Alternative Microfoundations for Strategic Classification

Meena Jagadeesan [1] **Celestine Mendler-Dünner** [1] **Moritz Hardt** [1]

Abstract

When reasoning about strategic behavior in a machine learning context it is tempting to combine *standard microfoundations* of rational agents with the statistical decision theory underlying classification. In this work, we argue that a direct combination of these ingredients leads to brittle solution concepts of limited descriptive and prescriptive value. First, we show that rational agents with perfect information produce discontinuities in the aggregate response to a decision rule that we often do not observe empirically. Second, when any positive fraction of agents is not perfectly strategic, desirable stable points—where the classifier is optimal for the data it entails—no longer exist. Third, optimal decision rules under standard microfoundations maximize a measure of negative externality known as *social burden* within a broad class of assumptions about agent behavior. Recognizing these limitations we explore *alternatives* to standard microfoundations for binary classification. We describe desiderata that help navigate the space of possible assumptions about agent responses, and we then propose the *noisy response* model. Inspired by smoothed analysis and empirical observations, noisy response incorporates imperfection in the agent responses, which we show mitigates the limitations of standard microfoundations. Our model retains analytical tractability, leads to more robust insights about stable points, and imposes a lower social burden at optimality.

1. Introduction

Consequential decisions compel individuals to react in response to the specifics of the decision rule. This individual-level response in aggregate can disrupt both statistical patterns and social facts that motivated the decision rule, leading to unforeseen consequences. A similar conundrum in the context of macroeconomic policy making fueled the *microfoundations* program following the influential critique of macroeconomics by Lucas in the 1970s (Lucas Jr, 1976). Microfoundations refers to a vast theoretical project that aims to ground theories of aggregate outcomes and population forecasts in microeconomic assumptions about individual behavior. Oversimplifying a broad endeavor, the hope was that if economic policy were *microfounded*, it would better anticipate the response that the policy induces.

Predominant in neoclassical economic theory is the assumption of an agent that exhaustively maximizes a utility function on the basis of perfectly accurate information. This modeling assumption about agent behavior underwrites many celebrated results on markets, mechanisms, and games. Although called into question by behavioral economics and related fields (e.g. see (Camerer et al., 2004)), the assumption remains central to economic theory and has become standard in computer science, as well.

When reasoning about incentives and strategic behavior in the context of classification tasks, it is tempting to combine the predominant modeling assumptions from microeconomic theory with the statistical decision theory underlying classification. In the resulting model, agents have perfect information about the decision rule and compute best-response feature changes according to their utility function with the goal of achieving a better classification outcome. We refer to this agent model as *standard microfoundations*. Assuming that agents follow this model, the decision maker then chooses the decision rule that maximizes their own objective in anticipation of the resulting agent response. This is the conceptual route taken in the area of *strategic classification*, but similar observations may apply more broadly to the intersection of economics and learning.

1.1. Our work

We argue that standard microfoundations are a poor basis for studying strategic behavior in binary classification. We make this point through three observations that illustrate the limited descriptive power of the standard model and the problematic solution concepts it implies. In response, we explore the space of alternative agent models for strategic classification, and we identify desiderata that when satisfied by microfoundations lead to more realistic and robust

[1]University of California, Berkeley. Correspondence to: Meena Jagadeesan <mjagadeesan@berkeley.edu>.

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021.

insights. Guided by these desiderata, we propose *noisy response* as a promising alternative to the standard model.

Limitations of standard microfoundations. In strategic classification, agents respond strategically to the deployment of a binary decision rule f_θ specified by classifier weights θ. The decision-maker assumes that agents follow standard microfoundations: that is, agents have full information about f_θ and change their features so as to maximize their utility function. The utility function captures the benefit of a positive classification outcome, as well as the cost of feature change. Consequently, an agent only invests in changing their features if the cost of feature change does not exceed the benefit of positive classification.

Our first observation concerns the *aggregate response*—the distribution $\mathcal{D}(\theta)$ over feature, label pairs induced by a classifier f_θ. We show that in the standard model, the aggregate response necessarily exhibits discontinuities that we often do not observe in empirical settings. The problem persists even if we assume an approximate best response or allow for heterogeneous cost functions.

Our second observation reveals that, apart from lacking descriptive power, the standard model also leads to brittle conclusions about the solution concept of *performative stability*. Performative stability (Perdomo et al., 2020) refers to decision rules that are optimal on the particular distribution they entail. Stable points thus represent fixed points of *retraining methods*, which repeatedly update the classifier weights to be optimal on the data distribution induced by the previous classifier weights. We show that the existence of performatively stable classifiers breaks down whenever a positive fraction of randomly chosen agents in the population are non-strategic. This brittleness suggests that the standard model does not constitute a reliable basis for investigating dynamics of retraining algorithms.

Our last observation concerns the solution concept of *performative optimality*. Performative optimality (Perdomo et al., 2020) refers to a decision rule that exhibits the highest accuracy on the distribution it induces. The global nature of this solution concept means that finding performatively optimal points requires the decision maker to anticipate strategic feedback effects. We prove that relying on standard microfoundations to model strategic behavior leads to extreme decision rules that maximize a measure of negative externality called *social burden* within a broad class of alternative models. Social burden, proposed in recent work, quantifies the expected cost that positive instances of a classification problem have to incur in order to be accepted. Standard microfoundations thus produce optimal solutions that are least favorable in terms of social burden.

Alternative microfoundations. We systematically explore alternative assumptions on agent responses, encompassing general agent behavior that need not be fully informed, strategic, or utility maximizing. We formalize microfoundations as a randomized mapping $M : X \times Y \to \mathcal{T}$ that assigns each agent to a response type $t \in \mathcal{T}$. The response type t is associated with a response function $\mathcal{R}_t \colon X \times \Theta \to X$ specifying how agents of type t change their features x in response to each decision rule f_θ. Letting $\mathcal{D}_{XY}$ be the base distribution over features and labels prior to any strategic adaptation, the *aggregate response* to a classifier f_θ is given by the distribution $\mathcal{D}(\theta; M)$ over induced feature, label pairs $(\mathcal{R}_t(x, \theta), y)$ for a random draw $(x, y) \sim \mathcal{D}_{XY}$ and $t = M(x, y)$. The mapping M thus *microfounds* the distributions induced by decision rules, endowing the distributions with structure that allows the decision maker to deduce the aggregate response from a model of individual behavior.

