Performative Power*

Moritz Hardt[1], Meena Jagadeesan[2], and Celestine Mendler-Dünner[1]

[1]*Max-Planck Institute for Intelligent Systems, Tübingen*
[2]*University of California, Berkeley*

Abstract

Algorithmic systems, such as digital content recommendation platforms, exert influence over consumers and producers in the marketplace. To quantify this phenomenon, we introduce the notion of *performative power* which measures the ability of a firm operating an algorithmic system, to induce changes on a population of market participants. Performative power relates the empirical phenomenon of performativity to the economic concept of power.

We show that that low performative power ensures that *learning* from historical data is close to optimal. On the other hand, with high performative power, the firm can benefit from *steering* the population towards more profitable patterns, and thus the best optimization strategy for the firm can differ significantly. We then investigate the theoretical properties of performative power in the concrete setting of strategic classification: monopolies maximize performative power and disutility for the participant, while competition and outside options decrease performative power. We also investigate performative power from an empirical perspective: we construct an observational causal design to measure a platform's ability to shape consumption patterns. Finally, we discuss how performative power could be applied to competition policy and antitrust enforcement in digital marketplaces.

1 Introduction

Digital platforms pose a well-recognized challenge for antitrust enforcement. Traditional market definitions, along with associated notions of competition and market power, map poorly onto digital platforms—a core challenge is the difficulty of precisely modeling the interactions between the market participants, products, and prices. An authoritative report, published by the Stigler Committee [2019], details the many challenges associated with digital platforms, among them: *"Pinpointing the locus of competition can also be challenging because the markets are multisided and often ones with which economists and lawyers have little experience. This complexity can make market definition another hurdle to effective enforcement."* Published the same year, a comprehensive report from the European Commission calls for *"less emphasis on analysis of market definition, and more emphasis on theories of harm and identification of anti-competitive strategies."* [Crémer et al., 2019]

Our work responds to this call by developing a normative and technical proposal for reasoning about power in digital economies, while relaxing the reliance on market definition. Our running example is a digital content recommendation platform. The platform connects content creators with viewers, while monetizing views through digital advertisement. Key to the business strategy

*Authors in alphabetical order.

of a firm operating a digital content recommendation platform is its ability to predict revenue for content that it recommends or ranks highly. Often framed as a supervised learning task, the firm trains a statistical model on observed data to predict some proxy of revenue, such as clicks, views, or engagement. Better predictions enable the firm to more accurately identify content of interest and thus increase profit.

A second way of increasing profit is more subtle. The platform can use its predictions to steer participants towards modes of consumption and production that are easier to predict and monetize. For example, the platform could reward consistency in the videos created by content creators, so that the audience and the popularity of their videos becomes more predictable. Similarly, the platform could recommend addictive content to viewers, appealing to behavioral weaknesses in order to drive up viewer engagement. How potent such a strategy is depends on the extent to which the firm is able to steer participants, which we argue reveals a salient power relationship between the platform and its participants.

1.1 Our contribution

We introduce the notion of *performative power* that quantifies a firm's ability to steer a population of participants. Performative power is a causal statistical notion that directly quantifies how much participants change in response to updates to the predictive model operated by the platform. In doing so it avoids market specifics, such as the number of firms involved, products, and monetary prices. Neither does it require a competitive equilibrium notion as a reference point. Instead, it focuses on where rubber meets the road: the predictive model of the platform and its causal powers. We argue that the sensitivity of participant behavior to algorithmic changes in the platform provides an important indicator of the firm's power.

We build on recent developments in performative prediction [Perdomo et al., 2020] to articulate the fundamental difference between learning and steering in prediction. In particular, we show that under low performative power, a firm cannot do better than standard supervised learning on observed data. Intuitively, this means the firm optimizes its loss function *ex-ante* on data it observes without the ability to steer towards data it would prefer. We interpret this optimization strategy as analogous to the firm being a price-taker, an economic condition that arises under perfect competition in classical market models. We contrast this optimization strategy with a firm that performs *ex-post* optimization and takes advantage of the performative power it may have to achieve lower expected risk.

To better understand the particularities of our definition, we study performative power in the concrete market model of strategic classification. Strategic classification models participants as best-responding agents that change their features rationally in response to a predictor with the goal to achieve a better prediction outcome. In this simple setting, we show that the willingness of participants to invest in changing their features governs the performative power of the firm. We study the role of different economic factors by extending the standard model to incorporate competing firms and outside options. We highlight two key observations:

i) A *monopoly* firm maximizes performative power. In this case, participants are willing to incur a cost up to the utility of using the service in order to adjust to the firm's predictor.

ii) Performative power decreases in the presence of *competition* and *outside options*. In particular, when firms compete for participants, offering services that are perfect substitutes for each other, then even two firms can lead to zero performative power. This result stands in analogy with the classical Bertrand competition.

To complement our theoretical investigations, we propose an empirical approach to measure performative power in the context of a recommender system arranging content into display slots. This empirical approach, that we call *discrete display design (DDD)*, establishes a connection between performative power and the unilateral causal effect of position on consumption. To derive a lower bound on performative power, DDD constructs a hypothetical model update that carefully aggregates these unilateral causal effects across the population.

Finally, we examine the potential role of performative power in competition policy. We contrast performative power with traditional measures of market power, describe how performative power can capture complex behavioral patterns, and discuss the role that we envision performative power could play in ongoing antitrust debates.

1.2 Related work

Our notion of performative power is inspired by the phenomenon of performativity in prediction, discussed and formalized by Perdomo et al. [2020]. Performativity allows the predictor to influence the data-generating process, a dependency ruled out by the traditional theory of supervised learning. Thus, in *performative prediction* the predictor both fits patterns in data and impacts the population (and hence the data distribution it is being trained on). The work by Perdomo et al. [2020] has spawned many follow-up works studying optimization challenges and solution concepts in the presence of performativity, including [Mendler-Dünner et al., 2020; Izzo et al., 2021; Dong and Ratliff, 2021; Miller et al., 2021; Brown et al., 2020; Li and Wai, 2021; Ray et al., 2022; Jagadeesan et al., 2022; Wood et al., 2022]. Our work provides a complementary perspective to these existing scholarships in that we focus on the means of optimization and the role of *steering*, rather than the convergence of a particular learning algorithms to solutions of low risk. Instead of solely viewing performative feedback effects as a challenge for the learner, we argue that they reveal a salient power relationship between the decision maker and the population. Furthermore, our work demonstrates that sufficiently high performative power of a firm in a market is necessary for performative optimization approaches to be beneficial for the firm over typical supervised learning approaches on historical data.

The *strategic classification* setup we use for our case study was proposed in [Brückner et al., 2012; Hardt et al., 2016]. The standard assumption underlying strategic classification is that performative effects are induced by individuals manipulating their features so as to best respond to the deployment of a predictive model. This standard model of best responding individuals has received significant attention from both the machine learning community [Milli et al., 2019; Hu et al., 2019; Braverman and Garg, 2020; Kleinberg and Raghavan, 2019; Dong et al., 2018; Zrnic et al., 2021; Levanon and Rosenfeld, 2021] and the economics community [Frankel and Kartik, 2020; Ball, 2020; Hennessy and Goodhart, 2020; Frankel and Kartik, 2019]. While the perfect rationality and information assumptions underlying the standard model have been questioned in various works [e.g., Jagadeesan et al., 2021; Ghalme et al., 2021; Bechavod et al., 2021], the focus has exclusively been on describing participant behavior in response to a single firm acting in isolation. We believe that our extensions to incorporate additional market factors into the standard microfoundation model, such as outside options or the choice between competing firms, will be important for gaining a better understanding of the dynamics of strategic manipulation of data in digital economies. These factors not only reflect actual markets, but they can significantly alter the properties of the solutions concepts as we demonstrated in this work. Similarly, in performative prediction more broadly, existing works have predominantly focused on a single decision-maker and the performative effects it induces. Only very recently Narang et al. [2022] and Piliouras and

Yu [2022] analyzed settings with multiple firms, simultaneously applying retraining algorithms in performative environments. Similar to our analysis in Section 3, these works study the solution concept of a Nash equilibrium, however with a focus on proving convergence to equilibrium solutions, whereas we are interested in how these equilibria solutions interact and change with the performative power of the participating firms.

Beyond the related works in computer science discussed above, performative power also connects to areas, such as antitrust enforcement in digital marketplaces [Stigler Committee, 2019; Crémer et al., 2019], nudging and persuasion patterns studied in behavioral economics [Thaler and Sunstein, 2008; Fogg, 2002], and measures of market power in classical economies [Syverson, 2019]. In Section 6, we discuss in more detail how performative power relates to these research threads.

2 Performative power

Fix a set $\mathcal{U}$ of participants interacting with a designated firm, each associated with a data point $z(u)$. Fix a metric $\mathrm{dist}(z, z')$ over the space of data points. Let $\mathcal{F}$ denote the set of actions a firm can take. We think of an action $f \in \mathcal{F}$ as a predictor that the firm can deploy at a fixed point in time. For each participant $u \in \mathcal{U}$ and action $f \in \mathcal{F}$, we denote by $z_f(u)$ the potential outcome random variable representing the data of participant u if the firm were to take action f.

Definition 1 (Performative Power)**.** We define the *performative power* of the firm with respect to the population $\mathcal{U}$, action set $\mathcal{F}$, and potential outcome pairs $(z(u), z_f(u))$ for $u \in \mathcal{U}$, $f \in \mathcal{F}$ as

$$\mathrm{P} := \sup_{f \in \mathcal{F}} \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \mathbb{E}\left[\mathrm{dist}\left(z(u), z_f(u)\right)\right] ,$$

where the expectation is over the randomness in the potential outcomes.

The expression inside the supremum generalizes an average treatment effect, corresponding to scalar valued potential outcomes and the absolute value as metric. We could generalize other causal quantities such as heterogeneous treatment effects, but this avenue is not subject of our paper. The definition takes a supremum over possible actions a firm can take at a specific point in time. We can therefore lower bound performative power by estimating the causal effect of any given action $f \in \mathcal{F}$.

Having specified $\mathcal{F}$ and $\mathcal{U}$, estimating performative power amounts to causal inference involving the potential outcome variables $z_f(u)$ for unit $u \in \mathcal{U}$ and action $f \in \mathcal{F}$. In an observational design, a researcher might rely on participant data observed shortly before and after a platform updated a model (we discuss a concrete implementation of this type of strategy in Section 5). In an experimental design, the researcher would deploy a suitable chosen predictor to estimate the effect. Neither route requires understanding the market in which the firm operates. It is not even necessary to know the firm's objective function, how it optimizes its objective, and whether it successfully achieves its objective. In practice, the dynamic process that generates the potential outcome $z_f(u)$ may be highly complex, but this complexity does not enter the definition. Consequently, the definition applies to complex multisided digital economies that resist a clean definition. To make this abstract concept of performative power more concrete, let us instantiate it in a concrete example.

2.1 Running example: Digital content recommendation

Consider a digital content recommendation platform, such as the video sharing services YouTube or Twitch. The platform aims to recommend channels that generate a high revenue, and these recommendations are typically personalized to each viewer. Towards this goal, it is helpful for the platform to collect data and build a predictor f for predicting the value of recommending a channel c to a viewer with preferences p. Let $x = (x_c, x_p)$ be the features used for the prediction task that capture attributes x_c of the channel and the attributes x_p of the viewer preferences. Let y be the target variable, such as *watch time*, that acts as a proxy for the monetary value of showing a channel to a specific viewer. For concreteness, take the supervised learning loss $\ell(f(x), y)$ incurred by a predictor f to be the squared loss $(f(x) - y)^2$.

