more discussion.

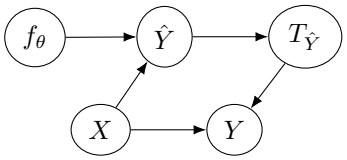


Figure 2: Decision released through $T_{\hat{Y}}$.

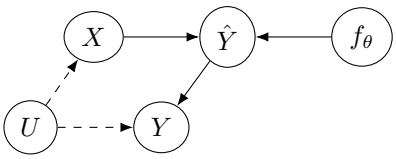


Figure 3: Noisy measurement of U.

Proposition 3. *Under Assumption 1, the function $h^*(x, \hat{y})$ is identifiable from observational data, and h_{SL} in (8) recovers h^* for $w(x, y) = 1$.*

Interestingly, in the context of predictions, a natural setting where Assumption 1 is satisfied is when f_θ implements an additive noise mechanism to preserve differential privacy [8]. For example, if the Laplace mechanism is used to release differentially private predictions, we have $T_{\hat{Y}} = f_\theta(X) + \eta$ where $\eta \sim \text{Lap}(0, b)$ for some $b > 0$ is sampled from a Laplace noise distribution centered at 0. This strategy ensures $P(T_{\hat{Y}} = t|X) > 0$ for all t which is sufficient for identification from observational data. Furthermore supervised learning works 'out-of-the-box' without sample reweighting.

Noisy measurement of covariates. The second setting where we can hope for identification is when the prediction is made based on covariates X that are noisy estimates of the underlying quality U that actually affects the outcome Y. Importantly, this underlying quantity does not need to be observed and $P(U)$ can be unknown. The setup is illustrated in Figure 3. For example, a student's college performance depends on their underlying scholastic ability, but predictions of performance (and perhaps admissions decisions) are made based on a noisy proxy like SAT score. In this setting we can achieve identification if we have information about the noise distribution of the measurement error $P(U|X = x)$ as the next result shows:

Proposition 4. *If the conditional densities $P(U|X = x)$ and $P(U|Y = y, \hat{Y} = \hat{y})$ are known and U satisfies overlap condition with respect to X then the function $\mathbb{E}[Y|\text{do}(\hat{Y} = \hat{y}), X]$ is identifiable from observational data, and h_{SL} recovers h^* for $w(x, \hat{y}) = P(U|X = x)/P(U|\hat{Y} = \hat{y})$.*

The result operates in a similar vein with the noise-induced randomization in regression discontinuity [9]. Both results exploit the measurement noise for identification. Proposition 4 tells us that we need to correctly adjust the weights in (8) based on the knowledge of the noise structure to recover h^*. The weights rely on knowledge of $P(U|X = x)$, although this can be relaxed using techniques from [9], see Appendix. In particlar, access to the heterogeneous noise bound $P(U|X = x)/P(U|X = x') \in [c_1(x, x'), c_2(x, x')]$ for some known functions c_1, c_2 is sufficient to obtain a consistent interval estimate for $\mathbb{E}[Y|\text{do}(\hat{Y} = \hat{y}), X]$.

3.3 Identification with separability

We discuss two more settings where we can achieve identification, even if the overlap assumption is violated. For these results, we assume the structural equation g describing the effect of X and $\hat{Y}$ on Y is separable. The following two results rely on incongruence between the effect of X on Y and the effect of X on $\hat{Y}$ that allows us to disentangle the two function components from observational data.

Overparameterization. We first consider the setting of incongruence in the functional complexity of the direct and the indirect causal effect of X on Y. This implies that $g_2(f_\theta(x))$ can not be expressed by any function $h \in \mathcal{H}$, and thus the model h_{SL} necessarily relies on $\hat{y}$ to describe the indirect link.

Proposition 5 (over-parameterized predictor)**.** *Consider the causal graph in Figure 1. Let $g(x, \hat{y}) = g_1(x) + g_2(\hat{y})$ where g_1 and $g_2 \circ f_\theta$ are linearly independent functions. Then, the risk minimizer of (8) with $w(x, y) = 1$ identifies h^*.*

An interesting setting where we can expect linear independence between the g_1 and $g_2 \circ f_\theta$ is when f_θ is overparameterized. Say f_θ represents a neural network trained on samples $(x, y) \sim \mathcal{D}(f_{\theta'})$, and is overparameterized in comparison to the underlying generative function $g'(x) = g_1(x) + g_2(f_\theta(x))$. Together with the natural assumption that performative shifts are simple and qualitatively different from the direct causal link (thus keeping $g_2 \circ f_\theta$ complex), this setting falls under the assumption of Proposition 5. Consider the following example to provide more intuition:

