Strategic Classification with Unknown User Manipulations

Tosca Lechner [1] **Ruth Urner** [2] **Shai Ben-David** [1 3]

Abstract

In many human-centric applications for Machine Learning instances will adapt to a classifier after its deployment. The field of strategic classification deals with this issue by aiming for a classifier that balances the trade-off between correctness and robustness to manipulation. This task is made harder if the underlying manipulation structure (i.e. the set of manipulations available at every instance) is unknown to the learner. We propose a novel batch-learning setting in which we use unlabeled data from previous rounds to estimate the manipulation structure. We show that in this batch-learning setting it is possible to learn a close to optimal classifier in terms of the strategic loss even without knowing the feasible manipulations beforehand. In line with recent advances in the strategic classification literature, we do not assume a best-response from agents but only require that observed manipulations are feasible.

1. Introduction

Consider the following scenario: a college or university has large amounts of records of students who at some point applied to the school, got admitted and then either succeeded or failed at obtaining a degree. Based on these records, the university sets (and publishes) admission criteria with the intent to admit students that are likely to successfully graduate. It then receives a set of applications for admission for the next year, some of which will lead to admission. In the next year (and upcoming years), the university will need to set and publish admission criteria again, its aim still being to attract and admit students that are likely to succeed. This scenario differs from a classic (statistical) decision-making setup in several ways: first, the entities that decisions are to be made for are human beings, and as such may actually adapt their application materials (as best as they can) to fit the published admission criteria. The decision maker may not know in advance in what ways the applicants can modify their credentials to get accepted. In addition, when it is time to publish a decision rule for the next round, the decision maker does not have feedback on the quality of last year's admission yet (since students usually take several years before they graduate or leave school without a degree). Thus the only information about the results from the last published decision rule was the set of (potentially strategically modified) applications.

Many human-centric real-life applications of machine learning, such as decisions on loan applications or bail recommendations, share characteristics with the above sketched scenario: there is a need for transparent classification and therefore a need (or maybe even a legal requirement) of publishing the decision rule to be used. This requirement for transparency, while in most scenarios well justified, has the effect that individuals might use this knowledge to adapt to or game the rules, i.e. they might change their feature vectors strategically in order to receive a desired outcome from the published classifier. However, this change of features often does not correspond to a change in their ground-truth label. Such feature manipulations then yield a loss in accuracy of the learned classifier after its publication. Moreover, often by the time the next round of decision making is due, the outcomes from the previous rounds are not known yet. That is, in addition to some labelled data to start with, a learner has access only to unlabelled data that potentially contains manipulated features in subsequent rounds.

The field of strategic classification, first proposed by Hardt et al (Hardt et al., 2016), studies the phenomenon of learning classifiers which are robust to strategic manipulations. The goal in strategic classification is to design a decision rule which is accurate as well as designed to withstand feature manipulations. There are two main motivations for discouraging such feature changes: either manipulated instances will be misclassified after the manipulation (resulting in false positives) or true positive instances are forced to misrepresent themselves in order to be classified correctly. This second consideration is also known as "social burden" (Milli et al., 2019) as individuals typically face a cost for this.

[1]Cheriton School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada [2]Lassonde School of Engineering, EECS Department, York University, Toronto, Ontario, Canada [3]Vector Institute, Toronto, Ontario, Canada. Correspondence to: Tosca Lechner <tlechner@uwaterloo.ca>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

It is a common assumption in the strategic classification literature, that the manipulation structure is known in terms of either a cost-of-manipulations function or a manipulation graph (the set of possible manipulations for each instance) and considers the strategic classification setup as a one-time decision making problem. However such knowledge of the feature manipulation capabilities is not always available, in particular not with exact accuracy. Furthermore, in most decision making setups where a decision rule (classifier) is learned from available data, the decision making is not a one-time event, but rather the learned decision rule is to be repeatedly employed over a long time, and ideally updated to adapt to potential changes in the data generation. In this work, we focus on the scenario where such changes are (only) the result of strategic feature adaptations to the published decision rules.

We propose a novel formalization of this batch-learning setting for strategic classification. In our framework, learning and data-generation proceed in rounds. Initially, the learner receives a labeled sample S_0 from some underlying (unmanipulated) distribution. Given this sample of labeled data, the learner decides on (and publishes) a classifier h^0. From then on, in each round t, the learner receives an unlabeled sample from the same data distribution, with the caveat that the features were strategically manipulated in response to h^{t-1}. The learner then decides on classifier h^t, based on labeled data sample S_0 and unlabeled samples $S_1, S_2, \dots S_{t-1}$. While the learner does not have access to the underlying feature manipulation capabilities of the instances, we assume that the true manipulation structure is a member of a class of possible such structures (graphs). We show that by exploiting the observed distribution shift in this batch-learning setting it is possible to learn the optimal classifier in terms of the so-called *strategic loss* (Lechner & Urner, 2022) even without knowing the underlying manipulation capabilities. For a wide range of combinations of *hypothesis classes* $\mathcal{H}$ and *manipulation graph classes* $\mathcal{G}$, we provide first positive results and analysis in this learning setup. More specifically, we derive bounds on sufficient sample sizes as well as the number of rounds for the learner to produce a classification rule with low loss. We focus in particular on graph classes which are totally ordered by inclusion, which captures the case in which it is unknown how manipulation costs compare to the value of being classified with the desired label. We show that in these cases batch-learning is possible if the VC-dimension of the loss-class of $(\mathcal{G} \times \mathcal{H})$ is finite. Roughly, the finiteness of the VC-dimension of the loss class, makes it possible to successfully estimate the distribution shift caused by a deployed hypothesis. The total order on the manipulation graphs allows to use this information to do a binary search on a discretized version of the hypothesis class.

Lastly, we show that for totally ordered $\mathcal{G}$ and $\mathcal{H}$ with finite VC-dimension it is possible to successfully improperly learn $\mathcal{H}$ with respect to $\mathcal{G}$ in the robustly realizable case with only access to unmanipulated data. In order to achieve this last result, we introduce a new paradigm, called maximally robust empirical risk minimization (MRERM) and use it to recreate the compression argument from (Montasser et al., 2019). MRERM picks a hypothesis that is robust with respect to the maximal graph that allows for robust realizability of the sample, in case such a maximal graph exists. However, such a maximal graph may not exist in some finite VC classes. We use the set-theoretic concept of ultrafilters to define an extension of the hypothesis class that is guaranteed to have an MRERM for every realizable sample and has the same VC dimension as our original class.

1.1. Related Work

The concern that learning outcomes might be compromised when agents adapt their feature vectors in response to published classification rules was first pointed out over a decade ago (Dalvi et al., 2004; Brückner & Scheffer, 2011). The area has received substantial interest from the research community in recent years, both in the context of adversarial robustness (Feige et al., 2015; Cullina et al., 2018; Montasser et al., 2019; 2021) and robustness to strategic feature manipulations. Hardt et al. formally introduced the setup where agents aim to improve their decision outcomes and termed it "strategic classification" (Hardt et al., 2016). In addition to the cost of induced misclassification, previous work has pointed out that changes to the decision boundary aiming to prevent false positives, may force true positive instances to manipulate their features for retaining their positive classification. This (also undesirable) effect has been summarized under the concept of "social burden" (Milli et al., 2019; Jagadeesan et al., 2021). It has also been shown that the cost of social burden might be disproportionately paid by underrepresented or disadvantaged sub-groups of a population (Milli et al., 2019; Hu et al., 2019). Recent work on strategic classification has pointed out that strategic feature modification can also be a positive effect, for example when applicants respond by studying better for tests and learning specific skills (Haghtalab et al., 2020), and addressed this phenomenon through a causality lense (Miller et al., 2020; Tsirtsis & Rodriguez, 2020; Shavit et al., 2020). Some recent works have further explored this interplay between gaming and improvement (Chen et al., 2021) and aligned incentives (Levanon & Rosenfeld, 2022).