We describe a collection of properties that are desirable for a model of agent responses to satisfy. The first condition, that we call *aggregate smoothness* rules out discontinuities arising from standard microfoundations. Aggregate smoothness requires that varying the classifier weights slightly must change the aggregate response smoothly. We find that this property alone is sufficient to guarantee the robust existence of stable points under mixtures with non-strategic agents.

The second condition, that we call the *expenditure constraint*, helps ensure that the model encodes realistic agent-level responses $\mathcal{R}_t$. At a high level, it requires that each agent does not spend more on changing their features than the utility of a positive outcome. This natural constraint gives rise to a large set of potential models. For any such model that satisfies a weak assumption, the social burden of the optimal classifier is no larger than the social burden of the optimal classifier deduced from standard microfoundations. Moreover, the optimal points are determined by local behavior, thus making it more tractable to find an approximately optimal classifier.

Noisy response. We identify *noisy response* as a compelling alternative to standard microfoundations for strategic classification. In this model, each agent best responds with respect to $\theta + \xi$, where ξ is an independent sample from a zero mean noise distribution. This model is inspired by *smoothed analysis* (Spielman and Teng, 2009) and encodes imperfection in the population's response to a decision rule by perturbing the manipulation targets of individual agents.

Noisy response satisfies many desirable properties. First, it satisfies aggregate smoothness, and thus leads to the robust existence of stable points. Moreover, the model satisfies the expenditure constraint, and thus encodes natural agent responses which can be used to reason about metrics such as social burden. When used to anticipate strategic feedback effects and compute optimal points, noisy response leads to strictly less pessimistic acceptance thresholds than those computed under standard microfoundations. In fact,

we show via simulations that a larger variance of the noise in the manipulation target leads to more conservative optimal thresholds. Finally, the aggregate distribution induced by noisy response can be estimated from individual experiments alone, without ever deploying a classifier.

1.2. Related work

Existing work on strategic classification has mostly followed standard microfoundations for modeling agent behavior in response to a decision rule, e.g., (Dalvi et al., 2004; Brückner and Scheffer, 2011; Hardt et al., 2016a; Khajehnejad et al.; Tsirtsis and Gomez-Rodriguez, 2020) to name a few. This includes works that focus on minimizing Stackelberg regret (Dong et al., 2018; Chen et al., 2020), quantify the price of transparency (Akyol et al., 2016), and investigate the benefits of randomization in the decision rule (Braverman and Garg, 2020). Investigations of externalities such as social cost (Milli et al., 2019; Hu et al., 2019) whether classifiers incentivize improvement as opposed to gaming (Kleinberg and Raghavan, 2019; Miller et al., 2020; Shavit et al., 2020; Haghtalab et al., 2020), and practical considerations for optimization (Levanon and Rosenfeld, 2021) have also mostly built on standard microfoundations.

A handful of works have suggested potential limitations of the standard strategic classification framework. Brückner et al. (2012) recognized that the standard model leads to very conservative Stackelberg solutions, and proposed resorting to Nash equilibria as an alternative solution concept. We instead advocate for altogether rethinking standard microfoundations. Concurrent and independent work by Ghalme et al. (2021) and Bechavod et al. (2021) relaxed the perfect information assumption in the standard model and studied strategic classification when the classifier is not fully revealed to the agents. In this work, we argue that the agents often do not perfectly respond to the classifier even when the decision rule is fully transparent.

Related work in economics also investigates strategic responses to decision rules. This line of work, initiated by Spence (1973), has investigated muddled information about individuals from heterogeneous gaming behavior (Frankel and Kartik, 2019), the role of commitment power of the decision maker (Frankel and Kartik, 2020), the aggregation of multi-dimensional features (Ball, 2020), and the performance of different training approaches (Hennessy and Goodhart, 2020). A notable work by Björkegren et al. (2020) investigates strategic behavior through a field experiment in the micro-lending domain. An important distinction is that these works tend to study regression, while we focus on classification. These settings appear to be qualitatively different in the context of strategic feedback effects; e.g. see note in (Hennessy and Goodhart, 2020).

Our work is conceptually related to recent work in economics that has recognized mismatches between the predictions of standard models and empirical realities, for example in macroeconomic policy (Stiglitz, 2018; Kaplan and Violante, 2018; Coibion et al., 2018) and in mechanism design (Li, 2017). These works, and many others, have explored incorporating richer behavioral and informational assumptions into the typical models used in economic settings.

1.3. Setup and basic notation

Let $X \subseteq \mathbb{R}^m$ denote the feature space, and let $Y = \{0, 1\}$ be the space of binary outcomes. Each agent is associated to a feature vector $x \in X$ and a binary outcome $y \in Y$ which represents their true label. A feature, label pair (x, y) need not uniquely describe an agent, and many agents may be associated to the same pair (x, y). The base distribution $\mathcal{D}_{XY}$ is a joint distribution over $X \times Y$ describing the population prior to any strategic adaption. We assume that $\mathcal{D}_{XY}$ is continuous and has zero mass on the boundary of X. We focus on binary classification where each classifier $f_\theta : X \to \{0, 1\}$ is parameterized by $\theta \in \mathbb{R}^d$, and the decision-maker selects classifier weights θ from $\Theta \subseteq \mathbb{R}^d$ which is a compact, convex set.[1] We adopt the notion of a distribution map $\mathcal{D}(\theta)$ from (Perdomo et al., 2020) to describe the distribution over $X \times Y$ induced by strategic adaptation of agents drawn from the base distribution in response to the classifier f_θ.

2. Limitations of standard microfoundations

In the strategic classification literature, the typical agent model is a *rational agent with perfect information*. The core assumption is that agents have perfect knowledge of the classifier and maximize their utility given the classifier weights. The utility consists of two terms: a reward for obtaining a positive classification and a cost of manipulating features. The reward is denoted $\gamma > 0$ and the manipulation cost is represented by a function $c : X \times X \to \mathbb{R}$ where $c(x, x')$ reflects how much agents need to expend to change their features from x to x'. A valid cost function satisfies a natural monotonicity requirement as stated in Assumption 1. Given a feature vector x and a classifier f_θ, agents solve the following utility maximization problem:

$$\arg\max_{x' \in X} \left[\gamma f_\theta(x') - c(x, x')\right]. \tag{1}$$

We will refer to this model as *standard microfoundations*.

