Learning Losses for Strategic Classification

Tosca Lechner,[1] **Ruth Urner,** [2]

[1] University of Waterloo
[2] York University
tlechner@uwaterloo.ca, ruth@eecs.yorku.ca

Abstract

Strategic classification, i.e. classification under possible strategic manipulations of features, has received a lot of attention from both the machine learning and the game theory community. Most works focus on analysing properties of the optimal decision rule under such manipulations. In our work we take a learning theoretic perspective, focusing on the sample complexity needed to learn a good decision rule which is robust to strategic manipulation. We perform this analysis by introducing a novel loss function, the strategic manipulation loss, which takes into account both the accuracy of the final decision and the vulnerability to manipulation. We analyse the sample complexity for a known graph of possible manipulations in terms of the complexity of the function class and the manipulation graph. Additionally, we address the problem of unknown manipulation capabilities of the involved agents. Using techniques from transfer learning theory, we define a similarity measure for manipulation graphs and show that learning outcomes are robust with respect to small changes in the manipulation graph. Lastly we analyse the (sample complexity of) learning of the manipulation capability of agents with respect to this similarity measure, providing a way to learn strategic classification with respect to an unknown manipulation graph.

1 Introduction

In many scenarios where a decision rule is learned from data, the publication of this decision rule has an effect on the distribution of the underlying population that may harm the quality of the rule. For example, applicants for a loan may change details in their bank account to receive a better score, people may join a gym or sports club without ever intending to participate, in order to get a better health insurance policy, or students may employ different strategies such as registering to volunteer, or joining rare clubs (without attending either) to appear better on college applications.

Effects and incentives resulting from strategic behavior in classification scenarios have received substantial attention from both machine learning and game theoretic perspectives in recent years (Hardt et al. 2016; Milli et al. 2019; Haghtalab et al. 2020; Tsirtsis and Rodriguez 2020; Zhang and Conitzer 2021). Most works study this as a two player game between an institution that publishes a decision rule and a population of best responding agents to be classified. Given the classifier, these agents may change their feature representations in order to obtain a more favorable classification outcome. To prevent the induced additional classification error the institution will publish a modified predictor, not transparently reflecting the underlying intent and potentially causing additional harm to sub-populations that may be less equipped to perform the required changes to their representations (Hu, Immorlica, and Vaughan 2019).

In this work, we propose a learning theoretic take on this scenario. In machine learning, it is common to model desiderata for a learning outcome in form of a loss function. The goal of the learning process is then to identify a predictor that minimizes this loss in expectation over a data-generating distribution. Thus, we here define a novel loss function for learning under strategic manipulations. The aim of this loss is to induce a combination of two (potentially competing) requirements: achieving low classification error taking into account that individuals being classified may manipulate their features, and discouraging such feature manipulations overall. Prior work has shown that these may be competing requirements (Zhang and Conitzer 2021), and our proposed loss function thus aims to induce a balanced combination of these requirements rather than strictly enforcing one and only observing the effect on the other (as is implied by frameworks that aim to minimize classification error under best-responding agents (Hardt et al. 2016; Milli et al. 2019) or enforcing incentive compatibility (Zhang and Conitzer 2021)).

To define our *strategic manipulation loss* we employ an abstraction of the plausible feature manipulations in form of a *manipulation graph* (Zhang and Conitzer 2021). An edge $\mathbf{x} \to \mathbf{x}'$ in this graph indicates that an individual with feature vector $\mathbf{x}$ may change their features to present as $\mathbf{x}'$ if this leads to a positive classification, for example since the utility of this change in classification exceeds the cost of the change between these vectors. We define our strategic loss in dependence of this graph and carefully motivate the proposed loss in terms of requirements and effects from previous literature. We then analyze the sample complexity of learning with this loss function. We identify sufficient conditions for proper learnability that take into account the interplay between a hypothesis class and an underlying ma-

nipulation graph. Moreover, we show that every class that has finite VC-dimension is learnable with respect to this loss by drawing a connection to results in the context of learning under adversarial perturbations (Montasser, Hanneke, and Srebro 2019). This effect may be surprising, since it presents a contrast to learning VC-classes with the sole requirement of minimizing classification error under strategic feature manipulations, which has been shown can lead to some VC-classes not being learnable (Zhang and Conitzer 2021). Thus, our analysis shows that balancing classification error with disincentivizing feature manipulations can reduce the complexity of the learning problem.

Moreover, we show that the quality of learning outcomes under our loss function is robust to inaccuracies in the manipulation graph. Such a robustness property is important, since an assumed graph might not exactly reflect agent's responses. In fact, it has recently been argued that the model of best-responding agents is not backed up by empirical observations on agent distributions after strategic responses (Jagadeesan, Mendler-Dünner, and Hardt 2021). Moreover, different sub-populations may have differences in their manipulation graphs (different capabilities to manipulate their features) or a manipulation graph may be inferred from data and therefore exhibit statistical errors. We introduce a novel distance measure between manipulation graphs by drawing connections to learning bounds in transfer learning (Ben-David et al. 2010; Mansour, Mohri, and Rostamizadeh 2009) and show that the strategic loss of a learned predictor when employing a different manipulation graph can be bounded in terms of this distance measure. Finally, we present some initial results on how manipulation graphs may be learned from data.

1.1 Related work

That learning outcomes might be compromised by agents responding to a published classification rules with strategic manipulations of their feature vectors was first pointed out over a decade ago (Dalvi et al. 2004; Brückner and Scheffer 2011) and has received substantial interest from the research community in recent years initiated by a study by Hardt et al. that differentiated the field from the more general context of learning under adversarial perturbations (Hardt et al. 2016). That study considered strategic responses being induced by separable cost functions for utility maximizing agents and studied the resulting decision boundaries for certain classes of classifiers. Recent years have seen a lot of interest in better understanding the interplay of various interests in settings where a decision rule is published and thereby has an effect on how the entities that are to be classified might present themselves to the decision make. In particular, various externalities to this scenario have been analyzed. A general cost to society formalized in form of "social burden" incurred by the costs of enforced feature manipulation has been shown to occur when institutions anticipate strategic responses (Milli et al. 2019; Jagadeesan, Mendler-Dünner, and Hardt 2021). Further, it has been demonstrated how such burden may be suffered to differing degrees by various subgroups of a population that may differ in their capabilities to adapt their features in ways that are favorable to them (Milli et al. 2019; Hu, Immorlica, and Vaughan 2019), raising concerns over fairness in such scenarios.