Example 3.1. *Let $g(x, \hat{y}) = \alpha x + \beta \hat{y}$ for some unknown α, β and let $\mathcal{H}$ bet the class of linear functions. The predictor f_θ has a quadratic term and is given by $f_\theta(x) = \gamma x^2 + \xi x$. Then, we can identify α, β from data points $(x, \hat{y}, y)$ by fitting a model $h(x, \hat{y}) = c_1 x + c_2 \hat{y}$ to the outcome y and inferring $\beta = c_2$ as well as $\alpha = c_1 - c_2$. This is possible despite lack of overlap because f_θ can not be expressed by a linear function, and the quadratic part is uniquely attributed to the indirect link.*

Discrete classification. A second natural source of incongruence which allows us to identify h^* is classification with discrete predictions. How to disentangle the continuous relationship between Y and X from the discrete relationship between X and $\hat{Y}$ has been studied in the context of regression discontinuity designs [23]. We will build on this literature for the following result.

Proposition 6 (discrete classification)**.** *Assume that the effect of X and $\hat{Y}$ on Y are separable $h^*(x, \hat{y}) = g_1(x) + g_2(\hat{y}), \forall x, \hat{y}$ for some differentiable functions g_1 and g_2. Further, suppose $\hat{Y}$ is a discrete random variable and X is a (possibly multi-dimensional) continuous random variable. Then, $h^*(x, \hat{y})$ is identifiable from observational data, and h_{SL} in (8) will recover h^* with $w(x, y) = 1$.*

The three results in the section illustrate why and when we could hope to answer our causal question with a model learned through supervised learning from observational data.

4 Social influence

The main theme of this work is about incorporating predictions $\hat{Y}$ into our meta model for how the outcome variable Y is generated. By relying on the causal graph in Figure 1 we have implicitly focused on explaining performative effects as the direct effect of a person's prediction on that person only. In this section we discuss a yet-unexplored dimension of performativity arising from indirect effects. Consider recommendations, which not only shape the consumption patterns of the individual they were targeted to, but through social media or verbal sharing, also affect friends of that individual.

In order to learn the causal effect of an intervention on individual behavior, it is common to perform simple, unilateral interventions on the prediction of a single individual and observe their response. See [4] for a representative field experiment. While this approach is sufficient for identifying the direct effect of a prediction $\hat{Y}_i$ for individual i on Y_i, it assumes away any kind of interference. This means spill-over effects will necessarily bias the estimator. This is akin to the important distinction between the individual causal effect and the overall causal effect in treatment effect estimation [42].

In the following we argue why ignoring such indirect effects might be less concerning when learning from historical data in performative settings, compared to building individualistic models offline. The intuition is that observational data collected under the deployment of f_θ we will implicitly contain information about indirect effects because the model has been deployed on the entire population. Formally, given a population of n individuals, the deployment of a predictive model f_θ results in n datapoints $\{Z_i\}_{i \in [n]}$ with

$$Z_i = \left(X_i,\ \hat{Y}_i,\ Y_i | \{X_i, \mathrm{do}((\hat{Y}_j)_{i \in [n]})\} \right),$$

where $\hat{Y}_i$ denotes the prediction $f_\theta(X_i)$ assigned to individual i resulting in label Y_i.