Assumption 1. A cost function $c : X \times X \to \mathbb{R}$ is *valid*, if it is continuous in both arguments, it holds that $c(x, x') = 0$ for $x = x'$, and c increases with distance[2] in the sense that $c(x, \bar{x}) < c(x, x')$ and $c(\bar{x}, x) < c(x', x)$ for every $\bar{x} \in X$ that

[1] We assume that for every $\theta \in \Theta$, the set $\{x \in X \mid f_\theta(x) = 1\}$ is closed, and the decision boundary is measure 0.

[2] We model a non-zero cost for all modifications to features, regardless of whether they result in positive classification or not.

lies on the line segment connecting the two points $x, x' \in X$.

2.1. Discontinuities in the aggregate response

A striking property of distributions induced by standard microfoundations in response to binary classifiers is that they are either trivial or discontinuous. The underlying cause is that agents behaving according to standard microfoundations either change their features exactly up to the decision boundary, or they do not change their features at all.

Proposition 1. *Given a base distribution $\mathcal{D}_{XY}$, let $\mathcal{D}(\theta)$ be the distribution induced by a classifier f_θ. Then, if $\mathcal{D}(\theta)$ is continuous and $\mathcal{D}(\theta) \neq \mathcal{D}_{XY}$, there does not exist a valid cost function c such that $\mathcal{D}(\theta)$ is an aggregate of agents following standard microfoundations.*

In addition to the discontinuities implied by Proposition 1, the aggregate response induced by standard microfoundations faces additional degeneracies. Namely, any non-trivial distribution arising from the standard model must have a region of zero density below the decision boundary.

These properties are unnatural in many applications.

Example 1 (Bank lending decisions and credit scores)**.** If lending decisions are based on FICO scores, then under standard microfoundations, the distribution over credit scores should exhibit a discontinuity at the threshold. However, this is not what we observe empirically. In particular, previous work (Hardt et al., 2016b) studied a FICO dataset from 2003, where credit scores range from 300 to 850, and a cutoff of 620 was commonly used for prime-rate loans. The observed distribution over credit scores appears continuous and is supported across the full range of scores.[3]

Example 2 (Yelp online ratings and rounding thresholds)**.** Restaurant ratings on Yelp are rounded to the nearest half star, and the star rating of a restaurant influences restaurant customer flows. Strategic adaptation can arise from restaurants leaving fake reviews. Under standard microfoundations, the distribution of restaurant ratings would exhibit discontinuities at the rounding thresholds. However, previous work (Anderson and Magruder, 2012) examined the distribution of restaurant ratings, and showed that there is no significant discontinuity in the density of restaurant ratings. at the rounding thresholds (see Figure 4 in their work).

Similar observations apply to New York high school exit exams where students and teachers are rewarded when student test scores meet designated cutoffs (Dee et al., 2019).[4]

It is important to note that the degeneracies of standard microfoundations arise from the fact that classification decisions are discrete and based on a *hard* decision. Agents who are not classified positively receive no reward: it does not matter how close to the decision boundary the agent is. This discontinuity in the utility does not arise in regression problems. However, in machine learning and statistical decision theory, binary classification is ubiquitous, and degeneracies that we have identified pertain to general settings where the decisions are binary.

The reader might imagine that common variations and generalizations of standard microfoundations can mitigate these issues. Unfortunately, the two variations of standard microfoundations that are typically considered—*heterogeneous cost functions* (Hu et al., 2019), and *approximate best response* (Miller et al., 2020)—result in similar degeneracies. Heterogeneity in the cost (or utility) function can only change whether or not an agent decides to change their features, but it does not change their target of manipulation. If agents approximately best-respond, and thus move to features x' that approximately maximize their utility, the model no longer leads to point masses at the decision boundary, but agents will never *undershoot* the decision boundary. This means that any nontrivial aggregate distribution must have a region of zero density below the decision boundary to comply with standard microfoundations or these variants.

Agent behavior that is inconsistent with standard microfoundations and these variants has been observed in field experiments.

Example 3 (Field Experiment (Björkegren et al., 2020))**.** The authors deployed an app in Kenya that mimicked aspects of "digital credit" applications to empirically investigate strategic behavior. Participants were rewarded if the app guessed that they were a high-income earner. When the participants were given access to the coefficients of the decision rule, they tended to change their features in the right direction, but exhibited a high variance in their responses—see Table 5 in their work. Moreover, when participants were given opaque access to the decision rule, agents often did not even change their features in the right direction.

2.2. Brittleness under natural model misspecifications

We describe two scenarios where relying on standard microfoundations to model agent behavior leads to undesirable properties under natural model misspecifications.

2.2.1. Stability as a fragile solution concept

Our first result demonstrates that the standard model does not lead to robust insights about performative stability. In

[3] There could be many reasons why we do not observe discontinuities, e.g. different lenders have different cutoffs, or features beyond credit scores are used. In any case, the typical strategic classification model does not describe this setting accurately.

[4] Interestingly, the distribution over test scores did appear discontinuous prior to 2012 as a result of *teachers* directly manipulating student test scores. Following reforms to the grading procedure, the test score distribution became continuous, demonstrating that strategic adaptation by *students* does not result in discontinuities.

particular, although performatively stable solutions are guaranteed to exist under standard microfoundations (see (Milli et al., 2019)), stable points do not exist when any positive fraction of randomly chosen individuals are non-strategic.

For our analysis, we consider a local relaxation of the notion of performative stability that corresponds to fixed points of *repeated gradient descent* (Perdomo et al., 2020). We say θ_{PS} is *locally stable* if θ_{PS} is a local minimum or a stationary point of the following optimization problem:

$$\min_{\theta \in \Theta} \mathbb{E}_{(x,y)\sim\mathcal{D}(\theta_{\mathrm{PS}})} \mathbb{1}\{y \neq f_\theta(x)\}. \tag{2}$$

Local stability is closely related to the concept of a pure strategy (local) *Nash equilibrium* in a simultaneous game between the strategic agents and the decision maker who responds to the observed distribution $\mathcal{D}(\theta)$.