Recent studies have extended the original game theoretic model of a classifier publishing intuition and best responding subjects. For example, a recent work studied how strategic modification may be a positive effect and how that should be taken into consideration by the institution (Haghtalab et al. 2020). Such a perspective has been connected to underlying causal relations between features and classification outcome and resulting strategic recommendations (Miller, Milli, and Hardt 2020; Tsirtsis and Rodriguez 2020). Further, a very recent study has explored how the model of a best responding agent may be relaxed to better reflect empirically observed phenomena (Jagadeesan, Mendler-Dünner, and Hardt 2021).

Much of previous work considers the scenario of classification with strategic agents on a population level. A few recent studies have also analyzed how phenomena observed on samples reflect the underlying population events (Haghtalab et al. 2020). Notably, very recent studies provided a first analyses of learning with strategically responding agents in a PAC framework (Zhang and Conitzer 2021; Sundaram et al. 2021). The former work studied the sample complexity of learning VC-classes in this setup and analyzed effects on sample complexity of enforcing incentive compatibility for the learned classification rules. Our work can be viewed as an extension of this analysis. We propose to combine aspects of incentive compatibility and minimizing negative externalities such as social burden in form of a novel loss function that may serve as a learning objective when strategic responses are to be expected.

Our sample complexity analysis is then hinging on techniques developed in the context of learning under adversarial perturbations, a learning scenario which has received considerable research attention in recent years (Feige, Mansour, and Schapire 2015; Cullina, Bhagoji, and Mittal 2018; Montasser, Hanneke, and Srebro 2019, 2021). While the learning problems are not identical, we present how strategic behaviour can be modeled as a form of "one-sided adversarial perturbation" and inheritance of resulting learning guarantees.

1.2 Overview on contributions

In Section 2 we review our notation and then introduce our new notion of strategic loss and motivate it. Our main contributions can be summarized as follows:

Strategic manipulation loss We propose a novel loss function for learning in the presence of strategic feature manipulations. We carefully motivate this loss by relating it to concepts of social burden and incentive compatibility (and their potential trade-offs with accuracy) in prior literature.

Sample complexity analysis We analyze (PAC type) learnability of VC-classes with the strategic loss. We provide sufficient conditions (and examples of when they are satisfied) for learnability with a proper learner. By drawing connections and adapting results from learning under adversarial perturbations to our setup, we also show that,

while proper learnability can not always be guaranteed, every VC-class is learnable under the strategic loss with an improper learner.

Robustness to inaccurate manipulation information We investigate the impact of using an approximate manipulation graph to yield a surrogate strategic loss function in cases where the true manipulation graph is not accessible. For this, we introduce a novel similarity measure on graphs and show that if graphs are similar with respect to our notion then they yield reasonable surrogate losses for each other (Theorem 6).

Learning the manipulation graph We explore the question whether it is possible to learn a manipulation graph that yields a good surrogate strategic loss. We identify a sufficient condition for a class of graphs $\mathcal{G}$ being learnable with respect to our previously defined similarity measure for graphs (Theorem 7), which in turn guaranteed the learning of a reasonable surrogate loss.

All proofs can be found in the appendix of the full version.

2 Setup

2.1 Basic Learning Theoretic Notions for Classification

We employ a standard setup of statistical learning theory for classification. We let $\mathcal{X} \subseteq \mathbb{R}^d$ denote the domain and $\mathcal{Y}$ (mostly $\mathcal{Y} = \{0, 1\}$) a (binary) label space. We model the data generating process as a distribution P over $\mathcal{X} \times \mathcal{Y}$ and let $P_{\mathcal{X}}$ denote the marginal of P over $\mathcal{X}$. We use the notation $(\mathbf{x}, y) \sim P$ to indicate that $(\mathbf{x}, y)$ is a sample from distribution P and $S \sim P^n$ to indicate that set S is a sequence (for example a training or test data set) of n i.i.d. samples from P. Further, we use notation $\eta_P(\mathbf{x}) = \mathbb{P}_{(\mathbf{x},y)\sim P}[y = 1 \mid \mathbf{x}]$ to denote the *regression* or *conditional labeling function* of P. We say that the distribution has *deterministic labels* if $\eta_P(\mathbf{x}) \in \{0, 1\}$ for all $\mathbf{x} \in \mathcal{X}$.

A *classifier* or *hypothesis* is a function $h : \mathcal{X} \to \mathcal{Y}$. A classifier h can naturally be viewed a subset of $\mathcal{X} \times \mathcal{Y}$, namely $h = \{(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y} \mid \mathbf{x} \in \mathcal{X},\ y = h(\mathbf{x})\}$. We let $\mathcal{F}$ denote the set of all Borel measurable functions from $\mathcal{X}$ to $\mathcal{Y}$ (or all functions in case of a countable domain). A *hypothesis class* is a subset of $\mathcal{F}$, often denoted by $\mathcal{H} \subseteq \mathcal{F}$. For a loss function $\ell : \mathcal{F} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ we denote the expected loss for a distribution P as $\mathcal{L}_P$ and the empirical loss for a sample S as $\mathcal{L}_S$. We use standard definitions like PAC learnability, sample complexity and approximation error. For further elaborations on these definitions we refer the reader to the appendix for an extended definitions section or to (Shalev-Shwartz and Ben-David 2014).

2.2 Strategic Classification

Learning objectives in prior work The possibilities for strategic manipulations of a feature vector are often modeled in terms of a cost function $\mathrm{cost} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_0^+$, so that $\mathrm{cost}(\mathbf{x}, \mathbf{x}')$ indicates how expensive it is for an individual with feature vector $\mathbf{x}$ to present as $\mathbf{x}'$. A natural minimal assumption a cost function should satisfy is $\mathrm{cost}(\mathbf{x}, \mathbf{x}) = 0$ for all feature vectors $\mathbf{x}$. It is then typically assumed that instances best-respond to a published classifier, in that the individual $\mathbf{x}$ would choose to pay the cost of presenting as $\mathbf{x}'$ as long as the cost doesn't exceed the utility that would be gained from the difference in classification outcome. Assuming the benefit of individual $\mathbf{x}$ receiving classification 1 over classification 0 is γ, the manipulation would happen if $\mathrm{cost}(\mathbf{x}, \mathbf{x}') \leq \gamma$ and $h(\mathbf{x}) = 0$ while $h(\mathbf{x}') = 1$ for a given classifier. That is, we can define the best response of an individual with feature vector $\mathbf{x}$ facing classifier h as

$$\mathrm{br}(\mathbf{x}, h) = \mathrm{argmax}_{\mathbf{x}' \in \mathcal{X}}[\gamma \cdot h(\mathbf{x}') - \mathrm{cost}(\mathbf{x}, \mathbf{x}')],$$

with ties broken arbitrarily, and assuming that, if the original feature vector $\mathbf{x}$ is among those maximizing the above, then the individual would choose to maintain the original features. An often assumed learning goal is then *performative optimality* (Perdomo et al. 2020; Jagadeesan, Mendler-Dünner, and Hardt 2021), which stipulates that a learner should aim to maximize accuracy on the distribution it induces via the agent responses. That is, this objective can be phrased as minimizing

$$\mathbb{E}_{(\mathbf{x},y)\sim P} \mathbb{1}\left[h(\mathrm{br}(\mathbf{x}, h)) \neq y\right]$$