When defining performative power, participants could either be viewers or content creators. The definition is flexible and applies to both. By selecting the units $\mathcal{U}$, which features to include in the data point z, and how to specify the distance metric dist, we can pinpoint the power relationship we would like to investigate.

Content creators. The predictor f can affect the type of videos that content creators stream on their channels. For example, content creators might strategically adjust various features of their content relevant for the predicted outcome, such as the length, type or description of their videos, to improve their ranking. Thus, by changing how it predicts the monetary value of a channel, the platform can induce changes in the content on the channel. To measure this source of power, we let the participants $\mathcal{U}$ be content creators and suppose that each content creator $u \in \mathcal{U}$ maintains a channel of videos. Let the data point $z(u)$ correspond to features x_c characterizing the channel c created by content creator u. Let dist be a metric over features of content. The resulting instantiation of performative power measures the changes in content induced by potential implementations $\mathcal{F}$ of the prediction function and thus captures a power relationship between the platform and the content creators. In Section 4, we investigate this form of performative power from a theoretical perspective by building on the setup of *strategic classification*.

Viewers. The predictor f can shape the consumption patterns of viewers. In particular, viewers tend to follow recommendations when deciding what content to consume (e.g. [Ursu, 2018]). Thus, by changing which content it recommends to a user, the platform can induce changes in the target variable: how much time the users spends watching content on a given channel. Let's suppose that we wish to investigate the effect of the predictor on viewer consumption of a certain genre of content (e.g. radical content). To formalize this source of power, we let the participants $\mathcal{U}$ be viewers. Let the data point $z(u)$ correspond to how long the viewer u spends watching content in the genre of interest. (More formally, let $z(u)$ for a user u with preferences p be equal to the *sum* of the target variable y (watch time) over pairs (x_c, x_p) where c is a channel within the genre of interest.) Let $\text{dist}(z, z') = |z - z'|$ capture the difference in watch time. The resulting instantiation of performative power measures the changes in consumption of a given genre of content induced by a set of prediction functions $\mathcal{F}$ the firm could implement. In Section 5, we propose an empirical approach to measure this quantity by establishing a formal connection to the causal effect of position.

3 Learning versus steering

The presence of performative power and the ability of a firm to steer participant behavior has direct consequences for the firm's optimization problem. High performative power offers an additional lever that the firm can exploit towards achieving their objective. Instead of identifying the best action f and treating the data of the population as given, it might be beneficial for the firm to *steer* the population towards data that it prefers. In the following we elucidate the role of performative power on the optimization strategy of a firm and the equilibria attained in an economy of predictors.

3.1 Optimization approaches

Let us focus on predictive accuracy as the optimization objective of the firm. Hence, the goal of the firm is to choose a predictive model f that suffers small loss $\ell(f(x), y)$ measured over instances (x, y). To elucidate the role of steering we distinguish between the *ex-ante* loss $\ell(f(x(u)), y(u))$ and the *ex-post* loss $\ell(f(x_f(u)), y_f(u))$. The former describes the loss that the firm can optimize when building the predictor. The latter describes the loss that the firm observes after deploying f. More formally, the *ex-post risk* that the firm suffers after deploying f on a population $\mathcal{U}$ is given by

$$\frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \ell(f(x_f(u)), y_f(u)) \,. \tag{1}$$

Expression (1) constitutes an instance of what Perdomo et al. [2020] term the *performative risk* of a predictor; the loss a predictor incurs on the distribution over instances it induces. To simplify notation we adopt the conceptual device of a distribution map from Perdomo et al. [2020]. The distribution map $\mathcal{D}(\theta)$ maps a predictive model, characterized by model parameters θ, to a distribution over data instances. To express our setting within this framework, we assume the firm's action space is to choose a parameter configuration θ for its predictor f_θ from an action space $\mathcal{F} = \Theta$. Let a data instance correspond to $z(u) = (x(u), y(u))$ for $u \in \mathcal{U}$ so we can capture performativity in the features as well as in the labels. Then, the *aggregate distribution* over data $\mathcal{D}(\theta)$ corresponds to the distribution over the potential outcome variable $z_\theta(u)$ after the firm takes action $\theta \in \Theta$, where the randomness comes from u being uniformly drawn from $\mathcal{U}$ as well as randomness in the potential outcomes. The firm's ex-post risk (1) from deploying predictor f_θ corresponds to the performative risk:

$$\mathrm{PR}(\theta) := \mathop{\mathbb{E}}_{z \sim \mathcal{D}(\theta)} [\ell(\theta; z)]$$

where the loss typically corresponds to the mismatch between the predicted label and the true label: $\ell(\theta; z) = \ell(f_\theta(x), y)$ for $z = (x, y)$.

In performative risk minimization, observe that θ arises in two places in the objective: in the distribution $\mathcal{D}(\theta)$ and in the loss $\ell(\theta; z)$. Thus, for any choice of model ϕ, we can decompose the performative risk $\mathrm{PR}(\theta)$ as:

$$\mathrm{PR}(\theta) = \mathrm{R}(\phi, \theta) + (\mathrm{R}(\theta, \theta) - \mathrm{R}(\phi, \theta)) \tag{2}$$

where $\mathrm{R}(\phi, \theta) := \mathbb{E}_{z \sim \mathcal{D}(\phi)} \ell(\theta; z)$ denotes the loss of a model θ on the distribution $\mathcal{D}(\phi)$, and thus $\mathrm{PR}(\theta) = \mathrm{R}(\theta, \theta)$. This seemingly trivial decomposition reveals the salient difference between ex-ante and ex-post optimization.

Ex-ante optimization. Ex-ante optimization focuses on predicting from historical data. Let ϕ by any previously chosen model, then employing supervised learning on historical data sampled from $\mathcal{D}(\phi)$ corresponds to minimizing the first term in the decomposition (2). For any ϕ, the resulting minimizer can be computed statistically and corresponds to:

$$\theta_{\mathrm{SL}} = \arg\min_{\theta \in \Theta} \mathrm{R}(\phi, \theta)$$

In the context of digital content recommendation, this ex-ante approach corresponds to training a predictor on the participant data collected on the platform at any previous point in time.

Ex-post optimization. In contrast to ex-ante optimization, *ex-post optimization* accounts for the impact of the model on the distribution, trades-off the two terms in (2), and directly optimizes the performative risk

$$\theta_{\mathrm{PO}} = \arg\min_{\theta \in \Theta} \mathrm{PR}(\theta)\,.$$

Solving this problem exactly, and finding the performative optimum θ_{PO} requires optimization over the distribution map $\mathcal{D}(\theta)$. This can either be achieved by explicitly modeling the distribution shifts [Hardt et al., 2016], by targeted exploration [e.g., Jagadeesan et al., 2022], or implicitly through A/B testing and selecting models based on ex-post performance.[1]

It holds that $\mathrm{PR}(\theta_{\mathrm{PO}}) \leq \mathrm{PR}(\theta_{\mathrm{SL}})$, because in ex-post optimization the firm can benefit from intentionally steering the population towards a distribution that it prefers. In the context of prediction this would be a distribution that is easier to predict—for example, a distribution with better Bayes optimal risk. High ex-post predictability as a strong incentive for a firm monetizing "prediction products" has been discussed in various works, including [Russell, 2019; Shmueli and Tafti, 2020].

Remark (Generalizing to other objectives)**.** In this section, we focused on the firm's *prediction* problem when describing performative effects. Nonetheless, the conceptual distinction between learning and steering applies to *general optimization objectives*. Ex-ante optimization corresponds to optimizing on historical data, whereas ex-post optimization corresponds to implicitly or explicitly optimizing over the counterfactuals.

3.2 Gain of ex-post optimization is bounded by a firm's performative power

We show that the gain of ex-post optimization over ex-ante optimization can be bounded by the firm's performative power with respect to the set of actions Θ and the data vector $z = (x, y)$. Intuitively, if the firm's performative power is low, then the distributions $\mathcal{D}(\theta)$ and $\mathcal{D}(\phi)$ for any $\theta, \phi \in \Theta$ are close to one another. This distributional closeness, coupled with a regularity assumption on the loss, means that the second term in (2) should be small. Thus, using the ex-ante approach of minimizing the first term produces a near-optimal ex-post solution, as we demonstrate in the following result:

Theorem 1. *Let* P *be the performative power of a firm with respect to the action set* Θ*. Let* L_z *be the Lipschitzness of the loss in* z *with respect to the metric* dist*. Let* θ_{PO} *be the ex-post solution and* θ_{SL} *be the ex-ante solution computed from* $\mathcal{D}(\phi)$ *for any past deployment* $\phi \in \Theta$*. Then, we have that:*

$$\mathrm{PR}(\theta_{\mathrm{SL}}) \leq \mathrm{PR}(\theta_{\mathrm{PO}}) + 4L_z\mathrm{P}.$$

[1] There might be some discrepancy between the offline metric used for training and the business objective used for A/B testing. For the sake of mapping the problem to performative prediction, we assume the two objectives are aligned and the offline metric closely captures the business incentives.

If ℓ is γ-strongly convex, we can further bound the distance between θ_{SL} and θ_{PO} in parameter space as:

$$\|\theta_{\mathrm{SL}} - \theta_{\mathrm{PO}}\|_2 \leq \sqrt{\frac{8L_z\mathrm{P}}{\gamma}}.$$

Theorem 1 illustrates that the gain achievable through ex-post optimization is bounded by performative power. Thus, a firm with small performative power cannot do much better than ex-ante optimization and might be better off sticking to classical supervised learning practices instead of engaging with ex-post optimization. Returning to our example of digital content platforms, this means that if a firm has low performative power, then it loses very little by solving the static problem of predicting target variables from historical data, rather than explicitly trying to shape the videos produced by content creators.

3.3 Ex-post optimization in an economy of predictors

The result in Theorem 1 studies the optimization strategy of a single firm in isolation. In this section, we investigate the interaction between the strategies of multiple firms that optimize simultaneously over the same population. We consider an idealized marketplace where C firms all engage in ex-post optimization and we assume all exogenous factors remain constant. Let $\mathcal{D}(\theta^1, \ldots, \theta^{i-1}, \theta^i, \theta^{i+1}, \ldots, \theta^C)$ be the distribution over $z(u)$ induced by each firm $i \in [C]$ selecting f_{θ^i}. We say a set of predictors $[f_{\theta^1}, \ldots, f_{\theta^C}]$ is a *Nash equilibrium* if and only if no firm has an incentive to unilaterally deviate from their predictor using ex-post optimization:

$$\theta^i \in \arg\min_{\theta \in \Theta} \mathop{\mathbb{E}}_{z \sim \mathcal{D}(\theta^1, \ldots, \theta^{i-1}, \theta, \theta^{i+1}, \ldots, \theta^C)} [\ell_i(\theta; z)].$$

Here ℓ_i denotes the loss function chosen by firm i. First, we show that at the Nash equilibrium, the suboptimality of each predictor f_{θ^i} on the induced distribution depends on the performative power of the firm.