To showcase that the existence of locally stable classifiers under standard microfoundations crucially relies on all agents following the modeling assumptions, we focus on the following simple 1-dimensional setting.

Setup 1 (1-dimensional)**.** Let $X \subseteq \mathbb{R}$ and consider a threshold functions $f_\theta(\cdot) = \mathbb{1}\{\cdot \geq \theta\}$ with $\theta \in \Theta \subseteq \mathbb{R}$. Let $\mu(x)$ be the conditional probability over $\mathcal{D}_{XY}$ of the true label being 1 given features x. Suppose that $\mu(x)$ is strictly increasing in x and there is an $\theta \in \mathrm{Int}(\Theta)$ such that $\mu(\theta) = 0.5$.

Proposition 2. *Consider Setup 1. Suppose that a p fraction of agents drawn from $\mathcal{D}_{XY}$ do not ever change their features, and a $1-p$ fraction of agents drawn independently from $\mathcal{D}_{XY}$ follow standard microfoundations with a valid cost function c. Then, we have the following properties:*

a) For $p \in \{0,1\}$, locally stable points exist. (For $p \in \{0,1\}$, let θ^p_{PS} be the smallest locally stable point.)

b) For $p \in (0,1)$, locally stable points do not exist.

c) For $p \in (0,1)$, RRM will oscillate between θ^1_{PS} and a threshold $\tau(p) \in (\theta^1_{\mathrm{PS}}, \theta^0_{\mathrm{PS}})$, where $\tau(p)$ is decreasing in p, approaching θ^1_{PS} as $p \to 1$ and θ^0_{PS} as $p \to 0$.

Proposition 2 implies that not only does the existence of locally stable points break down if a positive fraction $p \in (0,1)$ of randomly chosen agents are non-strategic, but also repeated risk minimization oscillates between two extreme points. The proof can be found in Appendix B.2. For illustration purposes, we have implemented a simple instantiation of Setup 1, and we visualize the trajectories of RRM for different values of p in Figure 1(a). The main insight is that retraining methods start oscillating substantially even when p is very close to 0 (only an ϵ fraction of agents are not following standard microfoundations). This sensitivity of the trajectory to natural deviations from the modeling assumptions suggests that standard microfoundations do not constitute a reliable model to study algorithm dynamics.

2.2.2. Negative externalities at optimality

Our next result shows that standard microfoundations do not constitute a good representative model of agent behavior for investigating qualitative properties of optimal solutions. We present a natural scenario where performatively optimal classifiers computed under standard microfoundations lead to the highest negative externalities within a broad class of alternative models for agent responses.

Recall that a performatively optimal solution corresponds to the best classifier for the decision maker from a global perspective, but it is not necessarily stable under retraining. These solutions are closely related to Stackelberg equilibria in the game between the agents and the decision maker. Formally, a classifier θ_{PO} is *performatively optimal* (Perdomo et al., 2020) if it minimizes the performative risk:

$$\theta_{\mathrm{PO}}(\mathcal{D}) := \mathrm{argmin}_{\theta\in\Theta} \mathbb{E}_{(x,y)\sim\mathcal{D}(\theta)} \mathbb{1}\{y \neq f_\theta(x)\}. \tag{3}$$

The key challenge of computing performative optima is that optimizing (3) requires the decision maker to anticipate the population's response $\mathcal{D}(\theta)$ to any classifier f_θ. A natural approach to model this response is to build on microfoundations and deduce properties of the distribution map from individual agent behavior.

While the decision-maker is unlikely to have a fully specified model for agent behavior at hand, we outline a few natural criteria that agent responses could reasonably satisfy. To formalize these criteria, we again focus on the 1-dimensional setting.

Property 1 (Expenditure monotonicity)**.** *For every $x \in X$, any agent a with true features x must have manipulated features $\mathcal{R}_a(x;\theta)$ in response to f_θ that satisfy:*

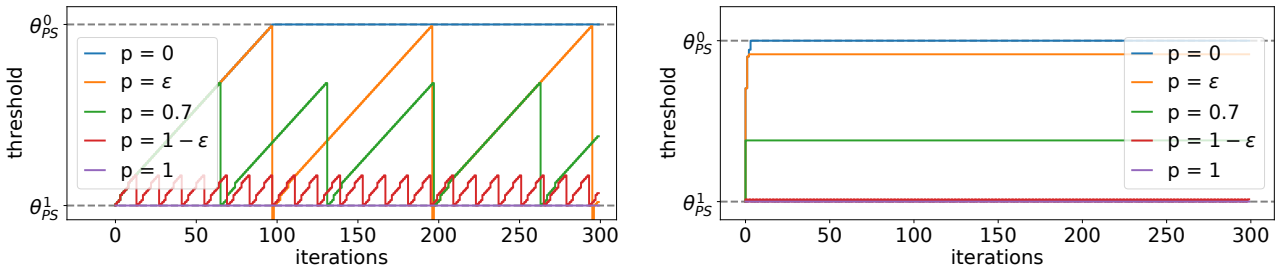
a) $c(x, \mathcal{R}_a(x;\theta)) < \gamma$ for every $\theta \in \Theta$.

b) $f_\theta(\mathcal{R}_a(x;\theta)) = 1 \implies f_{\theta'}(\mathcal{R}_a(x;\theta')) = 1 \ \forall \theta' \leq \theta$.

Property 1 describes agents that a) do not expend more on gaming than their utility from a positive outcome, and b) do not have their outcome worsened if the threshold is lowered. However, agents complying with Property 1 do not necessarily behave according to standard microfoundations. For example, Property 1 is satisfied by non-strategic agents who do not ever change their features and by the imperfect agents that we describe in Section 4.

We now show that within the class of microfoundations that exhibit Property 1, the standard model leads to an extreme acceptance threshold. The formal statement of our result can be found in Appendix B.3.