An alternative view on this setup, if the agent responses are deterministic, is to view the above as minimizing the binary loss of the *effective hypothesis* $\hat{h} : \mathcal{X} \to \{0, 1\}$ that is induced by h and the agents' best responses $\mathrm{br}(\cdot, \cdot)$ (Zhang and Conitzer 2021), defined as

$$\hat{h}(\mathbf{x}) = h(\mathrm{br}(\mathbf{x}, h)). \tag{1}$$

The goal of performative optimality has been combined with the notion of *social burden* that is induced by a classifier (Milli et al. 2019; **?**). This notion reflects that it is undesirable for a (truly) positive instance to be forced to manipulate its features to obtain a (rightfully) positive classification. This is modeled by considering the *burden* on a positive individual to be the cost that is incurred by reaching for a positive classification and the *social burden* incurred by a classifier to be the expectation with respect to the data-generating process over these costs:

$$\mathrm{brd}_P(h) = \mathbb{E}_{(\mathbf{x},y)\sim P}\left[\min_{\mathbf{x}' \in \mathcal{X}}\{\mathrm{cost}(\mathbf{x}, \mathbf{x}') \mid h(\mathbf{x}') = 1\} \mid y = 1\right]$$

It has been shown that optimizing for performative optimality (under the assumption of deterministic best-responses) also incurs maximal social burden (Jagadeesan, Mendler-Dünner, and Hardt 2021).

A new loss function for learning under strategic manipulations Arguably, to seek performative optimality (or minimize the binary loss over the effective hypothesis class) the cost function as well as the value γ (or function $\gamma : \mathcal{X} \to \mathbb{R}$) of positive classification needs to be known (or at least approximately known). To take best responses into account, a learner needs to know what these best responses may look like. In that case, we may ignore the details of the cost function and γ, and simply represent the collection of *plausible manipulations* as a directed graph structure $\mathcal{M} = (\mathcal{X}, E)$ over the feature space $\mathcal{X}$ (Zhang and Conitzer 2021). The

edge-set E consists of all pairs $(\mathbf{x}, \mathbf{x}')$ with $\text{cost}(\mathbf{x}, \mathbf{x}') \leq \gamma$, and we will also use the notation $\mathbf{x} \to \mathbf{x}'$ for $(\mathbf{x}, \mathbf{x}') \in E$, and write $\mathcal{M} = (\mathcal{X}, E) = (\mathcal{X}, \to)$. We note that this formalism is valid for both countable (discrete) and uncountable domains.

Given the information in the so obtained *manipulation graph* $\mathcal{M} = (\mathcal{X}, \to)$, we now design a loss function for classification in the presence of strategic manipulation that reflects both classification errors and the goal of disincentivizing manipulated features as much as possible. Our proposed loss function below models that, given that feature vector $\mathbf{x}$ can present as $\mathbf{x}'$, it is undesirable for a classifier to assign $h(\mathbf{x}) = 0$ and $h(\mathbf{x}') = 1$. This is independent of a true label y (e.g. if $(\mathbf{x}, y)$ is sampled from the data generating process). If the label $y = 0$ is not positive, the point gets misclassified when $\mathbf{x}$ presents as $\mathbf{x}'$. On the other hand, if the true label is 1, then either a true positive instance is forced to manipulate their features to obtain a rightly positive outcome (and this contributes to social burden), or, if the choice is to not manipulate the features, the instance will be misclassified (prior work has also considered models where true positive instance are "honest" and will not manipulate their features (Dong et al. 2018)). Here, we propose to incorporate both misclassification and contributions to social burden into a single loss function that a learner may aim to minimize.

Definition 1. *We define the* strategic loss $\ell^{\to} : \mathcal{F} \times \mathcal{X} \times \mathcal{Y} \to \{0, 1\}$ *as follows:*

$$\ell^{\to}(h, \mathbf{x}, y) = \begin{cases} 1 & \text{if } h(\mathbf{x}) \neq y \\ 1 & \text{if } h(\mathbf{x}) = 0 \\ & \quad \text{and } \exists \mathbf{x}' \text{ with } \mathbf{x} \to \mathbf{x}' \text{ and } h(\mathbf{x}') = 1 \\ 0 & \text{else} \end{cases}$$

Note that the first two cases are not mutually exclusive. The above loss function discretizes the social burden by assigning a loss 1 whenever a positive individual is required to manipulate features. As for the standard classification loss, the above point-wise definition of a loss function allows to define the *true strategic loss* $\mathcal{L}^{\to}_P(h)$ and *empirical strategic loss* $\mathcal{L}^{\to}_S(h)$ of a classifier with respect to a distribution P or a data sample S.

2.3 Comparison with Alternative Formalisms for Strategic Classification

To motivate our proposed loss, we here discuss several scenarios where, we'd argue, minimizing the strategic loss leads to a more desirable learning outcome than learning with a binary loss, while taking strategic manipulations into account. As discussed above, a common approach to modeling classification in a setting where strategic manipulations may occur is to assume that all agents will best-respond to a published classifier. That is, if $h(\mathbf{x}) = 0$, $h(\mathbf{x}') = 1$ and $\mathbf{x} \to \mathbf{x}'$, then the agent with initial feature vector $\mathbf{x}$ will *effectively* receive classification 1. A natural modeling is then to consider the effective hypothesis $\hat{h}$ induced h (see Equation 1) and aim to minimize the classification error with the *effective class* $\hat{\mathcal{H}} = \{f \mid f = \hat{h} \text{ for some } h \in \mathcal{H}\}$ (Zhang and Conitzer 2021). However it has been shown, that the VC-dimension of $\hat{\mathcal{H}}$ may be arbitrarily larger than the VC-dimension of $\mathcal{H}$, and may even become infinite (Zhang and Conitzer 2021). When learning this effective class $\hat{\mathcal{H}}$ with respect to the binary loss (which corresponds to aiming for performative optimality), this will imply that the class is not learnable. By contrast, we will show below that any class of finite VC-dimension remains learnable with respect to the strategic loss.