Proposition 2. *Suppose that the economy is in a Nash equilibrium $(\theta^1, \ldots, \theta^C)$, and firm i has performative power P_i with respect to the action set Θ. Let L_z be the Lipschitzness of the loss ℓ_i in z with respect to the metric* dist*. Then, it holds that:*

$$\mathop{\mathbb{E}}_{z \sim \mathcal{D}} [\ell_i(\theta^i; z)] \leq \min_{\theta} \mathop{\mathbb{E}}_{z \sim \mathcal{D}} [\ell_i(\theta; z)] + L_z \mathrm{P}_i \,,$$

where $\mathcal{D} = \mathcal{D}(\theta^1, \ldots, \theta^C)$ is the distribution induced at the equilibrium. If ℓ_i is γ-strongly convex, then we can also bound the distance between θ^i and $\arg\min_\theta \mathbb{E}_{z \sim \mathcal{D}}[\ell_i(\theta; z)]$ in parameter space.

Proposition 2 implies that if the performative power of all firms is small ($\mathrm{P}_i \to 0 \ \forall i$), then the equilibrium becomes indistinguishable from that of a static, non-performative economy with distribution $\mathcal{D}$ over content. However, there is an important distinction with the static setting: if the firms were to collude—for example, because of common ownership[2]—then they would be able to significantly shift the distribution. In particular, even if the performative power P_i of any given firm i is small, the aggregate performative power of a set of firms S can be much larger, potentially exceeding the aggregate power of the firms constituting S. As a result, low performative power only makes the optimization problem static if firms do not collude.

[2]Common ownership refers to the situation where many competitors are jointly held by a small set of large institutional investors [Azar et al., 2018].

To illustrate this, we consider a *mixture economy*, where all of the firms share a common loss function ℓ and performative power is uniformly distributed across firms. Let $z(u), z_\theta^M(u)$ denote the pair of counterfactual outcomes before and after the deployment of θ in a hypothetical monopoly economy where a single firm holds all the performative power. Let $\mathcal{D}^M(\theta)$ be the distribution map associated with the variables $z_\theta^M(u)$ for $u \in \mathcal{U}$. In a uniform mixture economy, we assume that each participant $u \in \mathcal{U}$ uniformly chooses between the C firms. The counterfactual $z_\theta(u)$ associated with one firm changing its predictor to θ is equal to $z(u)$ with probability $1 - 1/C$ and $z_\theta^M(u)$ otherwise. We can apply Proposition 2 to analyze the equilibria in the limit as $C \to \infty$.

Corollary 3. *Suppose that all firms share the same loss function $\ell_i = \ell$ for all i. Let θ^* be a symmetric equilibrium in the mixture economy with C platforms. As $C \to \infty$, it holds that:*

$$\mathop{\mathbb{E}}_{z\sim\mathcal{D}^M(\theta^*)}[\ell(\theta^*; z)] = \mathop{\mathbb{E}}_{z\sim\mathcal{D}(\theta^*,\ldots,\theta^*)}[\ell(\theta^*; z)] \to \min_\theta \mathop{\mathbb{E}}_{z\sim\mathcal{D}(\theta^*,\theta^*,\ldots,\theta^*)}[\ell(\theta; z)] = \min_\theta \mathop{\mathbb{E}}_{z\sim\mathcal{D}^M(\theta^*)}[\ell(\theta; z)].$$

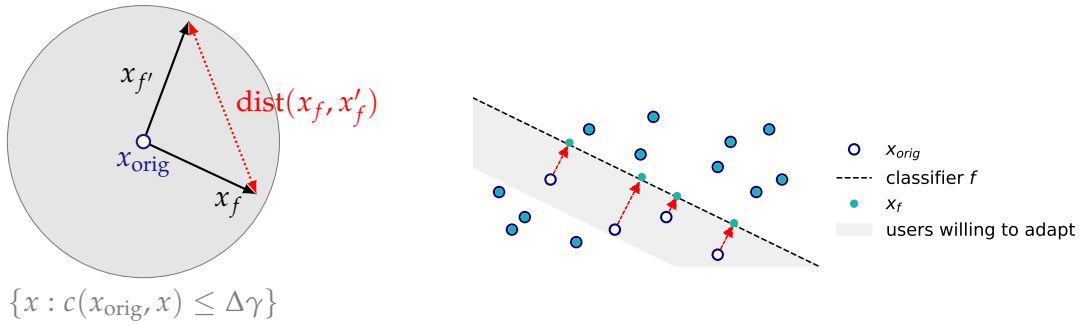
Corollary 3 demonstrates that a symmetric equilibrium approaches a *performatively stable point* of $\mathcal{D}^M$. In contrast, if a set of firms S collude, then the performative effect that they induce by all taking the same action is S times as large as the performative effect that any single firm could have induced with that action. Thus, their aggregate performative power is $\sum_{i\in S} P_i$. In the extreme case where all of the firms collude, they would benefit from all choosing a *performatively optimal point* of $\mathcal{D}^M$—which is also what the equilibrium would look like in a monopoly economy with a single firm. Since performatively optimal and performatively stable points can be arbitrarily far apart in general [Miller et al., 2021], a competitive economy can exhibit a significantly different equilibrium from that of the monopoly or collusive economy.

4 Performative power in strategic classification

Having elucidated the role of performative power in the firm's optimization strategy, we now turn to a stylized market model and investigate how performative power depends on the economy in which the firm operates. We use *strategic classification* [Hardt et al., 2016] as a test case for our definition. In strategic classification, participants strategically adapt their features with the goal of achieving a favorable classification outcome. Hence, performative power is determined by the degree to which a firm's classifier can impact participant features. Our mode of investigation is to use this concrete market setting to examine whether performative power exhibits the qualitative behavior that we would expect from a measure of power in the presence of competition and outside options.

4.1 Strategic classification setup

Let $x(u)$ be the *features* and $y(u)$ the *binary label* describing a participant $u \sim \mathcal{U}$. A firm chooses a binary predictor $f : \mathbb{R}^m \to \{0, 1\}$ and incurs loss $\ell(f(x), y) = |f(x) - y|$. The shifts in the population of participants in response to a change in the firm's predictor f is assumed to arise from participants strategically adjusting their features x in an attempt to achieve a better prediction outcome $f(x)$. Let $\mathcal{D}_{\text{base}}$ denote the base distribution over features and labels $(x_{\text{orig}}(u), y_{\text{orig}}(u))$ absent any strategic adaptation, which we assume is continuous and supported everywhere. Let $\mathcal{D}(f)$ be the distribution over potential outcomes $(x_f(u), y_f(u))$ that arises from deploying model f. We assume that participant u incurs a cost $c(x_{\text{orig}}(u), x)$ for changing their features to x. We further assume that c is a metric and that any feature change that deviates from the original features results in nonnegative cost for participants, but does not change the label $y_f(u)$ from the



(a) participant expenditor constraint

(b) participants adapting to classifier f

Figure 1: Illustrations for 1-dimensional strategic classification example

original label $y_{\mathrm{orig}}(u)$. The model is non-stateful and the cost for feature changes is measured relative to the *original* features.

Since performative effects surface as changes in participant features in strategic classification, we measure performative power over the data vector $z(u) = x(u)$. For instance, in our running example of digital content recommendations, participants correspond to content creators, and the data vector measures how much the content of each channel changes with changes in the recommendation algorithm. As a different example, suppose that firm is a hiring platform such as HireVue that uses video data from interviews to a compute a performance prediction for an applicant for a job. In this case, the participants would correspond to applicants, and the data vector would correspond to the applicant's voice patterns and interview responses.

When instantiating the definition of performative power, the choice of distance metric over $\mathcal{X} \times \mathcal{X}$ enables us to define how to weight specific feature changes. In the digital content recommendation example, if we are interested in the burden on *content creators*, an ideal distance metric would be aligned with the cost function c. However, if we are interested in measuring the impact of changes in content on viewers, a distance metric that more closely reflects the harm of specific content change for viewers might be more appropriate. We will keep this distance metric abstract in our analysis.

4.2 Performative power in the monopoly setting

We start by considering the typical strategic classification setup proposed by Hardt et al. [2016]. In this model, there exists only a single firm that offers utility γ to its participants for a positive classification. An implicit assumption is in this classical setting is that participants want to use the service independent of the classifier being deployed and the surplus utility corresponds to $\Delta\gamma = \gamma$. To better reflect real market situations, we extend this setting to incorporate *outside options*. These correspond to alternative services that are available to participants without any extra effort and offer utility level $\beta \geq 0$, which decreases the budget participants are willing to invest to $\Delta\gamma = \max(0, \gamma - \beta)$. Using this notion of surplus utility, we adopt the following standard rationality assumption on participant behavior.

Assumption 1 (Participant Behavior Specification)**.** Let $\Delta\gamma \geq 0$ be the surplus utility that a participant can expect from a positive classification outcome from classifier f over any outside option. Then, a participant $u \in \mathcal{U}$ with original features $x_{\mathrm{orig}}(u)$ will change their features

according to

$$x_f(u) = \arg\max_{x'} \left(\Delta\gamma f(x') - c(x_{\text{orig}}(u), x') \right).$$

Assumption 1 guarantees that a participant will change their features if and only if the cost of a feature change is no larger than $\Delta\gamma$. Furthermore, if a participant change their features, then they will expend the minimal cost required to achieve a positive outcome. This specification of participant behavior restricts the set of potential values that $x_f(u)$ to be close to the participant's original features $x_{\text{orig}}(u)$. We let $\mathcal{X}$ be the set of potential values of $x_f(u)$, given by:

$$\mathcal{X}(u; \Delta\gamma) := \{x : c(x_{\text{orig}}(u), x) \leq \Delta\gamma\}.$$

Given this constraint set on the counterfactual data points, we can bound performative power as follows

Lemma 4. *The performative power* P *of the firm with respect to any set of predictors $\mathcal{F}$ can be upper bounded as:*

$$\text{P} \leq \sup_{u \in \mathcal{U}} \text{diam}(\mathcal{X}(u; \Delta\gamma))$$

where the diameter is measured with respect to the metric dist*.*

The connection between the cost constraint and performative power is visualized in Figure 1(a). Note that the set $\mathcal{X}$ is defined with respect to the cost function c whereas the diameter of the set is measured in terms of the metric dist chosen for evaluating performative power. To relate the two measures we define

$$L := \sup_{x, x'} \frac{\text{dist}(x, x')}{c(x, x')}.$$

be the Lipschitz constant of dist with respect to c. This constant will show up in the next bound and can be evaluated for any choice of distance metric.

Corollary 5. *The performative of a firm in the monopoly setup can be bounded as:*

$$\text{P} \leq 2L\,\Delta\gamma. \tag{3}$$

where $\Delta\gamma$ measures the surplus utility offered by the service of the firm over outside options.

Thus, $\Delta\gamma > 0$ is a prerequisite for a firm to have any performative power, even in a monopoly economy. This qualitative behavior of performative power is in line with common intuition in economics that monopoly power relies on the firm offering a service that is superior to existing options. Interestingly, in the literature, it is typically assumed that $\Delta\gamma = \gamma$ and the firm is able to extract up to the full utility from participants close to the decision boundary. This means, in the worst case, participants invest effort but end up with a net utility of 0 in the classical setup. Outside options, however, restrict the firm's ability to extract this level of utility from participants, thereby reducing their performative power.

Heterogeneous participant base. A salient aspect of measuring performative power in the strategic classification setting is that different participants are typically impacted differently by a classifier, depending on their relative position to the decision boundary, as visualized in Figure 1(b). As a result of this heterogeneity, the upper bound in (3) is not tight, because the firm can not extract the full utility from all participants simultaneously.