While many previous studies in this area have taken a game theoretic perspective, some recent work has analyzed strategic classification in a PAC learning framework (Zhang & Conitzer, 2021; Sundaram et al., 2021; Lechner & Urner, 2022). Similarly, our work follows a new trend of not requiring agents to be cost-minimizing agents (Jagadeesan et al., 2021; Chen et al., 2020), as there is a potential limit

of the rationality of agents (Jagadeesan et al., 2021). In the PAC learning setting, the consideration of irrational agents is modelled by capturing the sets of possible manipulations in the *manipulation graph* (Zhang & Conitzer, 2021), which only distinguishes between feasible and infeasible manipulations. Using this notion of manipulation graph the objectives of discouraging strategic manipulations for the sake of avoiding misclassification and avoiding contributions to social burden have been jointly modelled in a loss function, the *strategic loss* (Lechner & Urner, 2022). We adopt this notion of loss and frame our learning goals in terms of the strategic loss function.

There has been some recent work on learning with respect to unknown manipulation structures in an online setting (Dong et al., 2018). The first results in this line of research was given in form of regret bounds for linear classifiers under the assumption that only instances of one label would manipulate (Dong et al., 2018). Similarly, Ahmadi et al introduce a version of the perceptron algorithm which takes possible feature manipulations into account (Ahmadi et al., 2021). They show finite mistakes bounds for their algorithm for both known and unknown cost functions under a linear separability assumptions, which is akin to our strategically robust realizability assumption. Both works (Dong et al., 2018; Ahmadi et al., 2021) do not require any knowledge of unmanipulated data in their setting but assume immediate label feedback for each (possibly manipulated) classified instance. Thus their results are complementary to our results in the strategic realizable case where we achieve robustness without having access to manipulated data. Furthermore, both works assume that agents are cost-minimizing, i.e., best-response. We also note that the notion of loss in those settings is slightly different, as they do not incorporate the notion of social burden into their success criterion. In the strategic PAC learning setting, there are known sufficient conditions for the strategic loss to be robust with respect to inaccuracies on the assumed manipulation structure (Lechner & Urner, 2022). Furthermore, PAC-learnability guarantees been shown with respect to an unknown manipulation (or perturbation) structure in both strategic classification (Lechner & Urner, 2022) and in adversarially robust classification (Montasser et al., 2021) with the assumption of an additional oracle. Both of these works require an oracle access that might be unrealisitic in real-world settings. While the oracle in the latter is more realistic and no further assumptions on the perturbation sets are needed, these learning guarantees additionally require the Littlestone dimension of the hypothesis class being used to be finite. This assumption is not fulfilled by most classes we consider in this paper (e.g., the simple class of thresholds classifiers as well as general finite VC-classes).

Finally, our framework bears some similarities with the setting of lifelong learning (Pentina & Lampert, 2014; Pentina & Ben-David, 2015; Balcan et al., 2015; Pentina & Urner, 2016; Balcan et al., 2020). In lifelong learning, a learning algorithm aims to perform well and adapt to a stream of related, but not identical learning tasks. Our setup distinguishes itself from standard lifelong learning goals in that the changes in input data are actually induced by the published decision rules from the previous round, while the actual target task remains the same.

1.2. Overview on our results

We consider a novel strategic batch-learning problem in which the manipulation graph is learned alongside the classification rule in order to achieve optimal classification (Definition 2.5). Importantly, we only assume prior knowledge of a graph class $\mathcal{G}$ which contains the true manipulation graph, but not exact knowledge of the true manipulation graph. We propose a formal learning protocol (Definition 2.2) and success criterion (Definition 2.5) for this setup and show that learning in this setting is possible for a wide variety of hypothesis classes $\mathcal{H}$ and graph classes $\mathcal{G}$.

In Section 3 we present possibility results for *proper learning* under the strategic batch learning protocol. As a warm-up, and to illustrate the intuition behind our techniques for a simple class, we start in Subsection 3.1 with presenting an algorithm (Algorithm 3.1) for the hypothesis class of thresholds and the class of manipulation graphs which allow manipulations within a fixed radius (while the radius of the underlying true manipulation graph is not known). We then show that this algorithm has sample complexity $O(\frac{\log(\frac{1}{\delta})}{\epsilon^2})$ and round complexity $O(1)$ in the (robustly) realizable case, as well sample complexity $O(\frac{\log(\frac{1}{\delta\epsilon})}{\epsilon^2})$ round complexity $O(\log(\frac{1}{\epsilon}))$ in the agnostic case (Observation 3.1 and Theorem 3.2).

In Subsection 3.2, we then move on to analyse proper strategic batch learning for general VC-classes. First, we show that if the joint loss class of some $\mathcal{G} \times \mathcal{H}$ with respect to the manipulation loss ℓ^{mani} (Definition 2.4) is finite, we can learn the manipulation structure for a particular hypothesis from $\mathcal{H}$ (Lemma 3.6 and Observation 3.7). We then use this to show a general learnability result for the strategic batch setting for classes with finite $\text{VC}((\mathcal{G} \times \mathcal{H})_{\ell^{\text{mani}}})$ and finite $\text{VC}(\mathcal{H})$. We furthermore give a generalization of Algorithm 3.1 in Algorithm 3.2 that works for arbitrary hypothesis classes $\mathcal{H}$ (with finite $\text{VC}((\mathcal{G} \times \mathcal{H})_{\ell^{\text{mani}}}) = d_1$ and finite $\text{VC}(\mathcal{H}) = d_2$) and totally ordered graph classes $\mathcal{G}$. We show that this algorithm has sample complexity $O(\frac{d_1+d_2+\frac{1}{\delta}}{\epsilon^2})$ and round complexity $O(1)$ in the realizable case (Observation 3.10) and sample complexity $O(\frac{d_1+d_2+\frac{1}{\delta\epsilon}}{\epsilon^2})$ and round complexity $O(\log(\frac{1}{\epsilon}))$ in the agnostic case (Theorem 3.11).

Finally, we also explore a more general, non-proper learning setup for cases where $\text{VC}((\mathcal{G} \times \mathcal{H})_{\ell^{\text{mani}}})$ is not necessarily

finite. In Section 4, we note that there are classes $\mathcal{H}$ of finite VC-dimension for which this $\mathrm{VC}((\mathcal{G} \times \mathcal{H})_{\ell^{\mathrm{mani}}})$ is not finite despite $\mathcal{G}$ being totally ordered and which cannot be learned with respect to such $\mathcal{G}$ by any proper learner. Extending techniques from the literature on learning under adversarial perturbations (Montasser et al., 2019), we show that every $\mathcal{H}$ with finite VC-dimension can be improperly learned with respect to any totally ordered graph class $\mathcal{G}$ in the realizable setting (Theorem 4.2), even if the actual underlying manipulation structure is not available to the learner.

Due to space limitations, we defer all technical proofs to the Appendix.