It has also been shown that the negative effects in terms of sample complexity of considering the effective hypothesis class can be avoided by considering only *incentive compatible* hypotheses in $\mathcal{H}$, that is outputting only such hypotheses that will not induce any feature manipulations in response to the published classifier (Zhang and Conitzer 2021). While this avoids the growths in terms of VC-dimension, it may prohibitively increase the approximation error of the resulting (pruned) class as we show in the example below. We would argue that this illustrates that low sample complexity, in itself, is not a sufficient criterion for learning success.

Example 1. *Consider $\mathcal{X} = \mathbb{N}$ and a manipulation graph that includes edges $n \to n+1$ and $n \to n-1$ for all $n \in \mathbb{N}$. This is a reasonable structure, considering that the cost of moving the (one-dimensional) feature by 1 is worth a positive classification outcome. However, the only two hypotheses that are incentive compatible in this case are the two constant functions $h_0 : \mathcal{X} \to \{0\}$ and $h_1 : \mathcal{X} \to \{1\}$. Thus, requiring incentive compatiblity forces the learner to assign all points in the space with the same label. This class, in fact, has low sample complexity. However, arguably, restricting the learning to such a degree (and suffering the resulting classification error, which will be close to 0.5 for distributions with balanced classes), is, in most cases not a reasonable price to pay for dis-incentivising feature manipulations.*

The following example illustrates how our loss function can be viewed as incorporating the notion of social burden directly into the loss.

Example 2. *Let's again consider a domain $\mathcal{X} = \mathbb{N}$ and a manipulation graph $\mathcal{M}$ with edges $n \to n+1$ for all $n \in \mathbb{N}$. We consider distributions that have support $\{(1,0), (2,0), (3,1), (4,1)\}$, thus only these four points have positive probability mass and a hypothesis class of thresholds $\mathcal{H} = \{h_a \mid a \in \mathbb{R}\}$, with $h_a(\mathbf{x}) = \mathbb{1}[\mathbf{x} \geq a]$. The true labeling on these distributions is $\eta(\mathbf{x}) = h_{2.5}(\mathbf{x})$. On all distributions, where all four points have positive mass the performatively optimal hypothesis (or effective hypothesis of minimal binary loss) however is $h_{3.5}$. The social burden incurred then is $\text{brd}_P(h_{3.5}) = P((3,1)) \cdot \text{cost}(3,4)$. It is important to note that the performative optimality of $h_{3.5}$ is independent of the distribution P over the points. A learner that minimizes the strategic loss, on the other hand, will take the distribution P into account and output $h_{2.5}$ if $P((2,0)) < P((3,1))$, while outputting $h_{3.5}$ if $P((2,0)) > P((3,1))$. If the difference in mass of these points (or the margin areas in a more general setting) is significant, then minimizing the strategic loss will opt for allowing a small amount of manipulation in turn for outputting a correct classification rule in case $P((2,0)) \ll P((3,1))$;*

and it will opt for changing the classification rule, accept a small amount of social burden in exchange for preventing a large amount of manipulations and resulting classification errors, in case $P((2,0)) \gg P((3,1))$. We would argue that this reflects a desirable learning outcome.

3 Learnability with the Strategic Loss

3.1 Warm up: Loss Classes and Learnability

It is well known that a class $\mathcal{H}$ is learnable (with respect to the set of all distributions) if the *loss class* induced by a $0/1$-valued loss function ℓ has finite VC-dimension. In the case of the classification loss, this is in fact a characterization for learnability (and the VC-dimension of the loss class is identical to the VC-dimension of the hypothesis class $\mathcal{H}$). In general, bounded VC-dimension of the loss class is a sufficient condition for learnability (the VC-dimension provides an upper bound on the sample complexity), but it is not a necessary condition (it doesn't, in general, yield a lower bound on the sample complexity of learning a class $\mathcal{H}$ with respect to some loss ℓ). We start by reviewing these notions for the classification loss and then take a closer look at the loss class induced by the strategic loss.

Let ℓ be a loss function and h be a classifier. We define the *loss set* $h_\ell \subseteq \mathcal{X} \times \mathcal{Y}$ as the set of all labeled instances $(\mathbf{x}, y)$ on which h suffers loss 1. The *loss class* $\mathcal{H}_\ell$ is the collection of all loss sets (in the literature, the loss class is often described as the function class of indicator functions over these sets). In the case of binary classification loss $\ell^{0/1}$, the loss set of a classifier h is exactly the complement of h in $\mathcal{X} \times \mathcal{Y}$. That is, in this case the loss set of h is also a binary function over the domain $\mathcal{X}$ (namely the function $\mathbf{x} \mapsto |h(\mathbf{x}) - 1|$). For the strategic loss on the other hand, the loss set of a classifier h is not a function, since it can contain both $(\mathbf{x}, 0)$ and $(\mathbf{x}, 1)$ for some points $\mathbf{x} \in \mathcal{X}$, namely if $h(\mathbf{x}) = 0$ and there exists an $\mathbf{x}'$ with $\mathbf{x} \to \mathbf{x}'$ and $h(\mathbf{x}') = 1$. For a class $\mathcal{H}$ we let $\mathcal{H}_{\ell^{0/1}}$ denote the loss class with respect to the binary loss and $\mathcal{H}_{\ell^\to}$ the loss class with respect to the strategic loss.

Definition 2. *Let $\mathcal{Z}$ be some set and $\mathcal{U} \subseteq 2^{\mathcal{Z}}$ be a collection of subsets of $\mathcal{Z}$. We say that a set $S \subseteq \mathcal{Z}$ is* shattered by *$\mathcal{U}$ if*

$$\{U \cap S \mid U \in \mathcal{U}\} = 2^S,$$

that is, every subset of S can be obtained by intersecting S with some set U from the collection $\mathcal{U}$. The VC-dimension *of $\mathcal{U}$ is the largest size of a set that is shattered by $\mathcal{U}$ (or ∞ if $\mathcal{U}$ can shatter arbitrarily large sets).*

It is easy to verify that for the binary loss, the VC-dimension of $\mathcal{H}$ as a collection of subsets of $\mathcal{X} \times \mathcal{Y}$ is identical with the VC-dimension of $\mathcal{H}_{\ell^{0/1}}$ (and this also coincides with the VC-dimension of $\mathcal{H}$ as a binary function class (Shalev-Shwartz and Ben-David 2014); VC-dimension is often defined for binary functions rather than for collection of subsets, however this is limiting for cases where the loss class is not a class of functions).

We now show that the VC-dimension of a class $\mathcal{H}$ and its loss class with respect to the strategic loss can have an arbitrarily large difference. Similar results have been shown for the binary loss class of the effective class $\hat{\mathcal{H}}$ induced by a manipulation graph (Zhang and Conitzer 2021). However the binary loss class of $\hat{\mathcal{H}}$ is different from the strategic loss class of $\mathcal{H}$ and, as we will see, the implications for learnability are also different.