Let us investigate the effect of heterogeneity in a concrete 1-dimensional setting where $\mathrm{dist}_X(x, x') = c(x, x') = |x - x'|$. Consider a set of actions $\mathcal{F}$ that corresponds to the set of all threshold functions. Suppose that the posterior $p(x) = \mathbb{P}[Y = 1 \mid X = x]$ satisfies the following regularity assumptions: $p(x)$ is strictly increasing in x with $\lim_{x\to-\infty} p(x) = 0$, and $\lim_{x\to\infty} p(x) = 1$. Now, let θ_{SL} be the supervised learning threshold, which is the unique value where $p(\theta_{\mathrm{SL}}) = 0.5$. We can then obtain the following bound on the performative power P with respect to any $\mathcal{F}$ assuming the firm's classifier is θ_{SL} in the current economy (see Proposition 10):

$$0.5\gamma \mathop{\mathbb{P}}_{\mathcal{D}_{\mathrm{base}}} \left[x \in [\theta_{\mathrm{SL}}, \theta_{\mathrm{SL}} + 0.5\Delta\gamma]\right] \leq \mathrm{P} \leq \Delta\gamma\,. \tag{4}$$

This bound illustrates how performative power in strategic classification depends on the fraction of participants that fall in between the old and the new threshold. As long as the density in this region is non-zero, a platform that offers $\Delta\gamma > 0$ utility will also have strictly positive performative power, providing a lower bound on P.

Ex-ante vs ex-post optimization. Let us contrast the two optimization approaches in this monopoly setting using the one-dimensional setting outlined above. For θ_{SL} being the supervised learning solution resulting from ex-ante optimization, the ex-post threshold lies at $\theta_{\mathrm{PO}} = \theta^{\mathrm{SL}} + \Delta\gamma$. As such, the ex-post solution leads to a higher acceptance threshold than ex-ante optimization. For any setting where the participants utility is decreasing in the threshold (e.g., the social burden proposed by Milli et al. [2019]), this implies that ex-post optimization creates stronger negative externalities for participants than ex-ante optimization. Furthermore, the effect grows with the performative power of the firm. In the extreme case of the monopoly setting with no outside options, ex-post optimization can leave participants with a net utility of 0 and thus can transfer the entire utility from these participants to the firm. This concrete setting provides an example where performative power is directly tied to participant harm.

4.3 Firms competing for participants

We next consider a dynamic model of *competition* where participants always choose the firm that offers higher utility. In this model of perfectly elastic demand, we demonstrate how the presence of competition significantly reduces the performative power of a firm. In particular, we will show that in an economy with two firms, each firm's performative power can drop to zero at equilibrium, even when the firm offers participants a nonzero utility for their service. This qualitative behavior of performative power is in line with well-known results on market power under Bertrand competition in economics, see e.g., [Baye and Kovenock, 2008].

To model competition in strategic classification, we specify participant behavior as follows: For a given predictive model f, participant u follows standard microfoundations (Assumption 1) and we assume that they select the firm for which they receive a higher utility. That is, if the first firm deploys f_1 and the second firm deploys f_2, then they will choose f_1 if

$$\max_{x'} \left(\gamma f_1(x') - c(x_{\mathrm{orig}}(u), x')\right) > \max_{x'} \left(\gamma f_2(x') - c(x_{\mathrm{orig}}(u), x')\right),$$

and choose f_2 otherwise. We assume that a participant tie-breaks in favor of the lower threshold, randomizing if the two thresholds are equal. After choosing firm $i \in \{1, 2\}$, they change their features according to $x_f(u) = \arg\max_{x'} \left(\gamma f_i(x') - c(x_{\mathrm{orig}}(u), x')\right)$.

For the results in this section, we focus on a 1-dimensional setup where $\mathcal{F}$ is the set of threshold functions. We assume that the cost function $c(x, x')$ is continuous in both of its arguments, strictly

increasing in x' for $x' > x$, strictly decreasing for $x' < x$, and satisfies $\lim_{x' \to \infty} c(x, x') = \infty$. Furthermore, we assume that the posterior $p(x) = \mathbb{P}_{\mathcal{D}_{\text{base}}}[Y = 1 \mid X = x]$ satisfies the following conditions: it is strictly increasing in x with $\lim_{x \to -\infty} p(x) = 0$, and $\lim_{x \to \infty} p(x) = 1$.

Proposition 6. *Suppose that the economy is at a symmetric state where both firms choose the same classifier θ. Let θ_{min} be the minimum threshold classifier in $\mathcal{F}(\theta)$, and for $\theta' \in \{\theta_{min}, \theta\}$, let θ'_M be the unique value such that $c(\theta'_M, \theta') = \gamma$ and $\theta'_M < \theta'$. Then, the performative power with respect to an action set $\mathcal{F}(\theta) \subseteq \mathcal{F}$ is upper bounded by:*

$$P \leq L \min(c(\theta_{min}, \theta), \gamma) + L\gamma \cdot \mathbb{P}_{\mathcal{D}_{base}}[x \in [(\theta_{min})_M, \theta_M)].$$

The performative power of a firm thus purely arises from how much larger the market threshold is than the minimum threshold a firm can deploy within their action set. We introduce the constraint $\mathcal{F}(\theta)$ on the action set of the firm in order to capture that not all actions are realistic for the firm to select. For example, one way that a firm could gain all of the participants and induce significant performative effects is by setting its threshold so that it accepts all participants. However, in doing so, the firm would accept many unqualified participants, and their profit or utility could become negative. Thus, to get a practically relevant measure of performative power, we constrain the set of actions $\mathcal{F}(\theta)$ to contain only predictors that result in a *nonnegative profit*, given that θ is chosen by the other firm. When working with historical data to evaluate performative power the choice of classifiers, we can observe naturally fall within this set, given the economic incentives of a firm.

Competition for positively labeled participants. Under a natural specification of the firm's utility functions, we show that the performative power can actually be driven down to zero in the presence of a single competitor. Therefore, we consider a strategic classification economy with two firms and participants selecting the firm that offers higher utility. Assume that a participant u contributes to a firm's utility if they choose the firm *and* the participant gets a positive prediction after strategic feature changes. In particular, the firm receives a utility of α for a positively predicted participant with positive label and a utility of $-\alpha$ for a negatively labeled participant. To ground this in an example, think of a hiring tool where a participant only contributes to a firm's utility if they select the firm and the firm decides to hire them. This contribution is positive if the participant is suitable for the job, and negative otherwise.

Now, let us focus on the state arising when the economy is at a Nash equilibrium (where both firms best-respond with respect to their utility functions taking their own performative effects into account). Using Proposition 6, we can analyze performative power of the resulting economy at equilibrium:

Corollary 7. *Suppose that the economy is at a symmetric equilibrium (θ^*, θ^*). Let $\mathcal{F}(\theta^*)$ denote the actions that a firm can take that achieve nonnegative utility assuming that the other firm chooses the classifier θ^*. If $L < \infty$, then:*

$$\mathrm{P} = 0,$$

where the performative power is measured with respect to the actions $\mathcal{F}(\theta^*)$.

The symmetric equilibrium in this economy occurs when both platforms play θ^* such that exactly half of the participants with features $x \geq \theta^*_M$ are positively labeled, where θ^*_M is the unique value such that $c(\theta^*_M, \theta^*) = \gamma$ and $\theta^*_M < \theta^*$ (see Proposition 11). More formally, it holds that

$$\mathbb{E}_{(x,y) \sim \mathcal{D}_{\text{base}}}[y = 1 \mid x \geq \theta^*_M] = \tfrac{1}{2}. \tag{5}$$

The key insight leading to Corollary 7 is that in equilibrium, the minimum threshold $\theta_{\min}$ within the set of non-negative utility solutions $\mathcal{F}(\theta^*)$ is equal to θ^*. Hence, even though both firms offer non-zero utility to their participants, they end up with 0 performative power at equilibrium because their services are perfect substitutes.

Notably, both platforms earn *zero utility* at the equilibrium state. This bears resemblance to a state of perfect competition where firms engage in price competition and at equilibrium no firm can make profit. In this sense, Corollary 7 provides an analogue to classical results on market power under Bertrand competition [Baye and Kovenock, 2008] that show how a state of zero power is reached in classical pricing economies with only two competing firms.

5 Discrete display design

Now that we have examined the theoretical properties of performative power, we turn to the question of *measuring* performative power from observational data. We focus on our running example of digital content recommendation and we propose an empirical approach to measure the recommender system's ability to shape consumption patterns. Our approach, that we call *discrete display design (DDD)*, investigates the arrangement of recommended content into an ordered set of display slots on the consumption patterns of viewers. The key ingredient of DDD is to establish a lower bound on performative power in terms of unilateral causal effects of position that have been examined in previous work.

Instantiating performative power. To instantiate performative power and pinpoint the said source of power, let the units $\mathcal{U}$ be the set of viewers. Let $\mathcal{C} = \{c_1, c_2, \ldots, c_{C-1}\} \cup \{\emptyset\}$ be the set of content available on the platform along with $\emptyset$ corresponding to an empty slot. We let $z(u)$ correspond to the empirical distribution over $\mathcal{C}$ of the content consumed by viewer u, represented as a histogram. More formally, let $z(u)$ be a vector on the C-dimensional probability simplex where the ith coordinate is the probability that viewer u consumes content c_i. The metric $\text{dist}(z, z')$ is taken to be ℓ_1 distance:

$$\text{dist}(z, z') = \sum_{i \in [m]} |z_i - z'_i|,$$

which captures the total-variation distance between probability distributions.

The decision space $\mathcal{F}$ of the firm corresponds to its decisions of assigning content to m display slots. It is natural to decompose this decision into a continuous score function s followed by a discrete conversion function c that maps scores into decisions. The score function $s : \mathcal{U} \to \mathbb{R}^C$ maps the viewer to a vector of scores that captures an estimate of the quality of the match between the viewer and each piece of content. The conversion function $c : \mathbb{R}^C \to \mathcal{C}^m$ takes as input the vector of scores and outputs an ordered list of the content corresponding to the top m scores that is displayed in the m display slots in order. We assume c is fixed, so the choice of s determines the firm's action $f = c \circ s$. Thus, we focus on the firm action space to be a set of score functions $\mathcal{S}$.

Construction of counterfactual. In the *discrete display design*, we investigate the effect of small perturbations to the current scoring function s_{curr}. Let δ be the maximum difference in the highest score and second highest score for any user, that is:

$$\delta = \max_{u \in U} \left(h^1(u) - h^2(u) \right),$$

where $h^1(u)$ and $h^2(u)$ are the highest and second highest values in the multi-set $\{s[u](i)\}_{1\leq i\leq C}$, respectively. More formally, let

$$\mathcal{S} = \left\{ s : \mathcal{U} \to \mathbb{R}^C \mid \|s(u) - s_{curr}(u)\|_\infty \leq \delta \ \forall u \in \mathcal{U} \right\}.$$

We denote the counterfactual variable induces by a score function $s \in \mathcal{S}$ as $z_s(u)$. In this notation, performative power with respect to the action set $\mathcal{S}$ is equal to:

$$P = \sup_{s\in\mathcal{S}} \frac{1}{\mathcal{U}} \sum_{u\in\mathcal{U}} \|z(u) - z_s(u)\|_\infty.$$

Small perturbations to the scoring function can lead to content in adjacent display slots being swapped. Let $s' \in S$ be a perturbation that leads to the items in the first two display slots being swapped. The key insight of our analysis is that we can lower bound performative power by the shift induced by changing s_{curr} to s' whose effect we can relate to the *causal effect of position* of display slots, also known as position bias.