2. Setup

Basic learning theoretic notions We start by providing some general notation. We adopt standard notation and terminology for machine learning concepts (Shalev-Shwartz & Ben-David, 2014). We consider a classification task given by an unknown ground-truth distribution P over $\mathcal{X} \times \{0,1\}$. We use the notation $P_{\mathcal{X}}$ to denote the marginal of P over the *feature space* $\mathcal{X}$. We denote the set of all finite sequences of feature vectors (eg. samples from $P_{\mathcal{X}}^m$) by $\mathcal{X}^*$, and the set of all finite sequences of labeled feature vectors (eg. samples from P^n) by $(\mathcal{X} \times \{0,1\})^*$. As is standard in PAC-type learning guarantees the learner is evaluated with respect to a fixed *hypothesis class* $\mathcal{H} \subset \mathcal{F} = \{0,1\}^{\mathcal{X}}$. The performance of a hypothesis is evaluated by means of a *loss function* $\ell : \mathcal{H} \times \mathcal{X} \times \{0,1\} \to \mathbb{R}$, and the goal is to learn a hypothesis h with small *expected loss* $\mathcal{L}_P(h) = \mathbb{E}_{(x,y)\sim P}[\ell(h,x,y)]$. The *approximation error* of $\mathcal{H}$ with respect to loss ℓ on distribution P is $\mathrm{opt}_P(\mathcal{H}) = \inf_{h\in\mathcal{H}} \mathcal{L}_P(h)$, and it indicates how suitable class $\mathcal{H}$ is for task P. We use superscripts to identify specific loss functions. The standard (binary) classification loss is denoted as $\ell^{0/1}$ (and $\mathcal{L}_P^{0/1}(h)$ denotes the corresponding expected loss). A (standard) *learner* is a function $\mathcal{A} : (\mathcal{X} \times \{0,1\})^* \to \{0,1\}^{\mathcal{X}}$ that takes in a labelled sample and outputs a hypothesis. The requirement for learnability of a class $\mathcal{H}$ with loss function ℓ in the PAC framework (Valiant, 1984) is the existence of function $m : (0,1)^2 \to \mathbb{N}$, and a learner $\mathcal{A}$ such that, for all $\epsilon, \delta \in (0,1)$, and all $m \geq m(\epsilon,\delta)$ we have

$$\mathbb{P}_{S\sim P^m}[\mathcal{L}_P(\mathcal{A}(S)) \leq \mathrm{opt}_P(H) + \epsilon] \geq 1 - \delta.$$

It is well known, that a binary hypothesis class is PAC-learnable with respect to loss $\ell^{0/1}$ if and only if its VC-dimension is finite (Blumer et al., 1989; Vapnik & Chervonenkis, 1971). Learnability in the *realizable setting* refers to the above guarantee under the additional condition that $\mathrm{opt}_P(\mathcal{H}) = 0$. And a learner $\mathcal{A}$ is called a *proper learner* for $\mathcal{H}$ if $\mathcal{A}(S) \in \mathcal{H}$ for all samples $S \in (\mathcal{X} \times \{0,1\})^*$.

Strategic classification In strategic classification, individuals (modelled as the members of the domain $\mathcal{X}$) will try to receive a preferred label (here $y = 1$) by manipulating their feature vectors according to some *admissible manipulation*. We model the set of admissible feature manipulations as a *manipulation graph* $g = (\mathcal{X}, E_g)$, where a manipulation from x to x' is admissible if and only if the (directed) edge (x,x') exists in E_g. We will denote the neighborhood set of a point $x \in \mathcal{X}$ according to graph g by $B_g(x) = \{x' \in \mathcal{X} : (x,x') \in E_g\}$. We will denote the true manipulation graph by $g^{\to}$. We do not assume this graph to be known during the learning process. Rather, we assume the learner has prior knowledge of some graph class $\mathcal{G}$ such that $g^{\to} \in \mathcal{G}$. The class of all manipulation graphs will be denoted by $\mathcal{G}_{\mathrm{all}}$.

We assume, that if an admissible manipulation for the preferred label (i.e. the label 1) is available to an instance x given a published classifier h, then some manipulation to a positively labeled instance will occur. However, we do not assume that this is necessarily a best-response manipulation, in the sense that the instance will "move as far as possible". The following definition formalizes this notion of classifier induced manipulations for a sample of instances.

Definition 2.1 (Classifier induced manipulation of a sample)**.** Let g be a manipulation graph and h be a hypothesis. We say $\pi : \mathcal{X} \to \mathcal{X}$ is a (g,h)-*induced manipulation* if

$$\pi(x) \begin{cases} = x & \text{if } h(x) = 1 \text{ or} \\ & B_g(x) \cap h^{-1}(1) = \emptyset \\ \in B_g(x) \cap h^{-1}(1) & \text{if } h(x) = 0 \text{ and} \\ & B_g(x) \cap h^{-1}(1) \neq \emptyset \end{cases}$$

Now for a labeled sequence $S = ((x_1,y_1),\ldots,(x_m,y_m))$ of instances and sequence $\Pi = (\pi_1,\ldots,\pi_m)$ of (g,h)-induced manipulations $\pi_1,\ldots,\pi_m$, we define the Π-manipulated sample $S^{\Pi} = ((\pi_1(x_1),y_1),\ldots(\pi_m(x_m),y_m))$. Similarly for an unlabeled sample $S = (x_1,x_2,\ldots x_m)$, the Π-manipulated sample is defined by $S^{\Pi} = (\pi_1(x_1),\ldots \pi_m(x_m))$.

Note that the above definition allows for repeated feature vectors $x_i = x_j$ (with $i \neq j$) in the sequence S to move to differing manipulated instances $\pi_i(x_i) \neq \pi_j(x_j)$. For simplicity, we will often just refer to the sequence Π as a (g,h)-*induced manipulation* without specifically referring to its components $\pi_1,\ldots,\pi_m$.

To model repeated decision-making scenarios (such as the university admission task outlined in the introduction), we introduce a formal batch-learning protocol. In our protocol, a learner receives one non-manipulated labelled sample from the distribution, publishes an initial hypothesis $\hat{h}^0$, and then, in each round t, successively observes strategically

manipulated (but unlabeled) samples in response to the last published hypothesis $\hat{h}^{t-1}$. Recall that we denote the true underlying manipulation graph by $g^{\rightarrow}$.

Definition 2.2 (Strategic batch-learning protocol)**.** Let $(m_i)_{i\in\mathbb{N}}$ be a sequence of sample sizes, $m_i \in \mathbb{N}$.

Round 0**:** The learner receives a labeled sample $S \sim P^{m_0}$ and, in response, publishes classifier $\hat{h}^0 : \mathcal{X} \to \{0,1\}$.

Round t **for** $t \geq 1$**:** The learner receives an unlabeled sample $S' = S^{\Pi} \in \mathcal{X}^{m_t}$, which is a Π-manipulated version of some an unlabeled sample $S \sim P_{\mathcal{X}}^{m_t}$, where Π is a $(g^{\rightarrow}, \hat{h}^{t-1})$-induced manipulation. In response, the learner publishes classifier $\hat{h}^t$.

We call a learner $\mathcal{A}$ that operates according to the above protocol, a *strategic batch learner*. Note that the learner receives only one labeled sample from the data-generating process, in the first round. And the only information about the underlying manipulation structure $g^{\rightarrow}$ it receives, are the unlabeled, manipulated samples in the subsequent rounds.