Observation 1. *For any $d \in \mathbb{N} \cup \{\infty\}$ there exists a class $\mathcal{H}$ and a manipulation graph $\mathcal{M} = (\mathcal{X}, \to)$ with $\mathrm{VC}(\mathcal{H}) = 1$ and $\mathrm{VC}(\mathcal{H}_{\ell^\to}) \geq d$.*

On the other hand, we prove that the VC-dimension of the strategic loss class $\mathcal{H}_{\ell^\to}$ is always at least as large as the VC-dimension of the original class.

Observation 2. *For any hypothesis class $\mathcal{H}$ and any manipulation graph $\mathcal{M} = (\mathcal{X}, \to)$, we have $\mathrm{VC}(\mathcal{H}) \leq \mathrm{VC}(\mathcal{H}_{\ell^\to})$.*

Standard VC-theory tells us that, for the binary classification loss, any learner that acts according to the ERM (Empirical Risk Minimization) principle is a successful learner for classes of bounded VC-dimension d. For a brief recap of the underpinnings of this result we refer the reader to the supplementary material or for further details to (Shalev-Shwartz and Ben-David 2014). In the case of general loss classes with values in $\{0, 1\}$, the VC-dimension does not characterize learnability. In particular, we next show that the VC-dimension of the strategic loss class does not imply a lower bound on the sample complexity.

Theorem 3. *For every $d \in \mathbb{N} \cup \{\infty\}$, there exists a hypothesis class $\mathcal{H}$ with $\mathrm{VC}(\mathcal{H}_{\ell^\to}) = d$ that is learnable with sample complexity $O(\log(1/\delta)/\epsilon)$ in the realizable case.*

3.2 Sufficient Conditions for Strategic Loss Learnability

In the previous section, we have seen that the loss class having a finite VC-dimension is a sufficient (but not necessary) condition for learnability with respect to the strategic loss. We have also seen that the VC-dimension of $\mathcal{H}_{\ell^\to}$ can be arbitrarily larger than the VC-dimension of $\mathcal{H}$. To start exploring what determines learnability under the strategic loss, we provide a sufficient condition for a class to be properly learnable with respect to the strategic loss.

Note that for a hypothesis h, the strategic loss set $h_{\ell^\to}$ can be decomposed into the loss set of h with respect to the binary loss and the component that comes from the strategic manipulations. Formally, we can define the *strategic component loss*.

Definition 3. *We let the strategic component loss with respect to manipulation graph $\to$ be defined as*

$$\ell^{\to,\perp}(h, \mathbf{x}) = \mathbb{1}\left[h(\mathbf{x}) = 0 \wedge \exists \mathbf{x}' : \mathbf{x} \to \mathbf{x}' : h(\mathbf{x}') = 1\right]$$

We note that $\ell^\to(h, x, y) \leq \ell^{0/1}(h, x, y) + \ell^{\to,\perp}(h, x)$. We will denote the true strategic component loss with respect to marginal distribution $P_\mathcal{X}$ as $\mathcal{L}^{\to,\perp}_{P_\mathcal{X}}$.

For the loss sets, we then get

$$h_{\ell^{0/1}} = \{(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y} \mid h(\mathbf{x}) \neq y\},$$

and

$$h_{\ell^{\to,\perp}} = \{(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y} \mid h(\mathbf{x}) = 0 \wedge \exists \mathbf{x}' : \mathbf{x} \to \mathbf{x}' \wedge h(\mathbf{x}') = 1\}.$$

This implies

$$h_{\ell^{\rightarrow}} = h_{\ell^{0/1}} \cup h_{\ell^{\rightarrow,\perp}}$$

for all classifiers $h \in \mathcal{F}$, and thereby

$$\mathcal{H}_{\ell^{\rightarrow}} = \{h_{\ell^{0/1}} \cup h_{\ell^{\rightarrow,\perp}} \mid h \in \mathcal{H}\}.$$

By standard counting arguments on the VC-dimension of such unions (see, for example Chapter 6 of (Shalev-Shwartz and Ben-David 2014) and exercises in that chapter), it can be shown this decomposition implies that $\mathrm{VC}(\mathcal{H}_{\ell^{\rightarrow}}) \leq d \log d$ for $d = \mathrm{VC}(\mathcal{H}_{\ell^{0/1}}) + \mathrm{VC}(\mathcal{H}_{\ell^{\rightarrow,\perp}}) = \mathrm{VC}(\mathcal{H}) + \mathrm{VC}(\mathcal{H}_{\ell^{\rightarrow,\perp}})$. Thus, if both the class $\mathcal{H}$ itself and the class of strategic components have finite VC-dimension, then $\mathcal{H}$ is properly learnable by any learner that is an ERM for the strategic loss:

Theorem 4. *Let $\mathcal{H}$ be a hypothesis class with finite $\mathrm{VC}(\mathcal{H}) + \mathrm{VC}(\mathcal{H}_{\ell^{\rightarrow,\perp}}) = d < \infty$. Then $\mathcal{H}$ is properly PAC learnable with respect to the strategic loss (both in the realizable and the agnostic case).*

Whether the class of strategic components has finite VC-dimension intrinsically depends on the interplay between the hypothesis class $\mathcal{H}$ and the graph structure of the manipulation graph. In Observation 1, we have seen that the graph structure can yield the strategic component sets to have much larger complexity than the original class. In the appendix, Section B, we provide a few natural examples, where the VC-dimension of the strategic components can be bounded.

Theorem 4 provides a strong sufficient condition under which any empirical risk minimizer for the strategic loss is a successful agnostic learner for a class of finite VC-dimension. We believe, in many natural situations the conditions in that theorem will hold, and analyzing in more detail which graph structure, combinations of graphs structures and hypothesis classes or classes of cost function lead to the strategic component sets having finite VC-dimension is an intriguing direction for further research.

We close this section with two results, both stated in Theorem 5, on the learnability under the strategic loss in the general case where the VC-dimension of the strategic component sets may be infinite. First, there are classes and manipulation graphs for which no proper learner is (PAC-) successful, even in the realizable case. Second, for any class of finite VC-dimension and any manipulation graph, there exists an improper PAC learner. These results follow by drawing a connection from learning under the strategic loss to learning under an adversarial loss (Montasser, Hanneke, and Srebro 2019). In the general adversarial loss setup, every domain instance $\mathbf{x}$ is assigned a set of potential perturbation $\mathcal{U}(\mathbf{x})$, and the adversarial loss of a hypothesis h is then defined as

$$\ell^{\mathcal{U}}(h, \mathbf{x}, y) = \mathbb{1}\left[\exists \mathbf{x}' \in \mathcal{U}(\mathbf{x}) : h(\mathbf{x}') \neq y\right].$$

The strategic loss can be viewed as a one-sided version of the adversarial loss, where the perturbation sets differ conditional on the label of a point, and where $\mathcal{U}(\mathbf{x}, 1) = \{\mathbf{x}\}$, while $\mathcal{U}(\mathbf{x}, 0) = \{\mathbf{x}' \in \mathcal{X} \mid \mathbf{x} \rightarrow \mathbf{x}'\}$. The following results on learnability with the strategic loss then follow by slight modifications of the corresponding proofs for learning under adversarial loss.