Definition 2 (Causal effect of position)**.** Let the treatment T be the action of flipping the content in the first and second display slots for a viewer, and let the potential outcome variable Y be an indicator for whether the viewer consumes the content that is initially in their first display slot. We call the corresponding average treatment effect

$$\beta := \left|\mathbb{E}[Y \mid \text{do}(T)] - \mathbb{E}[Y]\right|$$

the *causal effect of position*, where the expectation is taken over the population of viewers $\mathcal{U}$ and randomness in the potential outcomes.

To relate performative power to the causal effect of position, we place the following natural assumption on the counterfactual variables. The assumption requires that what content a user consumes only depends on the content recommended to them and not the recommendations to other users.

Assumption 2 (No interference across units)**.** For any $u \in \mathcal{U}$ and any pair of scoring functions $s_1, s_2 \in \mathcal{S}$, if $c(s_1(u)) = c(s_2(u))$ it also holds that $z_{s_1}(u) = z_{s_2}(u)$.[3]

Assumption 2 holds as long as there are no spill-over or peer effects that make the content shown to one user affect the consumption patterns of another user. It closely relates to the stable unit treatment value assumption (SUTVA) [Imbens and Rubin, 2015] prevalent in causal inference.

We are now ready to state the relationship between performative power and the causal effect β.

Theorem 8 (Bounding Performative power with DDD)**.** *Let* P *be the performative power with respect to* $\mathcal{S}$ *and* $\|\cdot\|_\infty$*, and let* β *be the causal effect of position. If Assumption 2 holds, then:*

$$\mathrm{P} \geq \beta.$$

Theorem 8 establishes a lower bound on performative power in terms of the causal effect of position. In the next section, we show we can explicitly leverage Theorem 8 in the context of search advertisement.

[3]If an approximate version of Assumption 2 holds (i.e. if $|z_{s_1}(u) - z_{s_2}(u)| < \delta'$, then our lower bound for performative power will apply up to a δ' additive error.

Example: Search advertisement. We explicitly estimate a lower bound on the performative power for the search advertisement marketplace studied by Narayanan and Kalyanam [2015]. They examine position effects in search advertising, where advertisements are displayed across a number of ordered slots whenever a keyword is searched. They show that the position effect β of display slot 1 versus display slot 2 is 0.0048.

To arrive at this number, Narayanan and Kalyanam [2015] implement a regression discontinuity approach to estimate the position effect. The input is a sample of data (k, p, z, y) where k is a keyword, $p \in \{1, 2\}$ is the position of the advertisement in the list of displayed content (the data is restricted to content in either position 1 or position 2), z is the AdRank score, and y is the click-through-rate (CTR). The following local linear regression estimator is applied to a subset of the data with appropriate window size λ chosen appropriately:

$$y = \alpha + \beta I[p = 1] + \gamma_1 z + \gamma_2 z I[p = 1] + f(k; \theta) + \epsilon_1 \tag{6}$$

The value β is an estimate of the *position effect* of the display slot as defined in Definition 2.

To connect this causal effect estimate to performative power, we treat each keyword as a distinct "viewer". The distribution $\mathcal{D}$ captures the distribution over keyword queries. Following the query, the viewer either clicks on the advertisement in one of the display slots or does not click on any advertisement. The value $z_s(u)[i]$ corresponds to the probability that the content c_i is consumed by user u under the scoring rule s. More specifically, for j such that $c_j = \emptyset$, the value $z_s(u)[j]$ corresponds to the probability that the viewer does not click on any advertisement. If c_i is displayed, $z_s(u)[i]$ corresponds to the click-through-rate.

In this instantation of performative power, we can apply Theorem 8 to see that $P \geq 0.0048$. To contextualize this value, we note that the mean click-through rate in display slot 2 is 0.023260. Thus, the lower bound 0.0048 is a 21% percent increase relative to the baseline click-through rate. The firm thus has a substantial ability to shape what advertisements users click on.

Discussion of discrete display design. We showed how the discrete display design can be leveraged to evaluate a lower bound on the performative power, focusing on a particular source of power: steering consumption patterns through arrangement of content display slots. The strategy we outline relies on the measurement of individual causal effects of position bias from prior work and reuses these values to derive a lower bound on performative power through a targeted construction of a counterfactual firm action. We believe that the high-level approach of relating performative power to unilateral causal effects can be generalized to other discrete decisions beyond display slots. We defer investigating DDD in more generality to future work.

6 Performative power in competition policy

Having investigated performative power from a theoretical and an empirical perspective, we discuss the potential role of performative power in competition policy and antitrust enforcement as a complementary tool to the existing machinery. We highlight that performative power fundamentally differs from traditional measures of market power, in that it relies on a directly observational statistic and does not require a precise specification of the market. As such it can be applied to obtain insights even in markets that resist a clean definition. In addition it is sensitive to spillover effects across market boundaries and behavioral aspects that are challenging to model. We note however that performative power focuses on measuring power rather than harm, which is important for how it should be used in antitrust enforcement.

6.1 Measures of market power in economics

Typical measures of market power in economic theory focus on classical pricing markets of homogeneous goods, where a firm's primary action is choosing a price to sell the good or the quantity of the good to sell. The scalar nature of these quantities enables them to be easily compared across different market contexts and firms. In addition, the utility of the firm and the utility of participants are inversely related: a higher price yields greater utility for the firm and lower utility for all participants. This simple relationship allows one to directly reason about participant welfare and profit of firms. However, the situation of a digital economy is much more complex than proposed by these classical models, as argued in [Stigler Committee, 2019; Crémer et al., 2019]. As a result, classical measures from competition theory that primarily focused on price effects struggle to accurately characterize these economies. Let us explain the challenges using two text-book definitions of market power:

Lerner index. The Lerner Lerner [1934] index quantifies the pricing power of a firm, measuring by how much the firm can raise the price above marginal costs. Marginal costs reflect the price that would arise in a perfectly competitive market. A major issue of applying this standard definition of market power in the context of digital economies is that it is not clear what the competitive reference state should look like, "We have lost the competitive benchmark,"[4] as Jacques Crémer said. Thus, measures based on profit margin cannot directly be adopted as a proxy for market power in these settings.

Market share. Measures such as the Herfindahl–Hirschman index (HHI), which is used by the US federal trade commission[5] to measure market competitiveness, is based on *market share*: the fraction of participants who participate in a given firm. However, the validity of market share as a proxy for power relies on a specific model of competition where the elasticity of demand is low. This model is challenging to justify[6] in the context of digital economies where opening an account on a platform is very simple and usually free of charge. In addition, not all participants with accounts on a digital platform are equally active and inactive participants should not factor into the market power of a firm the same way active participants do. Market share is not sufficiently expressive to make this distinction.

6.2 Complex consumer behavior

In addition to sidestepping the need to specify a model for competition, performative power is also agnostic to the behavior of participants. As outlined by the Stigler Committee [2019], "the findings from behavioral economics demonstrate an under-recognized market power held by incumbent digital platforms." In particular, behavioral aspects of consumers—such as tendencies for single-homing, vulnerability to addiction, and as well as the impact of framing and nudging on participant behavior [e.g. Thaler and Sunstein, 2008; Fogg, 2002]—can be exploited by firms in digital economies, but do not factor into traditional measures of market power. By focusing on changes in participant features, performative power has the potential to capture the effects of these behavioral patterns while again sidestepping the challenges of explicitly modeling them.

[4] Opening statement at the 2019 Antitrust and competition conference – digital platforms, markets, and democracy

[5] See https://www.justice.gov/atr/herfindahl-hirschman-index (retrieved January, 2022).

[6] This critique is similar to the disconnect between the Cournot model and the Bertrand model in classical economics [Bornier, 1992]. E.g., "concentration is worse than just a noisy barometer of market power" [Syverson, 2019].

6.3 From performative power to consumer harm

High performative power and the ability of a firm to steer participants naturally bears the potential for user harm, whenever there exists a misalignment between the utilities of the platform and the participants. In this context we want to highlight a recent work by Shmueli and Tafti [2020] that flags a similar concern and points to empirical evidence. However, in general, harm and power are two fundamentally distinct normative concepts. Performative power does not necessarily imply harm, but it can serve as an indicator of *potential* harm that can help flag market situations that merit further investigations by regulators.[7] However, for the purpose of antitrust enforcement it remains to establish a more formal connection between the two concepts.

That being said, in specific contexts, we can establish an exact correspondence between performative power and harm. By carefully choosing the parameters in the definition, performative power can be instantiated to more closely capture manifested harm to a particular (sub-)population of participants we wish to investigate. We implement this principle in Section 4 where we establish a connection between performative power and the utility of users. More generally, this connection can be achieved if the attributes $z(u)$ consists of the sensitive features that are impacted by the firm, the distance function is aligned with the utility function of participants, and the set $\mathcal{F}$ reflects actions that are taken by the firm.

7 Future directions

Our work builds a foundation for further investigations of the power exhibited by algorithmic systems, the implications for user welfare, and the development of technically grounded regulation for digital marketplaces. We outline some interesting directions for further inquiry. The first direction is *designing empirical approaches to measure different sources of performative power*. This includes measuring the ability of a recommender system to shape content production, consumption patterns or viewer preferences. Another direction of interest is *operationalizing performative power for antitrust regulation*. In particular, how does performative power relate to concepts in competition policy, and above what value should performative power be considered problematically high? Finally, it would be interesting to *investigate in more depth how standard algorithmic practices—such as retraining and A/B testing—are influenced by performative power.*

Acknowledgments

We're grateful to Martin C. Schmalz and Emilio Calvano for helpful feedback and pointers. MJ acknowledges support from the Paul and Daisy Soros Fellowship and the Open Phil AI Fellowship.

[7] In fact, the report of the European Commission on competition policy for the digital era [Crémer et al., 2019] suggests in several contexts where regulators should be suspicious of power and suggests putting burden of proof on the firm rather than the regulators.

References

Jose Azar, Martin C. Schmalz, and Isabel Tecu. Anticompetitive effects of common ownership. *The Journal of Finance*, 73(4):1513–1565, 2018.

Ian Ball. Scoring strategic agents. *ArXiv:1909.01888*, 2020.

Michael R. Baye and Dan Kovenock. *Bertrand competition*. Palgrave Macmillan UK, London, 2008.

Yahav Bechavod, Chara Podimata, Zhiwei Steven Wu, and Juba Ziani. Information discrepancy in strategic learning. *Arxiv:2103.01028*, 2021.

Jean Magnan de Bornier. The "Cournot-Bertrand Debate": A Historical Perspective. *History of Political Economy*, 24(3):623–656, 09 1992.

Mark Braverman and Sumegha Garg. The role of randomness and noise in strategic classification. In *Proc. 1st FORC 2020*, volume 156, pages 9:1–9:20, 2020.

Gavin Brown, Shlomi Hod, and Iden Kalemaj. Performative prediction in a stateful world. *arXiv preprint arXiv:2011.03885*, 2020.

Michael Brückner, Christian Kanzow, and Tobias Scheffer. Static prediction games for adversarial learning problems. *JMLR*, 13(1):2617–2654, September 2012.

Jacques Crémer, Yves-Alexandre de Montjoye, and Heike Schweitzer. *Competition Policy for the digital era : Final report*. Publications Office of the European Union, 2019.

Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 55–70, 2018.

Roy Dong and Lillian J Ratliff. Approximate regions of attraction in learning with decision-dependent distributions. *arXiv preprint arXiv:2107.00055*, 2021.

B. J. Fogg. Persuasive technology: Using computers to change what we think and do. *Ubiquity*, dec 2002. doi: 10.1145/764008.763957.