The goal is to output a hypothesis with low *strategic loss*:

Definition 2.3 (Strategic Loss (Lechner & Urner, 2022))**.** For a given manipulation graph g, the strategic loss $\ell^g : \mathcal{F} \times \mathcal{X} \times \mathcal{Y} \to \{0,1\}$ is defined by

$$\ell^g(h,x,y) = \begin{cases} 1 & \text{if } h(x) \neq y \\ 1 & \text{if } h(x) = 0 \text{ and} \\ & \quad B_g(x) \cap h^{-1}(1) \neq \emptyset \\ 0 & \text{otherwise.} \end{cases}$$

That is, a classifier h suffers strategic loss 1, if it misclassifies an instance (x,y), or if it assigns label 0 to x while there exists an admissible manipulation x' for x with $h(x') = 1$. The following loss captures the second condition only:

Definition 2.4 (Manipulation Loss)**.** The manipulation loss $\ell^{\text{mani}} : \mathcal{G}_{\text{all}} \times \mathcal{F} \times \mathcal{X} \to \{0,1\}$ is defined by

$$\ell^{\text{mani}}(g,h,x) = \begin{cases} 1 & \text{if } h(x) = 0 \text{ and} \\ & \quad B_g(x) \cap h^{-1}(1) \neq \emptyset \\ 0 & \text{otherwise} \end{cases}$$

We note that for a fixed manipulation graph g, the manipulation loss $\ell^{\text{mani}}(g,\cdot,\cdot)$ corresponds to the strategic component loss defined in (Lechner & Urner, 2022). Moreover, $\ell^g(h,x,y) \leq \ell^{\text{mani}}(g,h,x) + \ell^{0/1}(h,x,y)$.

We now define our success criterion for a strategic batch learner:

Definition 2.5. A strategic batch learner $\mathcal{A}$ is said to learn hypothesis class $\mathcal{H}$ under graph class $\mathcal{G}$ with sample complexity $m_{\mathcal{G},\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and round complexity $T_{\mathcal{G},\mathcal{H}} : (0,1)^2 \to \mathbb{N}$, if for every $\epsilon, \delta \geq 0$, and every P over $X \times \{0,1\}$ and every true manipulation graph $g^{\rightarrow} \in \mathcal{G}$, after $T = T(\epsilon,\delta)$ many rounds, $\mathcal{A}$ outputs hypothesis h^T satisfying $L_P^{g^{\rightarrow}}(h^T) \leq \inf_{h' \in H} L_P^{g^{\rightarrow}}(h') + \epsilon$, with probability at least $(1-\delta)$ over the sample generation.

We say a learner A is a successful strategic batch learner in the *realizable* case with sample complexity $m^{\text{real}}_{\mathcal{G},\mathcal{H}}$ and round complexity $T^{\text{real}}_{\mathcal{G},\mathcal{H}}$, if it satisfies the above criterion for all distributions $P \in \mathcal{P}^{\rightarrow}$, where $\mathcal{P}^{\rightarrow}$ denotes the set of all distributions over $\mathcal{X} \times \{0,1\}$ with $\inf_{h\in\mathcal{H}} L_{P'}^{g^{\rightarrow}}(h') = 0$.

We call the learner $\mathcal{A}$ *proper* for a hypothesis class H if it only outputs hypotheses $h^t \in \mathcal{H}$ from the class $\mathcal{H}$ (in every round t).

3. Proper Batch Learning

3.1. Class of Thresholds

We start by showing that learning with respect to an unknown manipulation graph for the hypothesis class of thresholds with a simple graph class is possible. Consider $\mathcal{X} = \mathbb{R}$. The class of thresholds is defined as $\mathcal{H}_{\text{thres}} = \{h_{a,0} : \mathbb{R} \to \{0,1\} : h_{a,0}(x) = 1 \text{ iff } x > a\} \cup \{h_{a,1} : \mathbb{R} \to \{0,1\} : h_{a,1}(x) = 1 \text{ iff } x \geq a\}$. We now look at the graph class consisting of fixed-radius manipulation graphs g_r, which for any x has outgoing edges to every x' with $x' - x \leq r$, i.e. $\mathcal{G}_{\text{f.r.}} = \{g_r = (\mathbb{R}, E_r) : (x,x') \in E_r \text{ iff } x' \leq x + r\}$.

We show that for this simple class there can be indeed a successful strategic batch-learner. We first note that under the robust realizability assumption, the learner only needs one round to learn a close to optimal classifier and does not require access to any manipulated samples.

Observation 3.1. *The Strategic Batch-Learning for Thresholds Algorithm (Algorithm 1) is a proper learner for the strategic-batch learning problem for* $\mathcal{H}_{\text{thres}}$ *and* $\mathcal{G}_{\text{f.r.}}$ *in the realizable case with sample complexity* $m^{\text{real}}_{\mathcal{H},\mathcal{G}} = O(\frac{\log(\frac{1}{\delta})}{\epsilon^2})$ *and round complexity* $T^{\text{real}}_{\mathcal{H},\mathcal{G}} = 1$.

Next, we show that there also is an algorithm that solves the strategic-batch learning problem for these classes in the agnostic case.

Theorem 3.2. *The Strategic Batch-Learning for Thresholds Algorithm (Algorithm 1) is a proper learner for the strategic-batch learning problem for* $\mathcal{H}_{\text{thres}}$ *and* $\mathcal{G}_{\text{f.r.}}$ *in the hypothesis-agnostic case with sample complexity* $m_{\mathcal{H},\mathcal{G}} = O(\frac{\log(\frac{1}{\delta\epsilon})}{\epsilon^2})$ *and round complexity* $T_{\mathcal{H},\mathcal{G}} = O(\log(\frac{1}{\epsilon}))$.

This is achieved by Algorithm 1 (formal proofs are provided in the appendix). In the first step the algorithm uses the labelled sample S_0 to generate candidate graphs (which are stored as an ordered list $\mathcal{G}^0$), in such a way that the sample losses on S_0 for the corresponding optimal hypotheses in-