Theorem 5. (Adaptation of Theorem 1 and Theorem 4 in (Montasser, Hanneke, and Srebro 2019)) *There exists a hypothesis class $\mathcal{H}$ with $\mathrm{VC}(\mathcal{H}) = 1$ that is not learnable with respect to the strategic loss by any proper learner $\mathcal{A}$ for $\mathcal{H}$ even in the realizable case. On the other hand, every class $\mathcal{H}$ of finite VC-dimension is learnable (by some improper learner).*

4 Strategic loss with respect to an approximate manipulation graph

In many situations one might not have direct access to the true manipulation graph $\mathcal{M} = (V, E)$, but only to some approximate graph $\mathcal{M}' = (V, E')$. In this section we will investigate how this change of manipulation graph impacts the corresponding loss function. We define a criterion for measuring the similarity of graphs with respect to hypothesis class $\mathcal{H}$ and show that similar graphs will yield similar strategic losses. That is, we show an upper bound on the true strategic loss of a hypothesis h (i.e., strategic loss with respect to the true manipulation graph) in terms of the graph similarity and the surrogate strategic loss of h (i.e., the strategic loss with respect to the approximate graph). We will use $\mathbf{x} \rightsquigarrow \mathbf{x}'$ to denote $(\mathbf{x}, \mathbf{x}') \in E'$. As the set of vertices V is always equal to $\mathcal{X}$ in our setting, the graphs $\mathcal{M}$ and $\mathcal{M}'$ are uniquely defined by $\rightarrow$ and $\rightsquigarrow$ respectively. We will therefore use $\rightarrow$ and $\mathcal{M}$, as well as $\rightsquigarrow$ and $\mathcal{M}'$ interchangeably.

We now define the distance between graphs with respect to a hypothesis class $\mathcal{H}$ by the impact a change of manipulation graph has on the strategic component loss of elements of $\mathcal{H}$. This definition and its later use is inspired by works on domain adaptation (Ben-David et al. 2010; Mansour, Mohri, and Rostamizadeh 2009).

Definition 4. *For two manipulation graphs, given by $\rightarrow$ and $\rightsquigarrow$, we let their $\mathcal{H}$-$P_{\mathcal{X}}$-distance be defined as*

$$d_{\mathcal{H},P_{\mathcal{X}}}(\rightarrow, \rightsquigarrow) = \sup_{h \in \mathcal{H}} \mathbb{E}_{\mathbf{x} \sim P_{\mathcal{X}}}[|\ell^{\rightarrow,\perp}(h, \mathbf{x}) - \ell^{\rightsquigarrow,\perp}(h, \mathbf{x})|]$$

We will now bound the strategic manipulation loss $\mathcal{L}_P^{\rightarrow}(h)$ with respect to the true graph $\rightarrow$ in terms of the strategic manipulation loss $\mathcal{L}_P^{\rightsquigarrow}(h)$ with respect to the approximate graph $\rightsquigarrow$ and the $\mathcal{H}$-$P_{\mathcal{X}}$-distance between $\rightarrow$ and $\rightsquigarrow$.

Theorem 6. *Let $\mathcal{H}$ be any hypothesis class and $\rightarrow, \rightsquigarrow$ two manipulation graphs. Then for any distribution P over $\mathcal{X} \times \mathcal{Y}$ and any $h \in \mathcal{H}$ we have*

$$\begin{aligned}\mathcal{L}_P^{\rightarrow}(h) &\leq \mathcal{L}_P^{0/1}(h) + \mathcal{L}_P^{\rightsquigarrow,\perp}(h) + d_{\mathcal{H},P_{\mathcal{X}}}(\rightarrow, \rightsquigarrow) \\ &\leq 2\mathcal{L}_P^{\rightsquigarrow}(h) + d_{\mathcal{H},P_{\mathcal{X}}}(\rightarrow, \rightsquigarrow).\end{aligned}$$

Furthermore, by rearranging the result, we get

$$\frac{1}{2}\mathcal{L}_P^{\rightsquigarrow}(h) - d_{\mathcal{H},P_{\mathcal{X}}}(\rightarrow, \rightsquigarrow) \leq \mathcal{L}_P^{\rightarrow}(h).$$

We note that the expression $d_{\mathcal{H},P_{\mathcal{X}}}(\rightarrow, \rightsquigarrow)$ is independent of the labelling and can therefore be estimated using data without any label information. Furthermore we have seen that small $d_{\mathcal{H},P_{\mathcal{X}}}(\rightarrow, \rightsquigarrow)$ tightens the upper as well as the lower bound on $\mathcal{L}_P^{\rightarrow}(h)$. Therefore, $d_{\mathcal{H},P_{\mathcal{X}}}(\rightarrow, \rightsquigarrow)$ is a suitable distance measure for approximating the structure of the

manipulation graph. In the following subsection we will explore learning $\rightsquigarrow$ with low $d_{\mathcal{H},P_\mathcal{X}}(\rightarrow,\rightsquigarrow)$ from finite samples.

5 Learning a manipulation graph

In the last section we have assumed to be given an approximate manipulation graph which we can use to learn a classifier with low strategic loss. We now want to go one step further and pose the goal of learning a manipulation graph $\rightsquigarrow$ from a predefined class of graphs $\mathcal{G}$ such that $\ell^{\rightsquigarrow,\perp}$ serves as good strategic surrogate loss for $\ell^{\rightarrow,\perp}$. From Theorem 6 we already know that $\ell^{\rightsquigarrow,\perp}$ is a good surrogate loss for $\ell^{\rightarrow,\perp}$ if $d_{\mathcal{H},P_\mathcal{X}}(\rightarrow,\rightsquigarrow)$ is small. This section will thus focus on learning an approximate manipulation graph $\rightsquigarrow\in\mathcal{G}$ with small $d_{\mathcal{H},P_\mathcal{X}}(\rightarrow,\rightsquigarrow)$.