Proposition 3 (Informal)**.** *Consider Setup 1. Let $\mathscr{D}$ be the class of distribution maps $\mathcal{D} : \Theta \to \Delta(X \times Y)$ that can be represented by a population of agents who all satisfy Property 1. Then under mild assumptions, for every distribution*



(a) SM mixed with non-strategic agents

(b) NR mixed with non-strategic agents

Figure 1. Convergence of retraining algorithm in a 1d-setting for different values of p with $\epsilon = 10^{-2}$. The population consists of 10^5 individuals. Half of the individuals are sampled from $x \sim \mathcal{N}(1, 0.33)$ with true label 1 and the other half is sampled from $x \sim \mathcal{N}(0, 0.33)$ with true label 0. θ^0_{PS} and θ^1_{PS} are defined as in Proposition 2 for standard microfoundations (and similarly for noisy response). The parameter of the noisy responses (NR) in (b) is taken to be $\sigma^2 = 0.1$.

map $\mathcal{D} \in \mathscr{D}$, it holds that

$$\theta_{\text{PO}}(\mathcal{D}_{SM}) \geq \theta_{\text{PO}}(\mathcal{D})$$

where $\mathcal{D}_{SM}$ is the distribution map induced by standard microfoundations.

A problematic implication of Proposition 3 is that standard microfoundations also maximize the negative externality called *social burden* (Milli et al., 2019):

$$\text{Burden}(\theta) := \mathbb{E}_{(x,y)\in\mathcal{D}_{XY}}[\min\{c(x,x') \mid f_\theta(x') = 1\} \mid y = 1].$$

Social burden quantifies the average cost that a positively labeled agent has to expend in order to be positively classified by f_θ. While previous work introduced and studied social burden within standard microfoundations, we use it to study implications of different modeling assumptions on agent behavior. In particular, the following corollary demonstrates that standard microfoundations lead to *worst possible social burden* across all microfoundations that satisfy Property 1.

Corollary 4. *Under the same assumptions as Proposition 3, for every distribution map $\mathcal{D} \in \mathscr{D}$, it holds that*

$$\text{Burden}(\theta_{\text{PO}}(\mathcal{D}_{SM})) \geq \text{Burden}(\theta_{\text{PO}}(\mathcal{D})).$$

where $\mathcal{D}_{SM}$ is the distribution map induced by standard microfoundations.

This result has implications for the likely situation where standard microfoundations do not exactly describe agent behavior. In particular, relative to the performative optimal point of the true agent responses, the solutions computed using standard microfoundations would not only experience suboptimal performative risk but also would cause unnecessarily high social burden. Thus, under natural modeling misspecification, it is hard for the decision-maker to justify using standard microfoundations. Implicit in our argument is the following moral stance: *given a set of criteria for what defines a plausible model for microfoundations, the decision-maker should not select the one that maximizes negative externalities*.

3. Alternative microfoundations

In this section, we depart from this classical approach and systematically search for models that are more appropriate for binary classification. We define the space of alternatives and collect a set of useful properties that we show are desirable for microfoundations to satisfy. These properties serve as a "compass" to guide our search for an alternative microfoundations for strategic classification in Section 4.

3.1. Defining the space of alternatives

The principle behind microfoundations for strategic classification is to equip the distribution map with structure by viewing the distribution induced by a decision rule as an aggregate of the responses of individual agents. We consider a space of alternative microfoundations that capture agent responses in full generality. We introduce a family of response types $\mathcal{T}$ that represents the space of all possible ways that agents can react to the classifier f_θ. The response type fully determines agent behavior through the *agent response function* $\mathcal{R}_t : X \times \Theta \to X$. In particular, an agent with true features x and response type t changes their features to $x' = \mathcal{R}_t(x, \theta)$ when the classifier f_θ is deployed.

Remark. Non-strategic agents and the standard microfoundations each correspond to one response type. In our framework a population of agents could exhibit a mixture of different types, or even be described by a *continuum* of types.

We formalize microfoundations through a *mapping* $M : X \times Y \to \mathcal{T}$ from agents to response types. We denote the set of possible mappings M by the collection $\mathcal{M}$ that consists of all[5] possible randomized functions $X \times Y \to \mathcal{T}$. Conceptually, a mapping $M \in \mathcal{M}$ sets up the rules of agent behavior. The response types directly specify agent responses, rather

[5]These mappings are subject to mild measurability constraints that we describe in Appendix A.2.

than specifying an underlying behavioral mechanism—this is an aspect that distinguishes our framework from the approach to microfoundations in economics. An advantage is that responses can be observed, whereas the behavioral mechanism is harder to infer.

Importantly, the mapping M coupled with the base distribution $\mathcal{D}_{XY}$ fully specifies the population's response to a classifier f_θ. In particular, for each $\theta \in \Theta$, the *aggregate response* $\mathcal{D}(\theta; M)$ is the distribution over $(\mathcal{R}_t(x,\theta), y)$ where $(x,y) \sim \mathcal{D}_{XY}$ and $t = M(x,y)$. We use the notation $\mathcal{D}(\cdot; M) : \Theta \to \Delta(X \times Y)$ to denote the aggregate response map induced by M. The mapping M thus provides sufficient information to reason about the performative risk; in addition, it also provides sufficiently fine-grained information about individuals to reason about metrics such as social burden.

Naturally, with such a flexible model, *any* distribution map can be microfounded. We provide a formal proof of this existence result in Appendix C.1. In the following subsections, we focus on narrowing down the space of candidate models and describe two properties that we believe microfoundations should satisfy.

3.2. Aggregate smoothness

The first property we describe pertains to the induced distribution and its interactions with the function class. This aggregate-level property rules out unnatural discontinuities in the distribution map. We call this property *aggregate smoothness*, and formalize it in terms of the *decoupled performative risk* (Perdomo et al., 2020).

Property 2 (Aggregate smoothness)**.** *Define the decoupled performative risk induced by M to be* $\mathsf{DPR}_M(\theta, \theta') := \mathbb{E}_{(x,y)\sim\mathcal{D}(\theta;M)}[\, \mathbb{1}\{y \neq f_{\theta'}(x)\}]$. *For a given base distribution $\mathcal{D}_{XY}$, a mapping M satisfies aggregate smoothness if the derivative of the decoupled performative risk with respect to θ' exists and is continuous in θ and θ' across all of Θ.*

Intuitively, the existence and the continuity of the partial derivative of $\mathsf{DPR}_M(\theta, \theta')$ with respect to θ' guarantee that[6]

a) *each distribution $\mathcal{D}(\theta; M)$ is sufficiently continuous (and has no point mass at the decision boundary)*,
b) *$\mathcal{D}(\theta; M)$ changes continuously in θ.*

We believe that these two continuity properties are natural and likely to capture practical settings, given the empirical evidence in Examples 1-3. A consequence of aggregate smoothness is that it is sufficient to guarantee the existence of locally stable points.