Algorithm 1 Strategic Batch-Learning for Thresholds

Input: parameters ϵ, ϵ'
receive sample $S_0 \sim P^m$
$L_0^{0/1} \leftarrow \inf_{h' \in \mathcal{H}_{\text{thres}}} L_{S_0}^{0/1}(h')$
for $i = 0, \ldots, \frac{1}{\epsilon}$ **do**
  $r_i \leftarrow \max\{r : \inf_{h' \in \mathcal{H}_{\text{thres}}} L_{S_0}^{g_r}(h') = L_0^{0/1} + i \cdot \epsilon\}$
end for
$\mathcal{G}^0 \leftarrow [g_{r_0}, g_{r_1}, \ldots, g_{r_{\frac{1}{\epsilon}}}]$
$\hat{h}^0 \leftarrow \text{argmin}_{h' \in \mathcal{H}_{\text{thres}}} L_{S_0}^{g_{r_0}}(h')$
publish $\hat{h}^0$
receive sample S_1, where $S_1' \sim P^m$ and $S_1 = S_1^{\Pi_1}$ for some sequence of $(g_{\rightarrow}, \hat{h}^0)$-induced manipulations Π_1.
$r_{\max} \leftarrow \max\{r : g_r \in \mathcal{G}^0 \text{ and } \mathcal{L}_{S_1}^{\text{mani}}(g_r, \hat{h}^0) = 0\}$
$r_{\min} \leftarrow \min\{r : g_r \in \mathcal{G}^0 \text{ and } \mathcal{L}_{S_0}^{\text{mani}}(g_r, \hat{h}^0) \geq \mathcal{L}_{S_0}^{\text{mani}}(g_{r_{\max}}, \hat{h}^0) - \epsilon' \text{ and } \mathcal{L}_{S_1}^{\text{mani}}(g, \hat{h}^0) = 0\}$
$\mathcal{G}^1 \leftarrow \{g_r \in \mathcal{G}^0 : r \in [r_{\min}, r_{\max}]\}$
$k_1 = \lfloor \frac{|\mathcal{G}_1|}{2} \rfloor$
$\hat{g}^1 \leftarrow \mathcal{G}^1[k_1]$, where $\mathcal{G}[k]$ refers to the k-th element of $\mathcal{G}$
$\hat{h}^1 \leftarrow \arg\min_{h \in \mathcal{H}'} L_{S_0}^{\hat{g}^1}(h)$
for rounds $t = 2, \ldots, T$ **do**
  publish $\hat{h}^{t-1}$
  receive sample S_t, where $S_t' \sim P^m$ and $S_t = S_t^{\Pi_t}$ for some sequence of $(g_{\rightarrow}, \hat{h}^{t-1})$-induced manipulations Π_t.
  $\mathcal{G}^t \leftarrow \{g \in \mathcal{G}^{t-1} : \mathcal{L}_{S_t}^{\text{mani}}(g, \hat{h}^{t-1}) = 0\}$
  $\mathcal{G}_0^t \leftarrow \{g \in \mathcal{G}^t : \mathcal{L}_{S_0}^{\text{mani}}(g, \hat{h}^{t-1}) \geq \max_{g' \in \mathcal{G}^t} \mathcal{L}_{S_0}^{\text{mani}}(g', \hat{h}^{t-1}) - \epsilon'\}$
  if $\hat{g}^{t-1} \in \mathcal{G}_0^t$ **then**
    $\mathcal{G}^t \leftarrow [\mathcal{G}^t[0], \ldots, \mathcal{G}^t[k_{t-1}]] \cap \mathcal{G}_0^t$
  else
    $\mathcal{G}^t \leftarrow \mathcal{G}_0^t$
  end if
  $k_t \leftarrow \lfloor \frac{|G^t|}{2} \rfloor$
  $\hat{g}^t \leftarrow \mathcal{G}^t[k_t]$
  $\hat{h}^t \leftarrow \min_{h \in \mathcal{H}'} L_{S_0}^{\hat{g}^t}(h)$
end for

creases in ϵ-steps. Choosing an appropriate sample size, we can guarantee that S_0 is ϵ-representative in terms of strategic loss ℓ^g for all $g \in \mathcal{G}_{\text{f.r.}}$. That is, the observed sample losses on S_0 are ϵ-close to the corresponding expected losses according to distribution P. This then guarantees that optimizing the sample loss for one of the generated candidates yields a close-to-optimal hypothesis on the ground-truth distribution.

We further use the fact, that for any $h \in \mathcal{H}_{\text{thres}}$, any distribution P over $\mathcal{X} \times \{0,1\}$ any sample S and any radii $r_1 \leq r_2$, we have that $L_P^{g_{r_1}}(h) \leq L_P^{g_{r_2}}(h)$ as well as $L_S^{g_{r_1}}(h) \leq L_S^{g_{r_2}}(h)$. Now let $g^{t-1} = g_r$ be the current candidate graph. Then the algorithm publishes a hypothesis $h^{t-1} = \text{argmin}_{h \in \mathcal{H}_{\text{thres}}} L_{S_0}^{g^{t-1}}$. There are two possibilities: (1) $L_{S_0}^{g^{\rightarrow}}(h^{t-1})$ and $L_P^{g^{\rightarrow}}(h^{t-1})$ are significantly higher than $L_{S_0}^{g^{t-1}}(h^{t-1})$. Therefore we have $r < r^{\rightarrow}$. Furthermore, with high probability, we will observe a manipulated sample S_t which was manipulated more than g^{t-1} would predict. Thus, g^{t-1} would not be in the updated sets of candidate graphs $\mathcal{G}_0^t$ consistent with the observed S_t. Similarly graphs $g_{r'}$ with $r' < r$ are eliminated from the candidate set. (2) $L_{S_0}^{g^{\rightarrow}}(h^{t-1})$ and $L_P^{g^{\rightarrow}}(h^{t-1})$ are not significantly higher than $L_{S_0}^{g^{t-1}}(h^{t-1})$. In this case, the observed sample S_t would be consistent with the current g^{t-1}. In this case all candidate graphs $g_{r''}$ with $r < r''$ are eliminated from the candidate set. In the case in which $r > r^{\rightarrow}$, this obviously does not pose a problem. Now consider the case in which $r < r^{\rightarrow}$. Then for $h^* = \arg\min_{h \in \mathcal{H}_{\text{thres}}} L^{g^{\rightarrow}}(h)$ we have that $L_P^{g_r}(h^*) \leq L_P^{g_r^{\rightarrow}}(h^*)$. Now assuming that S_0 is ϵ''-representative for P with respect to $\mathcal{H}_{\text{thres}}$ and loss ℓ^g for every $g \in \mathcal{G}_{g.r.}$, then $L_P^{g_r}(h^{t-1}) \leq L_{S_0}^{g_r}(h^{t-1}) + \epsilon'' \leq L_{S_0}^{g_r}(h^*) + \epsilon'' \leq L_P^{g_r}(h^*) + 2\epsilon'' \leq L_P^{g^{\rightarrow}}(h^*) + 2\epsilon''$. Thus, in this case, despite $r < r^{\rightarrow}$, the selected hypothesis is still close to optimal in terms of $\ell^{g^{\rightarrow}}$. Thus, the elimination of graphs with radius greater than r does not hinder success. Thus, the distinction of the two cases can be exploited by the algorithm to do a binary search on the candidate set.

We also note that the way the candidate hypotheses are picked, we always pick the maximal radius for a given sample loss. This leads to corresponding maximally robust hypotheses, allowing for the first hypotheses to be successful in the realizable case as shown in Observation 3.1.

3.2. General VC-Classes

We will now show that similar learning guarantees in our strategic batch learning setup are possible for more general hypothesis classes and graph classes. We start by addressing the problem of estimating the manipulation graph from two unlabeled samples, an un-manipulated and a manipulated sample, from the marginal distributions. To this aim, we define the loss class of the Cartesian product of a hypothesis

class $\mathcal{H}$ and a graph class $\mathcal{G}$.

Definition 3.3. The loss class of $\mathcal{G} \times \mathcal{H}$ with respect to ℓ^{mani} is defined as

$$(\mathcal{G} \times \mathcal{H})_{\ell^{\text{mani}}} =$$

$$\{\{x \in \mathcal{X} : \ell^{\text{mani}}(g, h, x) = 1\} : g \in \mathcal{G} \text{ and } h \in \mathcal{H}\}$$

Furthermore for a fixed h let the class $\mathcal{G}_{\ell^{\text{mani}},h}$ be defined as

$$\mathcal{G}_{\ell^{\text{mani}},h} = \{\{x \in \mathcal{X} : \ell^{\text{mani}}(g, h, x) = 1\} : g \in \mathcal{G}\}$$

We will define $\text{VC}(\mathcal{G}_{\ell^{\text{mani}},\mathcal{H}}) = \sup_{h \in \mathcal{H}} \text{VC}(\mathcal{G}_{\ell^{\text{mani}},h})$.