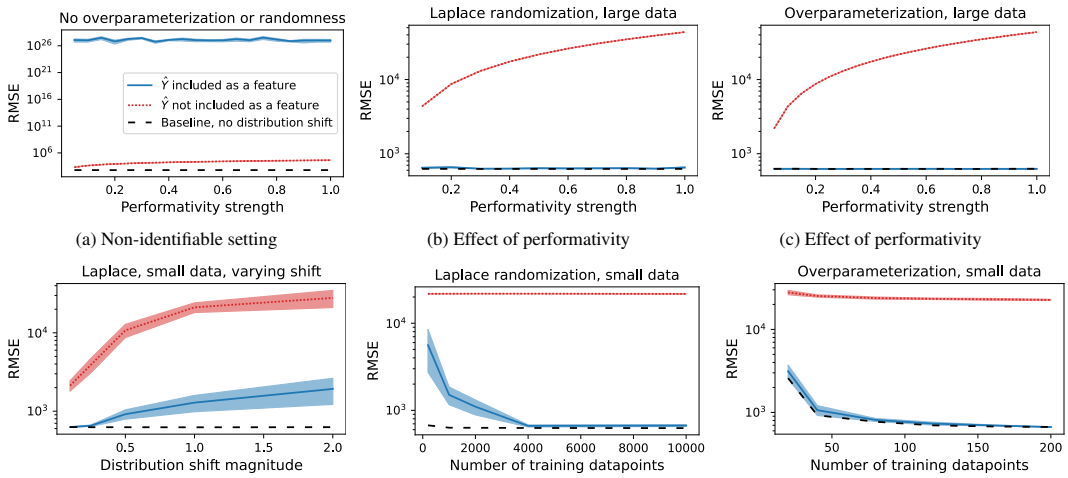
Now consider the following assumption that limits the dependency among units. It is inspired by the neighborhood interference assumption in [39] from the literature on peer effects in social networks: Let $N(i)$ denote the set of neighbors of individual i, then

$$Y_i \perp\!\!\!\perp \{\hat{Y}_j\}_{j \neq i} | G_i \quad \text{with} \quad G_i = \frac{1}{|N(i)|} \textstyle\sum_{j \in N(i)} \hat{Y}_j. \tag{9}$$

This assumes the spillover effect on individual i is mediated by the statistic G_i characterizing the interference pattern. Furthermore, interference depends on the average prediction in the neighborhood.

We highlight here that, in the context of predictions, *homophily* (the tendency for individuals to be similar to their neighbors) and smoothness of the prediction function together give us reason to expect strong correlations between outcomes of neighbors [12]. We propose the following assumption:

Assumption 2 (Homophily in predictions)**.** *Consider a network of n individuals and let $N(i)$ be the set of neighbors of individual i. Then, $|\mathbb{E}_{j \in N(i)} - \hat{Y}| \leq \delta$ for every $i \in [n]$ and some small $\delta > 0$.*



(a) Non-identifiable setting (b) Effect of performativity (c) Effect of performativity

(d) Effect of shift, with finite data (e) Effect of finite data (f) Effect of finite data

Figure 4: **Overparameterization and randomized decisions are each sufficient for learning models robust to distribution shift, even with limited data.** Shaded regions indicate standard error from 10 replicates with different random seeds. (a) Without any conditions to provide identifiability, RMSE is variable and large. With essentially infinite data (200,000 datapoints), randomized decisions (b) or overparameterization (c) along with access to $\hat{Y}$ leads to accurate prediction across different performativity strengths. (d) With finite data (1000 datapoints), increasing distribution shift magnitude increases the variance of h_{SL}. With moderate training dataset sizes, randomized decisions (e) and overparameterization (f) can achieve theoretically predicted accuracy.

Given Assumption 2, it follows that $\hat{Y}_i \approx G_i$, which implies that the interference effect can (at least partially) be explained by the prediction itself. This results in a machine learning model (8) that will implicitly pick up the interference effects from the training data in order to explain the total causal effect of f_θ on the outcome, which is helpful for prediction, despite our misspecified causal graph.

To illustrate the implications of Assumption 2, consider the following linear-in-means model studied by Manski [28], given as $Y_i = g(X_i) + \alpha\hat{Y}_i + \beta G_i$ for some $\alpha > \beta > 0$, where G_i is defined as in (9). This encodes a setting where the spillover effects are a weaker version of the direct effect, such as exposure probability to recommended content. Now, assume Assumption 2 holds for $\delta = 0$. Then, without explicitly measuring G_i, fitting a model to explain Y as a function of X and $\hat{Y}$ will produce $h(x, \hat{y}) = g(x) + (\alpha + \beta)\hat{y}$, which identifies $\mathbb{E}[Y|X, \text{do}((\hat{Y}_j = f_\theta(X_j)_{j\in[n]})]$ as desired.