In order to further specify our learning problem, we will now describe what the input of such a learning procedure will look like. For a manipulation graph $\rightarrow$, let $B_\rightarrow : \mathcal{X} \rightarrow 2^\mathcal{X}$ be the function that maps an instance $\mathbf{x}$ to its set of children, i.e., by $B_\rightarrow(\mathbf{x}) = \{\mathbf{x}' \in \mathcal{X} : \mathbf{x} \rightarrow \mathbf{x}'\}$. We note that a manipulation graph $\rightarrow$ is uniquely defined by $B_\rightarrow$. Thus we will sometimes use $B_\rightarrow$ and $\rightarrow$ interchangeably. The input to our learning procedure will be of the form of samples $S = \{(\mathbf{x}_1, B_\rightarrow(\mathbf{x}_1)), \ldots, (\mathbf{x}_n, B_\rightarrow(\mathbf{x}_n))\}$ from the true manipulation graph $\rightarrow$.

As a next step in formulating our learning problem, we will need to define a loss function. As stated above, our goal is to learn $\rightsquigarrow\in\mathcal{G}$ with small $d_{\mathcal{H},P_\mathcal{X}}(\rightarrow,\rightsquigarrow)$. As the definition of $d_{\mathcal{H},P_\mathcal{X}}(\rightarrow,\rightsquigarrow)$ contains a supremum over all $h \in \mathcal{H}$, we cannot use it as a loss directly (as a loss needs to be defined point-wise). However, we can formulate a loss that is closely related and will serve to guarantee low $d_{\mathcal{H},P_\mathcal{X}}(\rightarrow,\rightsquigarrow)$. Let the *graph loss* for a manipulation graph $\rightsquigarrow$, a domain point x, a manipulation set $B \subset \mathcal{X}$ and a hypothesis h as:

$$\ell^{\mathrm{gr}}(h,\rightsquigarrow,\mathbf{x},B) = \begin{cases} 1 & \text{if } h(\mathbf{x}) = 0 \wedge \exists \mathbf{x}' \in B : h(\mathbf{x}') = 1 \\ & \wedge \forall \mathbf{x}'' : \mathbf{x} \rightsquigarrow \mathbf{x}'' \text{ implies } h(\mathbf{x}'') = 0 \\ 1 & \text{if } h(\mathbf{x}) = 0 \wedge \forall \mathbf{x}' \in B : h(\mathbf{x}') = 0 \\ & \wedge \exists \mathbf{x}'' : \mathbf{x} \rightsquigarrow \mathbf{x}'' \text{ and } h(\mathbf{x}) = 1 \\ 0 & \text{otherwise} \end{cases}$$

This loss is indeed closely related to the $\mathcal{H}$-$P_\mathcal{X}$-distance as $\ell^{\mathrm{gr}}(h,\rightsquigarrow,\mathbf{x},B_\rightarrow(\mathbf{x})) = |\ell^{\rightarrow,\perp}(h,\mathbf{x}) - \ell^{\rightsquigarrow,\perp}(h,\mathbf{x})|$.

The *true graph loss* with respect to some marginal $P_\mathcal{X}$ and true manipulation graph $\rightarrow$ is then defined by

$$\mathcal{L}^{\mathrm{gr}}_{(P_\mathcal{X},\rightarrow)}(h,\rightsquigarrow) = \mathbb{E}_{\mathbf{x}\sim P_\mathcal{X}}[\ell^{\mathrm{gr}}(h,\rightsquigarrow,\mathbf{x},B_\rightarrow(\mathbf{x}))].$$

Furthermore for a sample $S = \{(\mathbf{x}_1, B_1), \ldots (\mathbf{x}_n, B_n)\}$ we define the *empirical graph loss* as

$$\mathcal{L}^{\mathrm{gr}}_S(h,\rightsquigarrow) = \sum_{(\mathbf{x}_i,B_i)\in S} \ell^{\mathrm{gr}}(h,\rightsquigarrow,\mathbf{x}_i,B_i).$$

Similar to previous sections, we now want to define a loss class for $\mathcal{H}\times\mathcal{G}$. We define $g(h,\rightsquigarrow)$ to be the set of all pairs $(\mathbf{x},B) \in \mathcal{X}\times 2^\mathcal{X}$ on which $\ell^{\mathrm{gr}}(h,\rightsquigarrow,\mathbf{x},B) = 1$. Then the *graph loss class* of $\mathcal{H}\times\mathcal{G}$ is defined as

$$(\mathcal{H}\times\mathcal{G})_{\ell^{\mathrm{gr}}} = \{g(h,\rightsquigarrow) : h \in \mathcal{H} \text{ and } \rightsquigarrow\in\mathcal{G}\}.$$

We will now show that if the VC-dimension of the loss class $(\mathcal{H}\times\mathcal{G})_{\ell^{\mathrm{gr}}}$ is finite, we can indeed learn $\mathcal{G}$ with respect to ℓ^{gr}. For some examples and more discussion on the VC-dimension with respect to the loss class $(\mathcal{H}\times\mathcal{G})_{\ell^{\mathrm{gr}}}$, we refer the reader to the appendix.

Lemma 1. *Let* $\mathrm{VC}((\mathcal{H}\times\mathcal{G})_{\ell^{\mathrm{gr}}}) = d$. *Then there is* $n_{\mathrm{graph}} : (0,1)^2 \mapsto \mathbb{N}$, *such that for any marginal distribution* $P_\mathcal{X}$ *and any manipulation graph* $\rightarrow$ *for a sample* $S = \{(\mathbf{x}_1, B_\rightarrow(\mathbf{x}_1)), \ldots, (\mathbf{x}_n, B_\rightarrow(\mathbf{x}_n))\}$ *of size* $n \geq n(\epsilon,\delta)$, *we have with probability at least* $1-\delta$ *over the sample generation* $S_\mathcal{X} = (\mathbf{x}_1,\ldots,\mathbf{x}_n) \sim P^n_\mathcal{X}$ *for any* $h \in \mathcal{H}$ *and any* $\rightsquigarrow\in\mathcal{G}$

$$|\mathcal{L}^{\mathrm{gr}}_{(P_\mathcal{X},\rightarrow)}(h,\rightsquigarrow) - \mathcal{L}^{\mathrm{gr}}_S(h,\rightsquigarrow)| < \epsilon.$$

Furthermore, $n_{\mathrm{graph}}(\epsilon,\delta) \in O(\frac{d+\log\frac{1}{\delta}}{\epsilon^2})$.

We note that the above lemma is agnostic in the sense that it did not require $\rightarrow\in\mathcal{G}$. We will now introduce an empirical version of the $\mathcal{H}$-$\mathcal{P}_\mathcal{X}$-distance. This will allow us to state the main theorem of this section and show that it is indeed possible to learn $\rightsquigarrow\in\mathcal{G}$ with low $d_{\mathcal{H},P_\mathcal{X}}(\rightarrow,\rightsquigarrow)$ if $\mathrm{VC}((\mathcal{H}\times\mathcal{G})_{\ell^{\mathrm{gr}}})$ is finite.