Alex Frankel and Navin Kartik. Muddled Information. *Journal of Political Economy*, 127(4): 1739–1776, 2019.

Alex Frankel and Navin Kartik. Improving Information via Manipulable Data. *Working Paper*, 2020.

Ganesh Ghalme, Vineet Nair, Itay Eilat, Inbal Talgam-Cohen, and Nir Rosenfeld. Strategic classification in the dark. In *Proceedings of the 38th International Conference on Machine Learning*, pages 3672–3681, 2021.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. ITCS '16, pages 111–122, 2016.

Christopher Hennessy and Charles Goodhart. Goodhart's law and machine learning. *SSRN*, 2020.

Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. The disparate effects of strategic manipulation. In *Proc. FAccT*, pages 259–268, 2019.

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

Zachary Izzo, Lexing Ying, and James Zou. How to learn when data reacts to your model: Performative gradient descent. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 4641–4650, 2021.

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, pages 4687–4697, 2021.

Meena Jagadeesan, Tijana Zrnic, and Celestine Mendler-Dünner. Regret minimization with performative feedback. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 9760–9785, 2022.

Jon Kleinberg and Manish Raghavan. How do classifiers induce agents to invest effort strategically? In *Proceedings of the 2019 ACM Conference on Economics and Computation*, EC '19, pages 825–844, 2019.

A. P. Lerner. The concept of monopoly and the measurement of monopoly power. *The Review of Economic Studies*, 1(3):157–175, 1934.

Sagi Levanon and Nir Rosenfeld. Strategic classification made practical. *Arxiv:2103.01826*, 2021.

Qiang Li and Hoi-To Wai. State dependent performative prediction with stochastic approximation. *ArXiv:2110.00800*, 2021.

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

John Miller, Juan C. Perdomo, and Tijana Zrnic. Outside the echo chamber: Optimizing the performative risk. *Arxiv:2102.08570*, 2021.

Smitha Milli, John Miller, Anca D. Dragan, and Moritz Hardt. The social cost of strategic classification. In *Proc. FAccT*, pages 230–239, 2019.

Adhyyan Narang, Evan Faulkner, Dmitriy Drusvyatskiy, Maryam Fazel, and Lillian J. Ratliff. Multiplayer performative prediction: Learning in decision-dependent games. *ArXiv:2201.03398*, 2022.

Sridhar Narayanan and Kirthi Kalyanam. Position effects in search advertising and their moderators: A regression discontinuity approach. *Marketing Science*, 34(3):388–407, may 2015. ISSN 1526-548X.

Juan C. Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *Proc. 37th ICML*, volume 119, pages 7599–7609, 2020.

Georgios Piliouras and Fang-Yi Yu. Multi-agent performative prediction: From global stability and optimality to chaos. *Arxiv:2201.10483*, 2022.

Mitas Ray, Lillian J Ratliff, Dmitriy Drusvyatskiy, and Maryam Fazel. Decision-dependent risk minimization in geometrically decaying dynamic environments. 2022.

S. Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin Publishing Group, 2019.

Galit Shmueli and Ali Tafti. "Improving" prediction of human behavior using behavior modification. *Arxiv:2008.12138*, 2020.

Stigler Committee. Final report: Stigler committee on digital platforms. available at https://research.chicagobooth.edu/stigler/media/news/committee-on-digitalplatforms-final-report, September 2019.

Chad Syverson. Macroeconomics and market power: Context, implications, and open questions. *Journal of Economic Perspectives*, 33(3):23–43, August 2019.

Richard H. Thaler and Cass R. Sunstein. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Yale University Press, 2008.

Raluca M. Ursu. The power of rankings: Quantifying the effect of rankings on online consumer search and purchase decisions. *Marketing Science*, 37(4):530–552, 2018.

Killian Wood, Gianluca Bianchin, and Emiliano Dall'Anese. Online projected gradient descent for stochastic optimization with decision-dependent distributions. *IEEE Control Systems Letters*, 6: 1646–1651, 2022.

Tijana Zrnic, Eric Mazumdar, Shankar Sastry, and Michael Jordan. Who leads and who follows in strategic classification? *Advances in Neural Information Processing Systems*, 34, 2021.

A Proofs

A.1 Auxiliary results

The proofs for Section 3 leverage the following lemma, which bounds the diameter of Θ with respect to Wasserstein distance in distribution map.

Lemma 9. *Let* P *be the performative power with respect to* Θ*. For any* $\theta, \theta' \in \Theta$*, it holds that* $\mathcal{W}(\mathcal{D}(\theta), \mathcal{D}(\theta')) \leq 2\mathrm{P}$.

Proof. Let θ_{curr} be the current classifier weights. We use the fact that for any weights $\theta'' \in \Theta$, it holds that $\mathcal{W}(\mathcal{D}(\theta_{\text{curr}}), \mathcal{D}(\theta'')) \leq \frac{1}{|U|} \sum_{u \in \mathcal{U}} \mathbb{E}[\text{dist}(z(u), z_{\theta''}(u))]$ where the expectation is over randomness in the potential outcomes. This follows from the definition of Wasserstein distance—in particular that we can instantiate the mass-moving function by mapping each participant to themselves. Thus, we see that:

$$\begin{aligned}
\mathcal{W}(\mathcal{D}(\theta), \mathcal{D}(\theta')) &\leq \mathcal{W}(\mathcal{D}(\theta), \mathcal{D}(\theta_{\text{curr}})) + \mathcal{W}(\mathcal{D}(\theta_{\text{curr}}), \mathcal{D}(\theta')) \\
&\leq \frac{1}{|U|} \sum_{u \in \mathcal{U}} \mathbb{E}[\text{dist}(z(u), z_{\theta}(u))] + \frac{1}{|U|} \sum_{u \in \mathcal{U}} \mathbb{E}[\text{dist}(z(u), z_{\theta'}(u))] \\
&\leq 2 \sup_{\theta'' \in \Theta} \frac{1}{|U|} \sum_{u \in \mathcal{U}} \mathbb{E}[\text{dist}(z(u), z_{\theta''}(u))] \\
&\leq 2\mathrm{P},
\end{aligned}$$

where the last line uses the definition of performative power that bounds the effect of any θ in the action set Θ on the participant data z. □

A.2 Proof of Theorem 1

Let ϕ be the previous deployment inducing the distribution on which the supervised learning threshold θ_{SL} is computed. Let θ^* be an optimizer of $\min_{\theta \in \Theta} \mathrm{R}(\theta_{\text{PO}}, \theta)$, where we recall the definition of the decoupled performative risk as $\mathrm{R}(\phi, \theta) := \mathbb{E}_{z \sim \mathcal{D}(\phi)} \ell(\theta; z)$. Then, we see that for any ϕ:

$$\begin{aligned}
&\mathrm{PR}(\theta^{\text{SL}}) - \mathrm{PR}(\theta_{\text{PO}}) \\
&= \Big(\mathrm{R}(\theta^{\text{SL}}, \theta^{\text{SL}}) - \mathrm{R}(\phi, \theta^{\text{SL}})\Big) + \mathrm{R}(\phi, \theta^{\text{SL}}) - \mathrm{R}(\theta_{\text{PO}}, \theta_{\text{PO}}) \\
&\leq \Big(\mathrm{R}(\theta^{\text{SL}}, \theta^{\text{SL}}) - \mathrm{R}(\phi, \theta^{\text{SL}})\Big) + \mathrm{R}(\phi, \theta^*) - \mathrm{R}(\theta_{\text{PO}}, \theta^*) \\
&\leq L_z \mathcal{W}(\mathcal{D}(\theta^{\text{SL}}), \mathcal{D}(\phi)) + L_z \mathcal{W}(\mathcal{D}(\phi), \mathcal{D}(\theta_{\text{PO}})) \\
&\leq 4 L_z P\,.
\end{aligned}$$

The first inequality follows because θ^* minimizes risk on the distribution $\mathcal{D}(\theta_{\text{PO}})$, while θ^{SL} minimizes risk on $\mathcal{D}(\phi)$. The second inequality follows from the dual of the Wasserstein distance where L_z is the Lipschitz constant of the loss function in the data argument z. The last inequality follows from Lemma 9.

Now, suppose that ℓ is γ-strongly convex. Then we have that:

$$\mathrm{R}(\theta, \theta_{\text{PO}}) - \mathrm{R}(\theta, \theta_{\text{SL}}) \geq \frac{\gamma}{2} \|\theta_{\text{PO}} - \theta_{\text{SL}}\|^2$$

Again applying Lemma 9,

$$\begin{aligned}
\mathrm{PR}(\theta_{\mathrm{SL}}) &= \mathrm{R}(\theta_{\mathrm{SL}}, \theta_{\mathrm{SL}}) \\
&\leq \mathrm{R}(\theta, \theta_{\mathrm{SL}}) + L_z \mathcal{W}(\mathcal{D}(\phi), \mathcal{D}(\theta_{\mathrm{SL}})) \\
&\leq \mathrm{R}(\phi, \theta_{\mathrm{SL}}) + 2L_z P \\
&\leq \mathrm{R}(\phi, \theta_{\mathrm{PO}}) - \frac{\gamma}{2}\|\theta_{\mathrm{PO}} - \theta_{\mathrm{SL}}\|^2 + 2L_z P \\
&\leq \mathrm{R}(\theta_{\mathrm{PO}}, \theta_{\mathrm{PO}}) + L_z \cdot \mathcal{W}(\mathcal{D}(\phi), \mathcal{D}(\theta_{\mathrm{PO}})) - \frac{\gamma}{2}\|\theta_{\mathrm{PO}} - \theta_{\mathrm{SL}}\|^2 + 2L_z P \\
&\leq \mathrm{PR}(\theta_{\mathrm{PO}}) + 4L_z P - \frac{\gamma}{2}\|\theta_{\mathrm{PO}} - \theta_{\mathrm{SL}}\|^2.
\end{aligned}$$

Using that $\mathrm{PR}(\theta_{\mathrm{PO}}) \leq \mathrm{PR}(\theta_{\mathrm{SL}})$, we find that

$$\frac{\gamma}{2}\|\theta_{\mathrm{PO}} - \theta_{\mathrm{SL}}\|^2 \leq 4L_z P\,.$$

Rearranging gives

$$\|\theta_{\mathrm{PO}} - \theta_{\mathrm{SL}}\| \leq \sqrt{\frac{8L_z P}{\gamma}}\,.$$

A.3 Proof of Proposition 2

Let's focus on firm i, fixing classifiers selected by the other firms. Let's take PR and R to be defined with respect to $\mathcal{D}(\cdot) = \mathcal{D}(\theta_1, \ldots, \theta_{i-1}, \cdot, \theta_{i+1}, \ldots, \theta_{\mathrm{C}})$. Let $\theta^* = \arg\min_\theta \mathrm{R}(\theta_i, \theta)$. We see that:

$$\begin{aligned}
\mathrm{PR}(\theta^i) &\leq \mathrm{PR}(\theta^*) \\
&\leq \mathrm{R}(\theta_i, \theta^*) + L_z \mathcal{W}(\mathcal{D}(\theta_i), \mathcal{D}(\theta^*)) \\
&\leq \min_\theta \mathrm{R}(\theta_i, \theta) + L_z \left(\frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \mathbb{E}[\mathrm{dist}(z(u), z_{\theta^*}(u))] \right) \\
&\leq \min_\theta \mathrm{R}(\theta_i, \theta) + L_z P_i.
\end{aligned}$$

Rewriting this, we see that:

$$\mathop{\mathbb{E}}_{z \sim \mathcal{D}}[\ell_i(\theta^i; z)] \leq \min_\theta \mathop{\mathbb{E}}_{z \sim \mathcal{D}}[\ell_i(\theta; z)] + L_z P_i.$$

If, in addition, ℓ_i is γ-strongly convex, then we know that:

$$L_z P_i \geq \mathop{\mathbb{E}}_{z \sim \mathcal{D}}[\ell_i(\theta^i; z)] - \min_\theta \mathop{\mathbb{E}}_{z \sim \mathcal{D}}[\ell(\theta; z)] \geq \frac{\gamma}{2}\|\theta^i - \min_\theta \mathop{\mathbb{E}}_{z \sim \mathcal{D}}[\ell_i(\theta; z)]\|^2.$$

Rearranging, we obtain that

$$\left\| \theta^i - \min_\theta \mathop{\mathbb{E}}_{z \sim \mathcal{D}}[\ell(\theta; z)] \right\|_2 \leq \sqrt{\frac{2L_z P_i}{\gamma}}\,.$$

A.4 Proof of Corollary 3

Let P be the performative power associated with the variables z_θ^M. We first claim that the performative power of any firm in the mixture model is at most P/C. This follows from the fact that for a given firm the potential outcome $z_\theta(u)$ is equal to $z(u)$ with probability $1 - 1/C$ and equal to $z_\theta^M(u)$ with probability $1/C$.