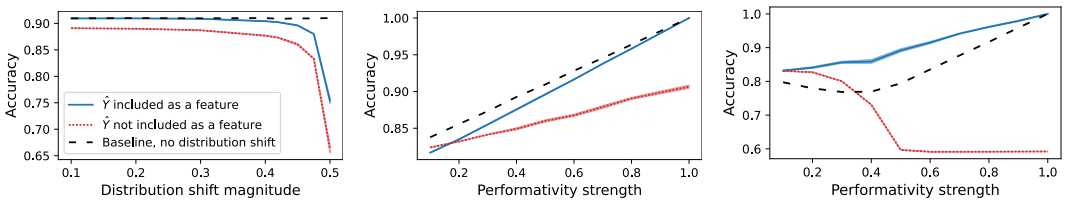
5 Empirical evaluation

In this section we investigate how well our identification strategy, training a supervised model, works in practice when faced with the limitations of finite data and model misspecifications. We generated semi-synthetic data for our experiments, using a Census income prediction dataset from `folktables.org` [6]. Appendix D contains qualitatively similar results on other Census datasets and the Kaggle credit scoring dataset [22], along with more experimental details. Using these datasets as starting points, we simulate a training dataset and test dataset with distribution shift as follows:

1. Choose two different predictors f_θ and f_ϕ to predict a feature of interest (e.g. income) from covariates X (e.g. age, occupation, education, etc.). In most of our experiments, f_θ is fit to the original dataset to minimize squared error, while f_ϕ is trained on random labels.
2. Posit a function g for simulating the performative effects.
3. Generate a *training* dataset of $(X, \hat{Y}, Y)$ tuples following the structural equation model in Figure 1, using X, g, f_θ. (This will overwrite the original labels Y in the original dataset).
4. Similarly generate a *test* dataset of $(X, \hat{Y}, Y)$ tuples following Figure 1, using X, g, f_ϕ.

We assess how well supervised methods learn transferable functional relationships by fitting a model h_{SL} to the training dataset and evaluating its accuracy on the test dataset.

Learning with finite data. Our theoretical results assume infinite training data, but supervised learning can still successfully learns transferable functions h_{SL} with few training data points. We design experiments to test randomized decisions with Laplace noise (Proposition 3) by setting g to



(a) Effect of larger distribution shifts (b) Standard setup: f_ϕ (test) fit to noisy labels (c) Reversed: f_θ (train) fit to noisy labels

Figure 5: **Performance degrades gracefully under model misspecification.** (a) Accuracy on the test distribution (higher is better) is plotted against distribution shift magnitude; supervised learning remains accurate until the train and test predictors, f_θ and f_ϕ, are uncorrelated (shift magnitude of 0.5). (b) Accuracy is plotted against performativity strength; despite model misspecification, accuracy is higher when $\hat{Y}$ is included as a feature, across most performativity strengths. (c) When the training set predictor f_θ is fit to random labels and is less accurate than f_ϕ, including $\hat{Y}$ as a feature universally improves accuracy.

be a linear function of X (with weights determined by a linear regression on the original dataset), summed with $\alpha\hat{Y}$, for hyperparameter α that quantifies the performativity strength. We optimize h_{SL} over linear functions, satisfying the assumptions in Proposition 3. To test overparameterization (Proposition 5), we additionally augment the input to f_θ with second-degree polynomial features to satisfy its assumptions. In addition, to varying training set size and strength of performativity α, we also vary the distance between predictions from f_θ and f_ϕ, i.e. distribution shift magnitude.

Figure 4(a) shows that without any conditions providing identifiability, our linear setting induces large, variable weights on the $\hat{Y}$ feature and h_{SL} has high error. Figure 4(b) and (c) show that with large data, supervised learning under randomized decisions or overparameterization succeeds in transferring out of distribution, matching in-distribution accuracy baselines. Figure 4(d) highlights that when data is limited, the error and variance of h_{SL} grows with the magnitude of distribution shift between f_θ and f_ϕ. Thus, for practitioners choosing which f_ϕ to deploy, performative effects are easier to estimate and anticipate if model updates are smaller. Figure 4(e) and (f) additionally show that only moderate dataset sizes are necessary for h_{SL} to be robust to distribution shift.

Learning despite model mis-specification. Our theoretical results in Section 3.3 depend on knowing the correct model class $\mathcal{H}$ to optimize h_{SL} on. Here we test the resiliency of discrete classification identifiability (Proposition 6) to model misspecification. Proposition 6 assumes that the underlying function g is a separable function of X and $\hat{Y}$; we break this assumption by constructing g as $g(x, \hat{y}) = g_1(x)$ with probability p, and $g(x, \hat{y}) = \hat{y}$ otherwise, where $g_1(x)$ maps x to its original label y in the original dataset, and $p \in [0, 1]$ is a hyperparameter for performativity strength that we vary. Similarly to the finite data experiments, we also vary the distance between predictions from f_θ and f_ϕ, i.e. distribution shift magnitude.