Observation 3.4.
- *The VC-dimension of* $(\mathcal{G}_{\text{f.r.}} \times \mathcal{H}_{\text{thres}})_{\ell^{\text{mani}}}$ *is 2.*
- *Let* $\mathcal{H}_{half} = \{h_w : w \in \mathbb{R}^d : h_w(x) = 1$ *iff* $x^T w \geq 0\}$ *the hypothesis class of linear half spaces and* $\mathcal{G}^d_{\text{f.r.}} = \{g_r = (\mathcal{X}, E_{g_r}) : (x, x') \in E_{g_r}$ *iff* $||x - x'||_2 \leq r\}$ *be the class of fixed-radius balls in* $\mathbb{R}^d$*. Then the VC dimension of* $(\mathcal{G}^d_{\text{f.r.}} \times \mathcal{H}_{half})_{\ell^{\text{mani}}}$ *is at most* $2d$*.*

We can now use this definition to derive a sample complexity bound for estimating the region of manipulation for any particular $h \in \mathcal{H}$ from one manipulated and one un-manipulated sample. We use the following notion of disagreement between two manipulation graphs:

Definition 3.5. Given a distribution D over a domain set $\mathcal{X}$ (a.k.a. a marginal distribution), a classifier h and two manipulation graphs, g, g'

$$Dis_{(D,h)}(g, g') =$$

$$D[\{x \in \mathcal{X} : \ell^{\text{mani}}(g, h, x) \neq \ell^{\text{mani}}(g', h, x)\}]$$

Lemma 3.6. *Let* $\mathcal{G}, \mathcal{H}$ *be such that* $\text{VC}(\mathcal{G}_{\ell^{\text{mani}},\mathcal{H}}) = d$*. Let* $A_{\text{graph}} : \mathcal{X}^* \times \mathcal{X}^* \times \mathcal{H} \to 2^{\mathcal{G}}$ *be a learner following the Empirical Manipulation Estimation Paradigm (as defined in Algorithm 2). Then* A_{graph} *has the following success guarantee for learning the manipulation graph:*

For every marginal distribution $P_{\mathcal{X}}$*, every* $g^{\rightarrow} \in \mathcal{G}$*, every* $h \in \mathcal{H}$ *and every sequence of* $(g^{\rightarrow}, h)$*-induced manipulations* Π*, and every* $\epsilon, \delta \in (0, 1)$*,*

if $m \geq C\frac{d+\log(\frac{1}{\delta})}{\epsilon^2}$ *(for some universal constant* C*) with probability at least* $(1 - \delta)$*, over samples* $S_1 \sim P^m_{\mathcal{X}}, S_2 \sim P^m_{\mathcal{X}}$*, for every* $\hat{g} \in A_{\text{graph}}(S_1, S_2^{\Pi}, h)$*,*

$$Dis_{(P_{\mathcal{X}},h)}(g^{\rightarrow}, \hat{g}) \leq \epsilon.$$

and

$$g^{\rightarrow} \in A_{\text{graph}}(S_1, S_2^{\Pi}, h).$$

The key tool for proving the above lemma is the notion of a sample S_1 being ϵ-representative with respect to $(\mathcal{G} \times \mathcal{H})_{\ell^{\text{mani}}}$. For any $g \in \mathcal{G}$ and $h \in \mathcal{H}$ the empirical manipulated loss over such a sample is ϵ-close to its true loss. Standard uniform convergence theory (see, e.g., (Shalev-Shwartz & Ben-David, 2014) Chapter 4) shows that, given a class of finite VC-dimension, for any data generating distribution, a large enough sample will be ϵ-representative with respect to that class.

Observation 3.7. *If an un-manipulated sample* S_1 *is* ϵ*-representative with respect to* $(\mathcal{G} \times \mathcal{H})_{\ell^{\text{mani}}}$*, then it can be indefinitely re-used by* A_{graph} *for any hypothesis* $h \in \mathcal{H}$ *and any manipulated* ϵ*-representative samples* S_2^{Π}*. Thus if* $\text{VC}(\mathcal{G} \times \mathcal{H})_{\ell^{\text{mani}}} = d$*, then* $m \geq C\frac{d+\log(\frac{1}{\delta})}{\epsilon^2}$ *(for some universal constant* C*) , implies that with probability* $1 - \delta$ *any* $S_1 \sim P^m$ *is repeatedly reusable by* A_{graph} *to guarantee* ϵ*-success as in the Lemma above. This allows us to reuse the initial unmanipulated sample in all subsequent steps.*

Algorithm 2 EmpiricalManipulationEstimation (realizable)

Input: graph class $\mathcal{G}$, hypothesis h, input samples S_1 and S_2^{Π},parameter ϵ
Output set of candidate manipulation graphs $\mathcal{G}_c$
$L_{max} \leftarrow \max_{g \in \mathcal{G}} \mathcal{L}^{\text{mani}}_{S_1}(g, h)$ s.t. $\mathcal{L}^{\text{mani}}_{S_2}(g, h) = 0$
$\mathcal{G}_c \leftarrow \{g \in \mathcal{G} : \mathcal{L}^{\text{mani}}_{S_1}(g, h) \in [L_{max} - \epsilon, L_{max}]\}$

To generalize the above algorithms to richer classes and higher data dimensions, will now define a partial order for the graph class. We will then show that if a graph class is totally ordered with respect to this partial order, we can give a similar algorithm to the one in the threshold case with a similar guarantee.

Definition 3.8. For a hypothesis class $\mathcal{H}$, let the $\ell^{\text{mani}} - \mathcal{H}$-induced partial order $\preceq_{\mathcal{H}}$ on manipulation graphs be defined by: $g_1 \preceq_{\mathcal{H}} g_2$, if and only if for every $h \in \mathcal{H}$ and every $x \in \mathcal{X}$, we have $\ell^{\text{mani}}(g_1, h, x) \leq \ell^{\text{mani}}(g_2, h, x)$.
A graph class $\mathcal{G}$ is *totally ordered* with respect to $\preceq_{\mathcal{H}}$ if for every distinct $g_1, g_2 \in \mathcal{G}$, we have that either $g_1 \preceq_{\mathcal{H}} g_2$ or $g_2 \preceq_{\mathcal{H}} g_1$. For a subset $A \subset \mathcal{G}$ of a totally ordered graph class, we define $max_{\preceq_{\mathcal{H}}}$ as the graph $g \in A$ with $g' \preceq g$ for all $g' \in A$.

Observation 3.9.
- *A graph class* $\mathcal{G}$ *is totally ordered with respect to the class of all hypotheses* $\mathcal{F}$ *if and only if for every distinct* $g_1, g_2 \in \mathcal{G}$ *either* g_1 *is a subgraph of* g_2 *or* g_2 *is a subgraph of* g_1*.*
- *For* $\mathcal{H}_1 \subset \mathcal{H}_2$ *and two graphs* g_1, g_2 $g_1 \preceq_{\mathcal{H}_2} g_2$ *implies* $g_1 \preceq_{\mathcal{H}_1} g_2$*. Thus, if a graph class* $\mathcal{G}$ *is totally ordered with respect to* $\mathcal{H}_2$ *it is also totally ordered with respect to* $\mathcal{H}_1$*.*
- *If* $\mathcal{G}$ *is totally ordered with respect to* $\mathcal{H}$*, then* $\text{VC}(\mathcal{G}_{\ell^{\text{mani}},\mathcal{H}}) = 1$.

- *There are $\mathcal{G}$ and $\mathcal{H}$, such that* $\mathrm{VC}(\mathcal{H}) = d$ *and $\mathcal{G}$ is totally ordered with respect to $\mathcal{H}$, but* $\mathrm{VC}((\mathcal{H} \times \mathcal{G})_{\ell^{\mathrm{mani}}}) = \infty$.