Let's focus on platform i, fixing classifiers selected by the other platforms. Let's take PR and R to be with respect to $\mathcal{D}(\cdot) = \mathcal{D}(\theta^*, \cdots, \theta^*, \cdot, \theta^*, \ldots, \theta^*)$. Now, we can apply Proposition 2 to see that

$$\mathrm{PR}(\theta^*) = \mathop{\mathbb{E}}_{z\sim\mathcal{D}^M(\theta^*)}[\ell(\theta^*; z)] \leq \min_\theta \mathop{\mathbb{E}}_{z\sim\mathcal{D}^M(\theta^*)}[\ell(\theta; z)] + \frac{L_z P}{C}.$$

Thus, in the limit as $C \to \infty$, it holds that

$$\mathop{\mathbb{E}}_{z\sim\mathcal{D}^M(\theta^*)}[\ell(\theta^*; z)] \to \min_\theta \mathop{\mathbb{E}}_{z\sim\mathcal{D}^M(\theta^*)}[\ell(\theta; z)]$$

as desired.

A.5 Proof of Lemma 4

By Assumption 1, we know that $x(u)$ and $x_f(u)$ are both in $\mathcal{X}(u; \Delta\gamma)$. If we let D be the diameter of $\mathcal{X}(u; \Delta\gamma)$, this means that $\mathrm{dist}(x(u), x_f(u)) \leq D$. We can upper bound performative power as follows:

$$\begin{aligned}
\mathrm{P} &\leq \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}} \mathbb{E}[\mathrm{dist}(x(u), x_f(u))] \\
&\leq \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}} D \\
&= D,
\end{aligned}$$

as desired.

A.6 Proof of Corollary 5

Applying Lemma 9, it suffices to show that the diameter of $\mathcal{X}(u; \Delta\gamma)$ can be upper bounded by $2L\Delta\gamma$. We see that for any $x, x' \in \mathcal{X}(u; \Delta\gamma)$, it holds that:

$$\begin{aligned}
\mathrm{dist}(x, x') &\leq L \cdot c(x, x') \\
&\leq L \cdot (c(x_{\mathrm{orig}}(u), x) + c(x_{\mathrm{orig}}(u), x')) \\
&\leq 2L\Delta\gamma,
\end{aligned}$$

using that c is a metric and using the definition of $\mathcal{X}(u; \Delta\gamma)$.

A.7 Statement and proof of Proposition 10

Proposition 10. *Suppose that* $\mathrm{dist}(x, x') = c(x, x') = |x - x'|$. *Let us consider a set of actions* $\mathcal{F}$ *that corresponds to the set of all threshold functions. Suppose that the posterior* $p(x) = \mathbb{P}[Y = 1 \mid X = x]$ *satisfies the following regularity assumptions:* $p(x)$ *is strictly increasing in* x *with* $\lim_{x\to-\infty} p(x) = 0$, *and* $\lim_{x\to\infty} p(x) = 1$. *Now, let* θ_{SL} *be the supervised learning threshold, which is the unique value where*

$p(\theta_{\text{SL}}) = 0.5$. *If the firm's classifier is θ_{SL} in the current economy, then performative power* P *with* $\mathcal{F}$ *can be bounded as:*

$$0.5\gamma \mathop{\mathbb{P}}_{\mathcal{D}_{base}} \left[x \in [\theta_{SL}, \theta_{SL} + 0.5\Delta\gamma]\right] \leq \text{P} \leq 2\Delta\gamma\,. \tag{7}$$

Proof. The upper bound follows from Corollary 7. For the lower bound, we take f to be the threshold classifier given by $\theta_{\text{SL}} + \Delta\gamma$. We see that for $x_{\text{orig}}(u) \in [\theta_{\text{SL}}, \theta_{\text{SL}} + \Delta\gamma]$, it holds that $x_f(u) = \theta_{\text{SL}} + \Delta\gamma$ and $x(u) = x_{\text{orig}}(u)$. This means that the performative power is at least:

$$\begin{aligned}
\text{P} &= \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \mathbb{E}[\text{dist}(x(u), x_f(u))] \\
&= \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \mathbb{E}[|x(u) - x_f(u)|] \\
&\geq \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} I[x_{\text{orig}}(u) \in [\theta_{\text{SL}}, \theta_{\text{SL}} + \Delta\gamma]]\, \mathbb{E}[|\theta_{\text{SL}} + \Delta\gamma - x_{\text{orig}}(u)|] \\
&\geq \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} I[x_{\text{orig}}(u) \in [\theta_{\text{SL}}, \theta_{\text{SL}} + 0.5\Delta\gamma]] \cdot 0.5\Delta\gamma \\
&\geq 0.5\Delta\gamma \mathop{\mathbb{P}}_{\mathcal{D}_{\text{base}}} [x \in [\theta_{\text{SL}}, \theta_{\text{SL}} + 0.5\Delta\gamma]],
\end{aligned}$$

as desired. □

A.8 Proof of Proposition 6

We prove Proposition 6.

Proof of Proposition 6. Consider a classifier $f \in \mathcal{F}(\theta)$ with threshold θ', and suppose that a firm changes their classifier to f. It suffices to show that:

$$\frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \mathbb{E}[\text{dist}(x(u), x_f(u))] \leq L \min(c(\theta_{\min}, \theta), \gamma) + L\gamma \cdot \mathbb{P}_{\mathcal{D}_{\text{base}}}[x \in [(\theta_{\min})_M, \theta_M]].$$

For technical convenience, we reformulate this in terms of the cost function c. Based on the definition of L, it suffices to show that:

$$\frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \mathbb{E}[c(x(u), x_f(u))] \leq \min(c(\theta_{\min}, \theta), \gamma) + \gamma \cdot \mathbb{P}_{\mathcal{D}_{\text{base}}}[x \in [(\theta_{\min})_M, \theta_M)]].$$

Case 1: $\theta' > \theta$**.** Participants either are indifferent between θ and θ' or prefer θ to θ'. Due to the tie breaking rule, the firm will thus lose all of its participants. Thus, all participants will switch to the other firm and adapt their features to that firm which has threshold θ. This is the same behavior as these participants had in the current state, so $x_f(u) = x(u)$ for all participants u. This means that

$$\frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \mathbb{E}[c(x(u), x_f(u))] = 0$$

as desired.

Case 2: $\theta' < \theta$**.** Participants either are indifferent between θ and θ' or prefer θ' to θ. Due to the tie breaking rule, the firm will thus gain all of the participants. We break into several cases:

$$\begin{cases} x_f(u) = x(u) = x_{\text{orig}}(u) & \text{if } x_{\text{orig}}(u) < \theta'_M \\ x_f(u) = \theta', x(u) = x_{\text{orig}}(u) & \text{if } x_{\text{orig}}(u) \in [\theta'_M, \min(\theta', \theta_M))] \\ x_f(u) = x(u) = x_{\text{orig}}(u) & \text{if } x_{\text{orig}}(u) \in (\theta', \theta_M) \\ x_f(u) = \theta', x(u) = \theta & \text{if } x_{\text{orig}}(u) \in (\theta_M, \theta') \\ x_f(u) = x_{\text{orig}}(u), x(u) = \theta & \text{if } x_{\text{orig}}(u) \in [\max(\theta', \theta_M), \theta] \\ x_f(u) = x(u) = x_{\text{orig}}(u) & \text{if } x_{\text{orig}}(u) \geq \theta. \end{cases}$$

The only cases that contribute to $\frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}} \mathbb{E}[c(x(u), x_f(u))]$ are the second, fourth, and fifth cases. Thus, we can upper bound $\frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}} \mathbb{E}[c(x(u), x_f(u))]$ by:

$$\underbrace{\frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}|x_{\text{orig}}(u)\in[\theta'_M,\min(\theta',\theta_M)]} \mathbb{E}[c(x(u), x_f(u))]}_{(A)} + \underbrace{\frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}|x_{\text{orig}}(u)\in(\theta_M,\theta')} \mathbb{E}[c(x(u), x_f(u))]}_{(B)}$$

$$+ \underbrace{\frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}|x_{\text{orig}}(u)\in[\max(\theta',\theta_M),\theta]} \mathbb{E}[c(x(u), x_f(u))]}_{(C)}$$

For (A), we see that

$$\begin{aligned} \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}|x_{\text{orig}}(u)\in[\theta'_M,\min(\theta',\theta_M)]} \mathbb{E}[c(x(u), x_f(u))] &= \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}|x_{\text{orig}}(u)\in[\theta'_M,\min(\theta',\theta_M))} \mathbb{E}[c(x_{\text{orig}}(u), \theta')] \\ &\leq \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}|x_{\text{orig}}(u)\in[\theta'_M,\min(\theta',\theta_M))} \mathbb{E}[c(\theta'_M, \theta')] \\ &= \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}|x_{\text{orig}}(u)\in[\theta'_M,\min(\theta',\theta_M))} \gamma \\ &= \gamma \cdot \mathbb{P}_{\mathcal{D}_{\text{base}}}[x \in [\theta'_M, \min(\theta', \theta_M)))] \\ &\leq \gamma \cdot \mathbb{P}_{\mathcal{D}_{\text{base}}}[x \in [\theta'_M, \theta_M)]. \end{aligned}$$

For (B), we see that:

$$\begin{aligned} \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}|x_{\text{orig}}(u)\in(\theta_M,\theta')} \mathbb{E}[c(x(u), x_f(u))] &= \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}|x_{\text{orig}}(u)\in(\theta_M,\theta')} \mathbb{E}[c(\theta, \theta')] \\ &= c(\theta, \theta') \cdot \mathbb{P}_{\mathcal{D}_{\text{base}}}[x \in (\theta_M, \theta')] \\ &= \min(c(\theta, \theta'), \gamma) \cdot \mathbb{P}_{\mathcal{D}_{\text{base}}}[x \in (\theta_M, \theta')]. \end{aligned}$$