Figure 5(a) shows that despite misspecification, the accuracy of h_{SL} remains close to in-distribution accuracy as long as distribution shifts are not too large (specifically, until f_θ and f_ϕ become uncorrelated, supporting our recommendation to update models gradually for policy optimization). Figure 5(b) highlights that h_{SL} is more prone to errors from model misspecification when performativity is weak. This is intuitive, since $\hat{Y}$ is correlated with the outcome Y, and a misspecified h_{SL} may erroneously attribute this to a causal link; in such cases, Figure 5(b) suggests that accuracy is slightly improved by dropping $\hat{Y}$ as a feature. However, this improvement is small, and Figure 5(c) shows that if the accuracy of f_θ is reduced (by fitting to noisy labels), this effect disappears. Thus, across most settings, including $\hat{Y}$ as a feature vastly outperforms not including it; even when practitioners are not sure if performativity is present, it is worthwhile to input predictions $\hat{Y}$ to the meta model.

6 Conclusion

This paper proposes a supervised learning approach to tackle the problem of predicting under performativity. We draw connections to causal inference to highlight that there are various natural settings where such a data-driven approach can result in transferable models that predict the outcome of an individual for unseen deployable predictive models. We believe that one of the most important lesson from this work is that there is high value to logging the state of the deployed predictor when collecting data for the purpose of supervised learning. To date, such data is rarely available in benchmark datasets, hindering the progress towards a better understanding of these important phenomena, essential for the reliable deployment of algorithmic systems in the social world.

References

[1] Susan Athey. Machine learning and causal inference for policy evaluation. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, page 5–6, 2015.

[2] Yahav Bechavod, Katrina Ligett, Zhiwei Steven Wu, and Juba Ziani. Gaming helps! learning from strategic interactions in natural dynamics, 2020. URL https://arxiv.org/abs/2002.07024.

[3] Xander M. Bezuijen, Peter T. van den Berg, Karen van Dam, and Henk Thierry. Pygmalion and employee learning: The role of leader behaviors. *Journal of Management*, 35(5):1248–1267, 2009.

[4] Daniel Björkegren, Joshua E. Blumenstock, and Samsun Knight. Manipulation-proof machine learning, 2020.

[5] Alex Chin. Regression adjustments for estimating the global treatment effect in experiments with interference, 2019.

[6] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[7] Dmitriy Drusvyatskiy and Lin Xiao. Stochastic optimization with decision-dependent distributions, 2020.

[8] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, 2014.

[9] Dean Eckles, Nikolaos Ignatiadis, Stefan Wager, and Han Wu. Noise-induced randomization in regression discontinuity designs, 2020. URL https://arxiv.org/abs/2004.09458.

[10] Dov Eden. Leadership and expectations: Pygmalion effects and other self-fulfilling prophecies in organizations. *The Leadership Quarterly*, 3(4):271–305, 1992.

[11] João Gama, Indre Zliobaite, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Comput. Surv.*, 46(4), 2014.

[12] Paul Goldsmith-Pinkham and Guido W. Imbens. Social networks and the identification of peer effects. *Journal of Business & Economic Statistics*, 31(3):253–264, 2013.

[13] M. Hardt, E. Price, and N. Srebo. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016.

[14] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, ITCS '16, pages 111–122, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4057-1.

[15] Moritz Hardt, Meena Jagadeesan, and Celestine Mendler-Dünner. Performative power, 2022. URL https://arxiv.org/abs/2203.17232.

[16] Keegan Harris, Daniel Ngo, Logan Stapleton, Hoda Heidari, and Zhiwei Steven Wu. Strategic instrumental variable regression: Recovering causal relationships from strategic responses, 2021.

[17] Keegan Harris, Dung Daniel T. Ngo, Logan Stapleton, Hoda Heidari, and Zhiwei Steven Wu. Strategic instrumental variable regression: Recovering causal relationships from strategic responses. *CoRR*, abs/2107.05762, 2021. URL https://arxiv.org/abs/2107.05762.

[18] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[19] Zachary Izzo, Lexing Ying, and James Zou. How to learn when data reacts to your model: Performative gradient descent. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4641–4650. PMLR, 18–24 Jul 2021.

[20] Meena Jagadeesan, Celestine Mendler-Dünner, and Moritz Hardt. Alternative microfoundations for strategic classification. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4687–4697. PMLR, 18–24 Jul 2021.