Theorem 5. *Given a base distribution $\mathcal{D}_{XY}$ and function class Θ, for any mapping M that satisfies aggregate smoothness, there exists a locally stable point.*

In fact, Theorem 5 implies that stable points exist under deviations from the model, as long as aggregate smoothness is preserved. Our next result shows that under weak assumptions on the base distribution this is the case for any mixture with non-strategic agents. For ease of notation, we formalize such a mixture through the operator $\Phi_p(M)$, where for $p \in [0,1]$, we let $\Phi_p(M(x,y))$ be equal to t_{NS} with probability p and equal to $M(x,y)$ otherwise.

Proposition 6. *Suppose that the non-performative risk $\mathbb{E}_{(x,y)\in\mathcal{D}_{XY}} \mathbb{1}\{f_\theta(x) = y\}$ is continuously differentiable for all $\theta \in \Theta$. Then, for any $p \in [0,1]$, aggregate smoothness of a mapping M is preserved under the operator $\Phi_p(M)$.*

Proposition 6, together with Theorem 5, implies the robust existence of locally stable points under mixtures with non-strategic agents, for any microfoundations model that satisfies aggregate smoothness.

Conceptually, our investigations in this section have been inspired by Perdomo et al. (2020) that demonstrated that regularity assumptions on the aggregate response can guarantee the existence of stable points for smooth, strongly convex loss functions. However, our results differ since we instead focus on the 0-1 loss function.

3.3. Constraint on manipulation expenditure

While aggregate smoothness focused on the population-level properties of the induced distribution, a model for microfoundations must also be descriptive of realistic agent-level responses in order to yield useful qualitative insights about metrics such as social burden or accuracy on subgroups. A minimal assumption is that an agent never expends more on manipulation than the utility of a positive outcome.

Property 3 (Expenditure constraint)**.** *Given a function class Θ and a cost function c, a mapping $M \in \mathcal{M}$ is expenditure-constrained if $c(x, \mathcal{R}_t(x,\theta)) \leq \gamma$ for every $\theta \in \Theta$ and $t \in$ Image(M).*

This constraint is implicitly encoded in standard microfoundations and many of its variants. We have previously encountered the expenditure constraint in Section 2.2, where we showed that if c is a valid cost function, then this property, together with a basic monotonicity requirement on feature manipulations, defines a set of microfoundations models among which the standard model achieves maximal social burden at optimality. In Section 4 we will focus on one model within this set which results in a *strictly* lower social burden than the standard model.

Remark (Reducing the complexity of estimating the distribution map)**.** An additional advantage of Property 3 is that it constrains each agent's range of manipulations. This can

[6] This correspondence between aggregate smoothness, and continuity of the distribution map can be made explicit for 1-dimensional features: see Appendix C.2.

significantly reduce the complexity of estimating the distribution map for a decision-maker who wants to compute a strategy robust classifier offline. While Zhang et al. (2021) directly specify the set of feature changes that an agent may make, we find that the expenditure constraint alone can reduce the complexity of estimating the distribution map. Namely, the expenditure constraint narrows down the set of agents who can cross the decision boundary, and so the decision-maker only needs to learn a small portion of the distribution map. We defer the formal result to Appendix E.

4. Microfoundations from imperfect agents

Using the properties established in the previous section as a guide, we propose an alternate model for microfoundations that naturally allows agents to undershoot or overshoot the decision boundary, while complying with aggregate smoothness and expenditure monotonicity. Furthermore, we show that this model, called *noisy response*, leads to strictly smaller social burden than the standard model while retaining analytical tractablility.

4.1. Noisy response

Noisy response captures the idea of an *imperfect agent* who does not perfectly best-respond to the classifier weights. This imperfection can arise from many different sources—including interpretability issues, imperfect control over manipulations, or opaque access to the classifier. Inspired by *smoothed analysis* (Spielman and Teng, 2009), we do not directly specify the source of imperfection but instead capture imperfection in an agnostic manner, by adding small random perturbations to the classifier weights targeted by the agents. Since smoothed analysis has been successful in explaining convergence properties of algorithms in practical (instead of worst case) situations, we similarly hope to better capture empirically observed strategic phenomena.

We define the relevant set of types $T_{\text{noisy}} \subset \mathcal{T}$ so that each type $t \in T_{\text{noisy}}$ is associated with noise $\eta_t \in \mathbb{R}^m$. An agent of type t perceives θ as $\theta + \eta_t$ and responds to the classifier f_θ as follows:

$$\mathcal{R}_t(x,\theta) := \arg\max_{x' \in X'} \Big[\gamma \cdot f_{\theta+\eta_t}(x') - c(x,x')\Big], \qquad (4)$$

where c denotes a valid cost function, $\gamma > 0$ denotes the utility of a positive outcome, and $X' \subseteq \mathbb{R}^d$ is a compact, convex set containing X.[7] For each $(x,y) \in X \times Y$, we model the distribution over noise across all agents with feature, label pair (x,y) as a multivariate Gaussian. To formalize this, we define a *randomized* mapping $M_\sigma : X \times Y \to \mathcal{T}$ as follows. For each (x,y), the random variable $M_\sigma(x,y)$ is defined so that if $t \sim M_\sigma(x,y)$, then η_t is distributed as $\mathcal{N}(\mathbf{0}, \sigma^2 I)$. This model results in the perceived values of θ across all agents with a given feature, label pair following a Gaussian distribution centered at θ. The noise level σ reflects the degree of imperfection in the population.[8]

Conceptually, our model of noisy response bears resemblance to models of *incomplete information* (Harsanyi, 1968) that are standard in game theory (but that have not been traditionally considered in the strategic classification literature). However, a crucial difference is that we advocate for modeling agents actions as imperfect even if the classifier is fully transparent, because we believe that imperfection can also arise from other sources. This is supported by the empirical study from Example 3 where agents act imperfectly even when the classifier weights are revealed.

4.2. Aggregate-level properties of noisy response

Intuitively, the noise in the manipulation target of noisy response smooths out the discontinuities of standard microfoundations, eliminating the point mass at the decision boundary and region of zero density below the boundary. We show this explicitly in a 1-dimensional setting.