For (C), we see that:

$$\begin{aligned}
\frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}|x_{\text{orig}}(u)\in[\max(\theta',\theta_M),\theta]}\mathbb{E}[c(x(u),x_f(u))] &= \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}|x_{\text{orig}}(u)\in[\max(\theta',\theta_M),\theta]}\mathbb{E}[c(x_{\text{orig}}(u),\theta)] \\
&\le \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}|x_{\text{orig}}(u)\in[\max(\theta',\theta_M),\theta]}\mathbb{E}[\min\left(c(\theta',\theta),c(\theta_M,\theta)\right)] \\
&= \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}|x_{\text{orig}}(u)\in[\max(\theta',\theta_M),\theta]}\mathbb{E}[\min\left(c(\theta',\theta),\gamma\right)] \\
&= \min(c(\theta',\theta),\gamma)\cdot\mathbb{P}_{\mathcal{D}_{\text{base}}}[x\in[\max(\theta',\theta_M),\theta]].
\end{aligned}$$

Putting this all together, we obtain that:

$$\frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}}\mathbb{E}[c(x(u),x_f(u))] \le \gamma\cdot\mathbb{P}_{\mathcal{D}_{\text{base}}}[x\in[\theta'_M,\theta_M)] + \min(c(\theta',\theta),\gamma)$$

as desired. Since $\gamma\cdot\mathbb{P}_{\mathcal{D}_{\text{base}}}[x\in[\theta'_M,\theta_M]] + \min(c(\theta',\theta),\gamma)$ is decreasing in θ', this expression is maximized when $\theta' = \theta_{\min}$. Thus we obtain an upper bound of:

$$\gamma\cdot\mathbb{P}_{\mathcal{D}_{\text{base}}}[x\in[(\theta_{\min})_M,\theta_M)] + \min(c(\theta_{\min},\theta),\gamma).$$

□

A.9 Statement and proof of Proposition 11

Proposition 11. *Consider the 1-dimensional setup described in Section 4.3. Then, a symmetric solution $[\theta^*,\theta^*]$ is an equilibrium if and only if θ^*_M satisfies (5), where θ^*_M is the unique value such that $c(\theta^*_M,\theta^*) = \gamma$ and $\theta^*_M < \theta^*$. Both firms earn zero utility at this equilibrium. Moreover, the set $\mathcal{F}(\theta^*)$ of actions that a firm can take at equilibrium that achieve nonnegative utility is exactly equal to $[\theta^*,\infty)$, assuming the other firm chooses the classifier θ^*.*

Proof. The proof proceeds in two steps. First, we establish that $[\theta^*,\theta^*]$ is an equilibrium; next, we show that $[\theta,\theta]$ is not in equilibrium for $\theta\neq\theta^*$.

Establishing that $[\theta^*,\theta^*]$ is an equilibrium and $\mathcal{F}(\theta^*) = [\theta^*,\infty)$. First, we claim that $[\theta^*,\theta^*]$ is an equilibrium. At $[\theta^*,\theta^*]$, each participant chooses the first firm with $1/2$ probability. The expected utility earned by a firm is:

$$\begin{aligned}
\frac{1}{2}\int_{\theta^*_M}^{\infty} p_{\text{base}}(x)(p(x)-(1-p(x))\mathrm{d}x &= \int_{\theta^*_M}^{\infty} p_{\text{base}}(x)(p(x)-0.5)\mathrm{d}x \\
&= \int_{\theta^*_M}^{\infty} p_{\text{base}}(x)p(x)\mathrm{d}x - 0.5\int_{\theta^*_M}^{\infty} p_{\text{base}}(x)\mathrm{d}x \\
&= \int_{\theta^*_M}^{\infty} p_{\text{base}}(x)\mathrm{d}x\left(\frac{\int_{\theta^*_M}^{\infty} p_{\text{base}}(x)p(x)\mathrm{d}x}{\int_{\theta^*_M}^{\infty} p_{\text{base}}(x)\mathrm{d}x} - \frac{1}{2}\right) \\
&= \left(\int_{\theta^*_M}^{\infty} p_{\text{base}}(x)\mathrm{d}x\right)\left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_{\text{base}}}[y=1\mid x\ge\theta^*_M] - \frac{1}{2}\right) \\
&= 0.
\end{aligned}$$

If the firm chooses $\theta > \theta^*$, then since the cost function is strictly monotonic in its second argument, participants either are indifferent between θ and θ^* or prefer θ to θ^*. Due to the tie breaking rule, the firm will thus lose all of its participants and incur 0 utility. Thus the firm has no incentive to switch to θ.

If the firm chooses $\theta < \theta^*$, then it will gain all of the participants. Let θ_M is the unique value such that $c(\theta_M, \theta) = \gamma$ and $\theta_M < \theta$. The firm's utility will be:

$$\begin{aligned}
&\int_{\theta_M}^{\infty} p_{\text{base}}(x)(p(x) - (1 - p(x))\mathrm{d}x \\
&\qquad = \int_{\theta_M}^{\theta_M^*} p_{\text{base}}(x)(p(x) - (1 - p(x))\mathrm{d}x + \int_{\theta_M^*}^{\infty} p_{\text{base}}(x)(p(x) - (1 - p(x))\mathrm{d}x \\
&\qquad = 2\int_{\theta_M}^{\theta_M^*} p_{\text{base}}(x)(p(x) - 0.5)\mathrm{d}x.
\end{aligned}$$

It is not difficult to see that at θ^*, it must hold that $p(\theta_M^*) \leq 0.5$. Since the posterior is strictly increasing, this means that $p(\theta_M) < p(\theta_M^*) = 0.5$, so the above expression is negative. This means that the firm will not switch to θ_M.

Moreover, this establishes that $\mathcal{F}(\theta*) = [\theta^*, \infty)$.

$[\theta, \theta]$ is not in equilibrium if θ_M^* does not satisfy (5). Let θ_M is the unique value such that $c(\theta_M, \theta) = \gamma$ and $\theta_M < \theta$.

If $\theta < \theta^*$, then the firm earns utility

$$\frac{1}{2}\left(\int_{\theta_M}^{\infty} p_{\text{base}}(x)(p(x) - (1 - p(x))\right)\mathrm{d}x\,,$$

which we already showed above was negative. Thus, the firm has incentive to change their threshold to above θ so that it loses the full participant base and gets 0 utility.

If $\theta > \theta^*$, then the firm earns utility

$$U = \frac{1}{2}\left(\int_{\theta_M}^{\infty} p_{\text{base}}(x)(p(x) - (1 - p(x))\right)\mathrm{d}x\,,$$

which is strictly positive. Fix $\epsilon > 0$, and suppose that the firm changes to a threshold θ' such that $c(\theta', \theta) = \epsilon$. Then it would gain all of the participants and earn utility:

$$\begin{aligned}
\int_{\theta_M'}^{\infty} p_{\text{base}}(x)(p(x) - (1 - p(x))\mathrm{d}x &= \int_{\theta_M'}^{\theta_M} p_{\text{base}}(x)(p(x) - (1 - p(x))\mathrm{d}x + \int_{\theta_M}^{\infty} p_{\text{base}}(x)(p(x) - (1 - p(x))\mathrm{d}x \\
&= \int_{\theta_M'}^{\theta_M} p_{\text{base}}(x)(p(x) - (1 - p(x))\mathrm{d}x + 2U.
\end{aligned}$$

We claim that this expression approaches $2U$ as $\epsilon \to 0$. To see this, note that $c(\theta_M', \theta) \to \gamma$ and so $\theta_M' \to \theta_M$ as $\epsilon \to 0$. This implies that $\int_{\theta_M'}^{\theta_M} p_{\text{base}}(x)(p(x) - (1 - p(x))\mathrm{d}x \to 0$ as desired. Thus, the expression approaches $2U > U$ as desired. This means that there exists ϵ such that the firm changing to θ' results in a strict improvement in utility. □

A.10 Proof of Corollary 7

We apply Proposition 6 to see that the performative power is upper bounded by

$$B := L \min(c(\theta_{\min}, \theta^*), \gamma) + L\gamma \cdot \mathbb{P}_{x \in \mathcal{D}_{\text{base}}}\left[x \in [(\theta_{\min})_M, \theta^*_M]\right]$$

where (θ, θ) is a symmetric state. Using Proposition 2, we see that $\mathcal{F}(\theta^*) = [\theta^*, \infty)$. This means that $\theta_{\min} = \theta^*$, and so $(\theta_{\min})_M = \theta^*_M$. Thus, $B = 0$ which demonstrates that the performative power is upper bounded by 0, and is thus equal to 0.

A.11 Proof of Theorem 8

First, we show that we can lower bound performative power as the aggregate effect induced by a series of *swapped* score functions, one for each viewer. The score function $\tilde{s}_v$ associated with viewer u swaps the scores of content that currently appears in the first two display slots for viewer u and keeps all other scores unchanged. More formally, let s_{curr} denote the score function associated with the current predictor. Let $u \in \mathcal{U}$, let $i_1(u)$ and $i_2(u)$ be the indices of the content shown in the first and second slot: that is, the indices such that $c_{i_1(u)} = (c \circ s_{curr})(u)[1]$ and $c_{i_2(u)} = (c \circ s_{curr})(u)[2]$. The score function $\tilde{s}_u$ be the score function that swaps the scores for viewer u for $i_1(u)$ and $i_2(u)$. More formally, for $u' \neq u$, let $\tilde{s}_u(u') = s(u')$ and for $u' = u$, we define:

$$\tilde{s}_u(u)[i] := \begin{cases} s_{curr}(u)[i_2(u)] & \text{if } i = i_1(u), \\ s_{curr}(u)[i_1(u)] & \text{if } i = i_2(u) \\ s_{curr}(u)[i] & \text{else.} \end{cases}.$$

We claim that:

$$P \geq \mathbb{E}_{v \sim \mathcal{D}}\left[|z(u)[i_1(u)] - z_{\tilde{s}_u}(u)[i_1(u)]|\right] \tag{8}$$

The intuition behind equation (8) is that the performative power is lower bounded by the shift induced by $\tilde{s}_v$ (i.e. the difference between the counterfactual variables $z(v)$ and $z_{\tilde{s}_v}(v)$). This means that $P \geq \mathbb{P}_{\mathcal{D}}[u]|z(u)[i_1(u)] - z_{\tilde{s}_u}(u)[i_1(u)]|$ for any $u \in \mathcal{U}$. Our key insight is that we can aggregate these effects across all viewers u. In particular, by applying Assumption 2, we can construct a *single* score function $\tilde{s}$ that induces these effects for all viewers at the same time.

More formally, the score function $\tilde{s}$ is defined so that $\tilde{s}(v) = \tilde{s}_v(v)$ for all $v \in \mathcal{V}$. To lower bound performative power, we observe that:

$$\begin{aligned} P &\geq \mathbb{E}_{v \sim \mathcal{D}} \|z(u) - z_{\tilde{s}}(v)\|_\infty \\ &\geq \mathbb{E}_{v \sim \mathcal{D}}\left[|z(v)[i_1(v)] - z_{\tilde{s}}(v)[i_1(v)]|\right] \\ &=_{(1)} \mathbb{E}_{v \sim \mathcal{D}}\left[|z(v)[i_1(v)] - z_{\tilde{s}_v}(v)[i_1(v)]|\right] \\ &= \mathbb{E}_{v \sim \mathcal{D}}\left[|z(v)[i_1(v)] - z_{\tilde{s}_v}(v)[i_1(v)]|\right]. \end{aligned}$$

where (1) follows from applying Assumption 2.

We then apply the definition of β to see that

$$P \geq \mathbb{E}_{v \sim \mathcal{D}}\left[|z(v)[i_1(v)] - z_{\tilde{s}_v}(v)[i_1(v)]|\right] = \beta.$$

as desired.