Definition 5. *Given a sample* $S_\mathcal{X} = \{(\mathbf{x}_1,\ldots,\mathbf{x}_n)\}$ *of domain elements* $\mathbf{x}_i$ *and two manipulation graphs* $\rightarrow$ *and* $\rightsquigarrow$ *we can define the empirical* $\mathcal{H}$-$S_\mathcal{X}$*-distance as*

$$d_{\mathcal{H},S_\mathcal{X}}(\rightarrow,\rightsquigarrow) = \sup_{h\in\mathcal{H}} \sum_{\mathbf{x}_i\in S_\mathcal{X}} \ell^{\mathrm{gr}}(h,\rightsquigarrow,\mathbf{x}_i,B_\rightarrow(\mathbf{x}_i))$$

Theorem 7. *Let* $\mathrm{VC}((\mathcal{H}\times\mathcal{G})_{\ell^{\mathrm{gr}}}) = d$. *Then there is* $n_{\mathrm{dist}} : (0,1)^2 \mapsto \mathbb{N}$, *such that for any marginal distribution* $P_\mathcal{X}$ *and any manipulation graph* $\rightarrow$ *for a sample* $S = \{(\mathbf{x}_1, B_\rightarrow(\mathbf{x}_1)), \ldots, (\mathbf{x}_n, B_\rightarrow(\mathbf{x}_n))\}$ *of size* $n \geq n(\epsilon,\delta)$, *we have with probability at least* $1-\delta$ *over the sample generation* $S_\mathcal{X} = (\mathbf{x}_1,\ldots,\mathbf{x}_n) \sim P^n_\mathcal{X}$ *for any* $\rightsquigarrow\in\mathcal{G}$

$$d_{\mathcal{H},P_\mathcal{X}}(\rightarrow,\rightsquigarrow) < d_{\mathcal{H},S_\mathcal{X}}(\rightarrow,\rightsquigarrow) + \epsilon.$$

Furthermore, $n_{\mathrm{dist}}(\epsilon,\delta) \in O(\frac{d+\log\frac{1}{\delta}}{\epsilon^2})$.

Combining Theorem 7 and Theorem 6 we can thus conclude that it is indeed possible to learn $\rightsquigarrow\in\mathcal{G}$ such that using $\ell^{\rightsquigarrow}$ as a surrogate loss function guarantees a good approximation on the true strategic loss $\ell^{\rightarrow}$.

6 Conclusion

In this paper we introduced a new strategic loss, which incentivizes correct classification, but also robustness to strategic manipulation. We also incorporate the idea of social burden into our notion of loss. We differentiated this loss from previous formulations designed to mitigate strategic manipulation. In particular, we showed that optimizing for our strategic loss can yield satisfactory classification rules, even if there is no incentive-compatible hypothesis in the class that performs well on the classification task at hand. In addition, the loss formulation yields desirable effects in terms of sample complexity. Our work opens various avenues for further investigations and we hope it will inspire follow up studies on the connections between a hypothesis class and the underlying manipulation graphs, effects of these connections, as well as learnability of the manipulation graph.

References

Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; and Vaughan, J. W. 2010. A theory of learning from different domains. *Mach. Learn.*, 79(1-2): 151–175.

Brückner, M.; and Scheffer, T. 2011. Stackelberg games for adversarial prediction problems. In *Proceedings of the 17th ACM International Conference on Knowledge Discovery and Data Mining SIGKDD*, 547–555.

Cullina, D.; Bhagoji, A. N.; and Mittal, P. 2018. PAC-learning in the presence of adversaries. In *Advances in Neural Information Processing Systems*, 230–241.

Dalvi, N. N.; Domingos, P. M.; Mausam; Sanghai, S. K.; and Verma, D. 2004. Adversarial classification. In *Proceedings of the Tenth ACM International Conference on Knowledge Discovery and Data Mining SIGKDD*, 99–108.

Dong, J.; Roth, A.; Schutzman, Z.; Waggoner, B.; and Wu, Z. S. 2018. Strategic Classification from Revealed Preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation, EC*, 55–70.

Feige, U.; Mansour, Y.; and Schapire, R. 2015. Learning and inference in the presence of corrupted inputs. In *Conference on Learning Theory*, 637–657.

Haghtalab, N.; Immorlica, N.; Lucier, B.; and Wang, J. Z. 2020. Maximizing Welfare with Incentive-Aware Evaluation Mechanisms. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI*, 160–166.

Hardt, M.; Megiddo, N.; Papadimitriou, C. H.; and Wootters, M. 2016. Strategic Classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science, ITCS*, 111–122.

Hu, L.; Immorlica, N.; and Vaughan, J. W. 2019. The Disparate Effects of Strategic Manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT**, 259–268.

Jagadeesan, M.; Mendler-Dünner, C.; and Hardt, M. 2021. Alternative Microfoundations for Strategic Classification. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, 4687–4697.

Mansour, Y.; Mohri, M.; and Rostamizadeh, A. 2009. Domain Adaptation: Learning Bounds and Algorithms. In *The 22nd Conference on Learning Theory, COLT*.

Miller, J.; Milli, S.; and Hardt, M. 2020. Strategic Classification is Causal Modeling in Disguise. In *Proceedings of the 37th International Conference on Machine Learning, ICML*, 6917–6926.

Milli, S.; Miller, J.; Dragan, A. D.; and Hardt, M. 2019. The Social Cost of Strategic Classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT**, 230–239.

Montasser, O.; Hanneke, S.; and Srebro, N. 2019. VC Classes are Adversarially Robustly Learnable, but Only Improperly. In *Conference on Learning Theory, COLT*, 2512–2530.

Montasser, O.; Hanneke, S.; and Srebro, N. 2021. Adversarially Robust Learning with Unknown Perturbation Sets. In *Conference on Learning Theory, COLT 2021*, 3452–3482.

Perdomo, J. C.; Zrnic, T.; Mendler-Dünner, C.; and Hardt, M. 2020. Performative Prediction. In *Proceedings of the 37th International Conference on Machine Learning, ICML*, 7599–7609.

Shalev-Shwartz, S.; and Ben-David, S. 2014. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press.

Sundaram, R.; Vullikanti, A.; Xu, H.; and Yao, F. 2021. PAC-Learning for Strategic Classification. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, 9978–9988.

Tsirtsis, S.; and Rodriguez, M. G. 2020. Decisions, Counterfactual Explanations and Strategic Behavior. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems NeurIPS*.

Zhang, H.; and Conitzer, V. 2021. Incentive-Aware PAC Learning. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI*, 5797–5804.