Proposition 7. *Let $X \subseteq \mathbb{R}$, and let $\Theta \subseteq \mathbb{R}$ be a model class of threshold functions. For any $\sigma \in (0,\infty)$, the mapping M_σ satisfies aggregate smoothness.*

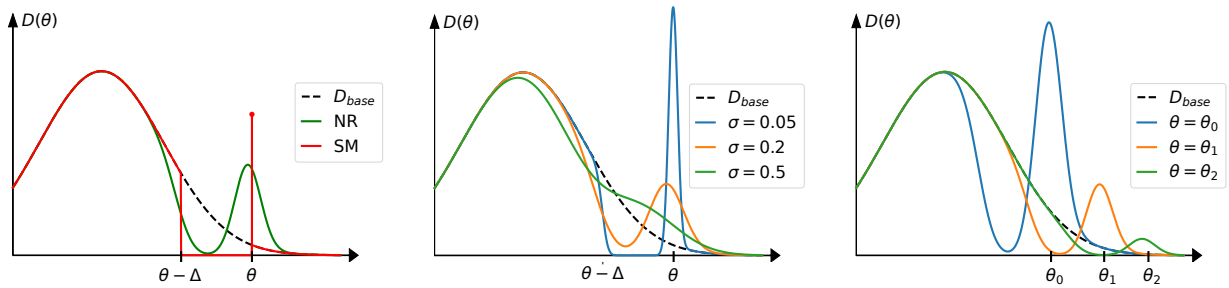
Remark. Proposition 7 implies that noisy response inherits the robust existence of stable points from Theorem 5. Furthermore, Figure 1(b) illustrates that noisy response mitigates the oscillations of repeated retraining that we observed for standard microfoundations.

To visualize the aggregate-level properties of noisy response and compare them to standard microfoundations we depict the respective density functions for a 1-dimensional Gaussian base distribution in Figure 2(a). The distribution $\mathcal{D}(\theta)$ can be bimodal, since agents closer to the threshold θ are more likely to change their features. The shape of the response distribution also changes with σ as illustrated in Figure 2(b). As $\sigma \to 0$, the aggregate response of a population of noisy response agents approaches that of standard microfoundations, so noisy response can approximate the aggregate response of standard microfoundations to arbitrary accuracy. Finally, the distribution map of noisy response changes *continuously* with θ, as visualized in Figure 2(c).

In fact, the distribution map induced by noisy response is *Lipschitz* in total-variation distance for any valid cost function. For smooth and strongly convex loss functions, this implies convergence of repeated retraining (Perdomo et al., 2020; Mendler-Dünner et al., 2020).

[7] We assume that c is defined on all of $X' \times X'$, and $c(x,x') > \gamma$ for all $x \in X$ and all x' that are on the boundary of X'.

[8] While we focus on Gaussian noise in the perception function throughout this work, the outlined benefits of noisy response also apply to other *parameterized* noise distributions.



(a) $\mathcal{D}(\theta)$ of NR vs. SM (b) $\mathcal{D}(\theta)$ of NR for different σ (c) $\mathcal{D}(\theta)$ of NR for different θ

Figure 2. Probability density of the aggregate response $\mathcal{D}(\theta)$ in a 1d-setting, where the base distribution $\mathcal{D}_X$ is a Gaussian with $x \sim \mathcal{N}(0, 0.5)$. We illustrate (a) $\mathcal{D}(\theta)$ for a population of agents that follow noisy response (NR) compared to standard microfoundations (SM), (b) how $\mathcal{D}(\theta)$ of NR changes for different θ, (c) variations in $\mathcal{D}(\theta)$ of NR for different values of σ.

4.3. Trade-off between imperfection and social burden

Apart from satisfying desirable aggregate-level properties, noisy response also satisfies the expenditure monotonicity requirement in Property 1. By Corollary 4, this implies an upper bound on the induced social burden at optimality for Setup 1. In the following we present a stronger result and show that in certain cases the social burden of noisy response is *strictly* lower.

Corollary 8. *Consider Setup 1. Let M_{SM} be the mapping associated with standard microfoundations, let $\sigma \in (0, \infty)$, and let the cost function be of the form $c(x_1, x_2) = |x_1 - x_2|$. Suppose that $[\theta_{SL}, \theta_{SL} + 1] \in \Theta \cap X$, where θ_{SL} is defined so that $\mu(\theta_{SL}) = 0.5$. Then, it holds that:*

$$\mathsf{Burden}(\theta_{\mathrm{PO}}(M_\sigma)) < \mathsf{Burden}(\theta_{\mathrm{PO}}(M_{\mathrm{SM}})).$$

In fact, the social burden for fuzzy perception can be well below the social burden of standard microfoundations. To demonstrate this, we visualize the social burden across a variety of different parameters of σ and p in Figure 3. The dashed lines indicate the reference values for standard microfoundations (SM) and for a population of non-strategic agents (NS). We observe that the social burden decreases with the fraction p of non-strategic agents in the population. Furthermore, if every agent follows noisy response ($p = 0$), the social burden is decreasing in σ. Thus, as the degree of imperfection in agents responses increases, the negative externalities of noisy response are increasingly smaller compared to those of the standard microfoundations.

Remark (Additional property of noisy response)**.** Since noisy response defines a *parameterized* model, the complex task of learning agent behavior is reduced to a parameter estimation problem for σ. This noise parameter can often be estimated via *individual experiments*, i.e., gathering information about individuals without deploying a classifier. (We refer to (Björkegren et al., 2020) for a related field experiment.) Estimating σ enables the decision-maker to estimate the distribution map, and thus estimate performative optima.

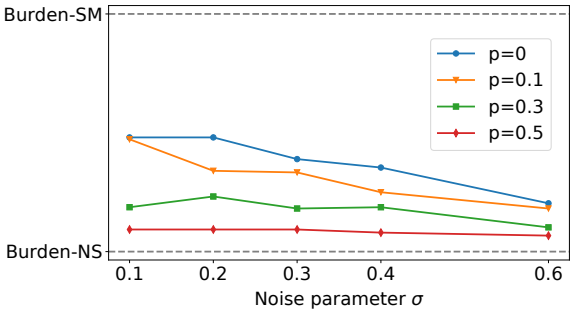


Figure 3. Social burden of optimal points in a 1d-setting for different values of σ and p (fraction of non-strategic agents). The population is sampled from a Gaussian mixture as in Figure 1.