We can now generalize our threshold algorithm to an algorithm for totally ordered graph classes (Algorithm 3). This algorithm essentially works in the same way: It first identifies the maximal graph g_0 for which there is a hypothesis with maximum accuracy that is still robust with respect to g_0. The corresponding optimal hypothesis is the first hypothesis $\hat{h}^0$ published by the algorithm. For the robust-realizable case, this is sufficient to guarantee success (see Observation 3.10), as we get uniform convergence of $\mathcal{H}$ both in terms of ℓ^m with respect to $\mathcal{G}$ and in terms of $0/1$-loss, which is sufficient to guarantee uniform convergence in terms of the strategic loss. For the agnostic case other candidate graphs are generated in a similar fashion. The upper bound on the strategic loss is increased by ϵ-steps, and for each such increment the maximal graph g_i, which allows for a classifier $h \in \mathcal{H}$ with corresponding loss $L^{g_i}_{S_0}(h)$ to be smaller than that bound, is identified. For these $\frac{1}{\epsilon}$ many candidate graphs $\{g_0, \ldots, g_{\frac{1}{\epsilon}}\}$ we can then perform a kind of binary search by always publishing the optimal classifier with respect to the current median classification graph. We can then update the set of candidate graphs in each round by observing the manipulations caused by the published classifiers. We are guaranteed to always observe manipulations if robustness was under-estimated, giving the algorithm sufficient feedback for the binary search to terminate successfully. This yields the guarantee in Theorem 3.11.

Observation 3.10. *Let* $\mathrm{VC}(\mathcal{H} \times \mathcal{G})_{\ell^{\mathrm{mani}}} = d_1$ *and* $\mathrm{VC}(\mathcal{H}) = d_2$. *Furthermore let $\mathcal{G}$ be totally ordered with respect to $\mathcal{H}$. Then Algorithm 3 is a successful proper strategic batch learner in the realizable case with sample complexity* $m^{\mathrm{real}}_{\mathcal{H},\mathcal{G}}(\epsilon, \delta) = O(\frac{(d_1+d_2)\log(d_1+d_2)+\frac{1}{\delta}}{\epsilon^2})$ *and round complexity* $T^{\mathrm{real}}_{\mathcal{H},\mathcal{G}}(\epsilon, \delta) = 1$.

Theorem 3.11. *Let* $\mathrm{VC}(\mathcal{H} \times \mathcal{G})_{\ell^{\mathrm{mani}}} = d_1$ *and* $\mathrm{VC}(\mathcal{H}) = d_2$. *Furthermore, let $\mathcal{G}$ be totally ordered with respect to $\mathcal{H}$. Then Algorithm 3 is a successful proper strategic batch learner with sample complexity* $m_{\mathcal{H},\mathcal{G}}(\epsilon, \delta) = O(\frac{(d_1+d_2)\log(d_1+d_2)+\log(\frac{1}{\delta\epsilon})}{\epsilon^2})$ *and round complexity* $T_{\mathcal{H},\mathcal{G}}(\epsilon, \delta) = O(\log(\frac{1}{\epsilon}))$.

4. Improper Learning

As noted in Observation 3.9, it can be the case that $\mathrm{VC}(\mathcal{H})$ and $\mathrm{VC}(\mathcal{G})_{\ell^{\mathrm{mani}},\mathcal{H}}$ are finite, but $\mathrm{VC}(\mathcal{G} \times \mathcal{H})_{\ell^{\mathrm{mani}}}$ is still infinite. In particular, this is the case for any $\mathcal{G}$ that contains a $g \in \mathcal{G}$ such that $\mathrm{VC}(\mathcal{H}_{\ell^g})$ is infinite, where $\mathcal{H}_{\ell^g}$ is the loss class of $\mathcal{H}$ with respect to the strategic loss ℓ^g. Furthermore, it has been shown that there are hypothesis classes $\mathcal{H}$ with finite VC-dimension but infinite VC-dimension of the loss class $\mathcal{H}_{\ell^g}$, which are not properly strategically robust learn-

Algorithm 3 Strategic Batch-Learning for totally ordered graph classes

Input: parameters ϵ, ϵ', hypothesis class $\mathcal{H}$, graph class $\mathcal{G}$
receive sample $S_0 \sim P^m$
$L^{0/1}_0 \leftarrow \inf_{h' \in \mathcal{H}} L^{0/1}_{S_0}(h')$
for $i = 0, \ldots, \frac{1}{\epsilon}$ **do**
  $g_i \leftarrow \max_{\preceq_{\mathcal{H}}} \{g : \inf_{h' \in \mathcal{H}} L^g_{S_0}(h') = L^{0/1}_0 + i \cdot \epsilon\}$
end for
set $\mathcal{G}^0 = [g_0, g_1, g_2, \ldots, g_{\frac{1}{\epsilon}}]$
$\hat{h}^0 \leftarrow \arg\min_{h \in \mathcal{H}} L^{g_0}_{S_0}(h)$
publish $\hat{h}^0$
receive sample S_1, where $S'_1 \sim P^m$ and $S_t = S'^{\Pi_t}_t$ for some sequence of $(g_{\rightarrow}, h^0)$-induced manipulations Π_1.
$g_{\max} \leftarrow \max_{\preceq_{\mathcal{H}}} \{g \in \mathcal{G}^0 : \mathcal{L}^{\mathrm{mani}}_{S_1}(g, \hat{h}^0) = 0\}$
$g_{\min} \leftarrow \min_{\preceq_{\mathcal{H}}} \{g \in \mathcal{G}^0 : \mathcal{L}^{\mathrm{mani}}_{S_0}(g, \hat{h}^0) \geq \mathcal{L}^{\mathrm{mani}}_{S_0}(g, \hat{h}^0) - \epsilon'$ and $\mathcal{L}^{\mathrm{mani}}_{S_1}(g, \hat{h}^0) = 0\}$]
$\mathcal{G}^1 \leftarrow \{g \in \mathcal{G}^0 : g \preceq_{\mathcal{H}} g_{\max}$ and $g_{\min} \preceq_{\mathcal{H}} g\}$
$k_1 = \lfloor \frac{|G_1|}{2} \rfloor$
$\hat{g}^1 \leftarrow \mathcal{G}^1[k_1]$
$\hat{h}^1 \leftarrow \arg\min_{h \in \mathcal{H}'} L^{\hat{g}^1}_{S_0}(h)$
for rounds $t = 2, \ldots, T$ **do**
  publish $\hat{h}^{t-1}$
  receive sample S_t, where $S'_t \sim P^m$ and $S_t = S'^{\Pi_t}_t$ for some sequence of $(g^{\rightarrow}, h^{t-1})$-induced manipulations Π_t.
  $\mathcal{G}^t \leftarrow \{g \in \mathcal{G}^{t-1} : \mathcal{L}^{\mathrm{mani}}_{S_t}(g, \hat{h}^{t-1}) = 0\}$
  $\mathcal{G}^t_0 \leftarrow \{g \in \mathcal{G}^t : \mathcal{L}^{\mathrm{mani}}_{S_0}(g, \hat{h}^{t-1}) \geq \max_{g' \in \mathcal{G}^t} \mathcal{L}^{\mathrm{mani}}_{S_0}(g', \hat{h}^{t-1}) - \epsilon'\}$
  if $\hat{g}^{t-1} \in \mathcal{G}^t_0$ **then**
    $\mathcal{G}^t \leftarrow [\mathcal{G}^t[0], \ldots, \mathcal{G}^t[k_{t-1}]] \cap \mathcal{G}^t_0$
  else
    $\mathcal{G}^t \leftarrow \mathcal{G}^t_0$
  end if
  $k_t \leftarrow \lfloor \frac{|G^t|}{2} \rfloor$
  $\hat{g}^t \leftarrow \mathcal{G}^t[k_t]$
  $\hat{h}^t \leftarrow \min_{h \in \mathcal{H}'} L^{\hat{g}^t}_{S_0}(h)$
end for

able by any learner (Lechner & Urner, 2022). The following observation is a corollary of these results from the literature.