[21] Meena Jagadeesan, Tijana Zrnic, and Celestine Mendler-Dünner. Regret minimization with performative feedback. *Arxiv*, abs/2202.00628, 2022.

[22] Kaggle. Give me some credit, 2011. URL https://www.kaggle.com/competitions/GiveMeSomeCredit/data.

[23] David S. Lee and Thomas Lemieux. Regression discontinuity designs in economics. *Journal of Economic Literature*, 48(2), June 2010.

[24] Scott Levin, Matthew Toerper, Eric Hamrock, Jeremiah S. Hinson, Sean Barnes, Heather Gardner, Andrea Dugas, Bob Linton, Tom Kirsch, and Gabor Kelen. Machine-learning-based electronic triage more accurately differentiates patients with respect to clinical outcomes compared with the emergency severity index. *Annals of Emergency Medicine*, 71(5):565–574.e2, 2018. ISSN 0196-0644. doi: https://doi.org/10.1016/j.annemergmed.2017.08.005. URL https://www.sciencedirect.com/science/article/pii/S0196064417314427.

[25] Yang Liu, Yatong Chen, and Jiaheng Wei. Induced domain adaptation, 2021.

[26] Donald MacKenzie. *An engine, not a camera: How financial models shape markets*. Mit Press, 2008.

[27] Nikhil Malik. Does machine learning amplify pricing errors in housing market? : Economics of ml feedback loops. *Information Systems & Economics eJournal*, 2020.

[28] Charles F. Manski. Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies*, 60(3):531–542, 1993. ISSN 00346527, 1467937X. URL http://www.jstor.org/stable/2298123.

[29] Gustavo Manso. Feedback effects of credit ratings. *Journal of Financial Economics*, 109(2):535–548, 2013.

[30] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*, 2009. URL http://www.cs.mcgill.ca/%7Ecolt2009/papers/003.pdf#page=1.

[31] Celestine Mendler-Dünner, Juan C. Perdomo, Tijana Zrnic, and Moritz Hardt. Stochastic optimization for performative prediction. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, 2020.

[32] John Miller, Smitha Milli, and Moritz Hardt. Strategic classification is causal modeling in disguise. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6917–6926. PMLR, 13–18 Jul 2020.

[33] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, NY, USA, 2nd edition, 2009. ISBN 052189560X, 9780521895606.

[34] Judea Pearl. Causal inference in statistics: An overview. *Statist. Surv.*, 3:96–146, 2009. doi: 10.1214/09-SS057.

[35] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 7599–7609. PMLR, 13–18 Jul 2020.

[36] Reilly Raab and Yang Liu. Unintended selection: Persistent qualification rate disparities and interventions. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[37] Robert Rosenthal and Lenore Jacobson. *Pygmalion in the classroom*. The Urban Review, 1968.

[38] Yonadav Shavit, Benjamin Edelman, and Brian Axelrod. Causal strategic linear regression. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8676–8686. PMLR, 13–18 Jul 2020.

[39] Michael E Sobel. What do randomized studies of housing mobility demonstrate? causal inference in the face of interference. *Journal of the American Statistical Association*, 101(476): 1398–1407, 2006. doi: 10.1198/016214506000000636.

[40] Gloria B. Solomon, David A. Striegel, John F. Eliot, Steve N. Heon, Jana L. Maas, and Valerie K. Wayda. The self-fulfilling prophecy in college basketball: Implications for effective coaching. *Journal of Applied Sport Psychology*, 8(1):44–59, 1996.

[41] George Soros. *The Alchemy of Finance*. John Wiley Sons,, 2015.

[42] Eric J Tchetgen Tchetgen and Tyler J VanderWeele. On causal inference in the presence of interference. *Statistical Methods in Medical Research*, 21(1):55–75, 2012.

[43] Eric J Tchetgen Tchetgen and Tyler J VanderWeele. On causal inference in the presence of interference. *Statistical Methods in Medical Research*, 21(1):55–75, 2012. doi: 10.1177/0962280210386779. URL https://doi.org/10.1177/0962280210386779. PMID: 21068053.

[44] Yixin Wang and David M. Blei. The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528):1574–1596, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We analyzed practical limitations in our empirical evaluation.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix for complete proofs.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Appendix D.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix D.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Figures 4 and 5.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix D.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes] See Appendix D.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Code to replicate the experiments is available in the supplemental material.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See Appendix D.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] See Appendix D.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]