5. Discussion

Traditional approaches for decision making in strategic environments are either individual-level like *strategic classification*, or population-level like *performative prediction*. In this work, we combine these two perspectives. We take advantage of microfoundations to endow the distribution map with structure, but we also keep in mind the aggregate-level properties they imply. Taking this holistic view enabled us to identify degeneracies with standard microfoundations in the context of binary classification. Furthermore, it inspired noisy response as a promising alternative microfoundations. While we have focused on strategic classification in this work, we believe that synthesizing the individual-level and aggregate-level perspectives can lead to interesting insights for the intersection of economics and learning more broadly.

Acknowledgments

We thank Jacob Steinhardt and Tijana Zrnic for feedback. We acknowledge support from the Paul and Daisy Soros Fellowship, Swiss National Science Foundation Postdoc.Mobility Fellowship, and NSF Award 1750555.

References

Emrah Akyol, Cedric Langbort, and Tamer Basar. Price of transparency in strategic machine learning. *Arxiv:1610.08210*, 2016.

Michael Anderson and Jeremy Magruder. Learning from the crowd: Regression discontinuity estimates of the effects of an online review database. *The Economic Journal*, 122 (563):957–989, 2012.

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

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

Daniel Björkegren, Joshua E. Blumenstock, and Samsun Knight. Manipulation-proof machine learning. *Arxiv:2004.03865*, 2020.

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

Michael Brückner and Tobias Scheffer. Stackelberg games for adversarial prediction problems. In *Proc.* 17*th ACM KDD*, page 547–555, 2011.

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

Colin F. Camerer, George Loewenstein, and Matthew Rabin. *Advances in Behavioral Economics*. Princeton University Press, 2004.

Yiling Chen, Yang Liu, and Chara Podimata. Learning strategy-aware linear classifiers. In *Proc.* 33*rd NeurIPS*, volume 33, pages 15265–15276, 2020.

Olivier Coibion, Yuriy Gorodnichenko, and Rupal Kamdar. The formation of expectations, inflation, and the phillips curve. *Journal of Economic Literature*, 56(4):1447–1491, 2018.

Nilesh N. Dalvi, Pedro M. Domingos, Mausam, Sumit K. Sanghai, and Deepak Verma. Adversarial classification. In *Proc.* 10*th KDD*, page 99–108, 2004.

Thomas S. Dee, Will Dobbie, Brian A. Jacob, and Jonah Rockoff. The causes and consequences of test score manipulation: Evidence from the new york regents examinations. *American Economic Journal: Applied Economics*, 11(3):382–423, July 2019. doi: 10.1257/app.20170520.

Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. Strategic classification from revealed preferences. In *Proc. EC*, page 55–70, 2018.

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.

Dimitris Gatzouras. On images of borel measures under borel mappings. *Proceedings of the American Mathematical Society*, 130(9):2687–2699, 2002.

Ganesh Ghalme, Vineet Nair, Itay Eilat, Inbal Talgam-Cohen, and Nir Rosenfeld. Strategic classification in the dark. *Arxiv:2102.11592*, 2021.

Nika Haghtalab, Nicole Immorlica, Brendan Lucier, and Jack Z. Wang. Maximizing welfare with incentive-aware evaluation mechanisms. In *Proc.* 29*th IJCAI*, pages 160–166, 2020.

Moritz Hardt, Nimrod Megiddo, Christos H. Papadimitriou, and Mary Wootters. Strategic classification. In *Proc.* 7*th ITCS*, page 111–122. ACM, 2016a.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Proc.* 29*th NeurIPS*, pages 3315–3323, 2016b.

John C. Harsanyi. Games with incomplete information played by "bayesian" players, i-iii. part ii. bayesian equilibrium points. *Management Science*, 14(5):320–334, 1968.

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*, page 259–268, 2019.

Greg Kaplan and Giovanni L. Violante. Microeconomic Heterogeneity and Macroeconomic Shocks. *Journal of Economic Perspectives*, 32(3):167–194, 2018.

Moein Khajehnejad, Behzad Tabibian, Bernhard Schölkopf, Adish Singla, and Manuel Gomez-Rodriguez. Optimal decision making under strategic behavior. *Arxiv:1905.09239*.

Jon Kleinberg and Manish Raghavan. How do classifiers induce agents to invest effort strategically? In *Proc. EC*, page 825–844, 2019.

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

Shengwu Li. Obviously strategy-proof mechanisms. *American Economic Review*, 107(11):3257–3287, 2017.

Robert E Lucas Jr. Econometric policy evaluation: A critique. In *Carnegie-Rochester conference series on public policy*, volume 1, pages 19–46. North-Holland, 1976.

Celestine Mendler-Dünner, Juan Perdomo, Tijana Zrnic, and Moritz Hardt. Stochastic optimization for performative prediction. *Proc.* 33*rd NeurIPS*, 33:4929–4939, 2020.

John Miller, Smitha Milli, and Moritz Hardt. Strategic classification is causal modeling in disguise. In *Proc.* 37*th ICML*, volume 119, pages 6917–6926, 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*, page 230–239, 2019.

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

Yonadav Shavit, Benjamin L. Edelman, and Brian Axelrod. Causal strategic linear regression. In *Proc.* 37*th ICML*, volume 119, pages 8676–8686, 2020.

Michael Spence. Job market signaling. *The Quarterly Journal of Economics*, 87(3):355–374, 1973.

Daniel A. Spielman and Shang-Hua Teng. Smoothed analysis: An attempt to explain the behavior of algorithms in practice. *Commun. ACM*, 52(10):76–84, October 2009.

Joseph E Stiglitz. Where modern macroeconomics went wrong. *Oxford Review of Economic Policy*, 34(1-2):70–106, 01 2018.

Stratis Tsirtsis and Manuel Gomez-Rodriguez. Decisions, counterfactual explanations and strategic behavior. *Arxiv:2002.04333*, 2020.

Hanrui Zhang and Vincent Conitzer. Incentive-aware PAC learning. *Proc.* 35*th AAAI*, 2021.

Hanrui Zhang, Yu Chen, and Vincent Conitzer. Automated mechanism design for classification with partial verification. *Proc.* 35*th AAAI*, 2021.