Observation 4.1. *There are classes $\mathcal{H}$ of finite VC-dimension and graph classes $\mathcal{G}$ which are totally ordered with respect to $\mathcal{H}$, such that there is no proper successful batch-learner for $\mathcal{H}$ and $\mathcal{G}$ for any finite sample and round complexity, not even in the realizable case.*

However, in the PAC-learning setting with a fixed hypothesis class it has been shown that any class of finite VC dimension can be improperly strategically robustly learned (Montasser et al., 2019; Lechner & Urner, 2022). We can generalize these positive results in the realizable case to a setting where the manipulation graph is unknown, but the learner is given the prior knowledge that the true manipulation graph comes from a known, totally ordered graph class. We exploit the fact that here we assume realizability in the strategically robust sense and that therefore any sample will be robustly realizable with probability 1. Now let $\mathcal{G}$ be a totally ordered graph class and $\mathcal{H}$ a hypothesis class. We define the maximally robust graph in $\mathcal{G}$ with respect to a sample S and class $\mathcal{H}$ as

$$\mathrm{MG}_{\mathcal{H},\mathcal{G}}(S) = \max_{\preceq_{\mathcal{H}}}\{g \in \mathcal{G} : \min_{h\in\mathcal{H}} L_S^g(h) = 0\}.$$

In cases where this maximum does indeed exist, we then define a *maximally robust empirical risk minimizer (MRERM)* with respect to $\mathcal{H}$ and $\mathcal{G}$ as the hypothesis in $\mathcal{H}$ minimizing the strategically robust loss with respect to $\mathrm{MG}_{\mathcal{H},\mathcal{G}}(S)$, i.e.,

$$\mathrm{MRERM}_{\mathcal{H},\mathcal{G}}(S) \in \arg\min_{h\in\mathcal{H}} L_S^{MG_{\mathcal{H},\mathcal{G}}(S)}(h).$$

In cases where the maximum graph does not exist, we need to instead pick a hypothesis $h_S^{\max}$ that is simultaneously robust with respect to every graph in $\mathcal{G}$ for which S is $\mathcal{H}$-realizable to achieve our guarantee. Such a hypothesis might not always exist within $\mathcal{H}$. However, we can extend $\mathcal{H}$ to a class $\mathcal{H}'$, with $\mathrm{VC}(\mathcal{H}') = \mathrm{VC}(\mathcal{H})$ and such that for every finite sample $S \subset \mathcal{X}$ the class $\mathcal{H}'$ contains a hypothesis $h_S^{\max}$. In order to define such a hypothesis class and general MRERM rigorously we need to use the set-theoretic concept of ultrafilters. For an explanation and proof that such $\mathcal{H}'$ and MRERM always exist, we refer the reader to the appendix.

We note that in the realizable setting, the maximal graph used for the estimation here will always overestimate the robustness of the true manipulation graph $g^{\rightarrow}$. We use this fact to define a successful improper learner and prove the following theorem based on techniques from the literature on PAC learning with respect to adversarial perturbations (Montasser et al., 2019).

Theorem 4.2. *Let $\mathrm{VC}(\mathcal{H})$ be finite and let $\mathcal{G}$ be totally ordered with respect to $\mathcal{H}$. Then there is a strategically robust (improper) PAC-learner (i.e., a PAC-learner with respect to $\ell^{g^{to}}$-loss) which is successful for every $g^{\rightarrow} \in \mathcal{G}$ in the strategically robustly realizable case (i.e. when $\inf_{h\in\mathcal{H}} L_P^{g^{\rightarrow}}(h) = 0$).*

Note that the learner is successful, even without knowing the true manipulation graph $g^{\rightarrow}$ and without receiving any local perturbation sets as input.

5. Ethics discussion

As machine learning applications seem to be infiltrating all aspects of society as well as individual people's lives, it becomes increasingly important to develop tools and frameworks to analyze and provide performance and safety guarantees for diverse settings beyond the standard one-time supervised learning task from iid data. We view our work in line with studies that aim to provide solid foundations for non-standard learning settings and make such foundations applicable to more realistic application scenarios. We believe that developing a more thorough understanding and dependable theory will ultimately benefit machine learning practitioners as well as policymakers that need to shape the legal landscape in which machine learning practitioners operate.

While our work does not include implementations of algorithms or algorithmic frameworks (and as such can not be abused "directly"), we do believe that the algorithms we develop will be beneficial in scenarios of repeated learning and decision-making tasks with strategic agents. Whether such an application is morally commendable highly depends on the actors and objects of the application (as it does in any decision-making scenario that involves or affects human lives). We acknowledge that (as has been pointed out in the literature (Hu et al., 2019)) policies that are designed to discourage or prevent strategic responses to decision rules, might disproportionately affect underrepresented and/or disadvantaged segments of society. Developing mechanisms to address (and potentially counteract) such effects is an important complementary task to our study, which is however not a part of this submission.

Acknowledgements

We would like to thank Vinayak Pathak and Alex Bie for helpful discussions. Tosca Lechner was supported by a Vector Research Grant and a Waterloo Apple PhD Fellowship in Data Science and Machine Learning. Ruth Urner is also a faculty affiliate member of Toronto's Vector institute. The last two authors' research is supported through NSERC discovery grants.

References

Ahmadi, S., Beyhaghi, H., Blum, A., and Naggita, K. The strategic perceptron. In Biró, P., Chawla, S., and Echenique, F. (eds.), *EC '21: The 22nd ACM Conference on Economics and Computation, Budapest, Hungary, July 18-23, 2021*, pp. 6–25. ACM, 2021.

Balcan, M., Blum, A., and Vempala, S. S. Efficient representations for lifelong learning and autoencoding. In *Proceedings of The 28th Conference on Learning Theory, COLT*, pp. 191–210, 2015.

Balcan, M., Blum, A., and Nagarajan, V. Lifelong learning in costly feature spaces. *Theor. Comput. Sci.*, 808:14–37, 2020.

Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. K. Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4):929–965, 1989.

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

Chen, Y., Liu, Y., and Podimata, C. Learning strategy-aware linear classifiers. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NeurIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Chen, Y., Wang, J., and Liu, Y. Linear classifiers that encourage constructive adaptation, 2021.

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

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

Dong, J., Roth, A., Schutzman, Z., Waggoner, B., and Wu, Z. S. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation, EC*, pp. 55–70, 2018.

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

Haghtalab, N., Immorlica, N., Lucier, B., and Wang, J. Z. Maximizing welfare with incentive-aware evaluation mechanisms. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI*, pp. 160–166, 2020.

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

Hu, L., Immorlica, N., and Vaughan, J. W. The disparate effects of strategic manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT**, pp. 259–268, 2019.

Jagadeesan, M., Mendler-Dünner, C., and Hardt, M. Alternative microfoundations for strategic classification. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, pp. 4687–4697, 2021.

Lechner, T. and Urner, R. Learning losses for strategic classification. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022*, pp. 7337–7344, 2022.

Levanon, S. and Rosenfeld, N. Generalized strategic classification and the case of aligned incentives. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12593–12618. PMLR, 17–23 Jul 2022.

Miller, J., Milli, S., and Hardt, M. Strategic classification is causal modeling in disguise. In *Proceedings of the 37th International Conference on Machine Learning, ICML*, pp. 6917–6926, 2020.

Milli, S., Miller, J., Dragan, A. D., and Hardt, M. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT**, pp. 230–239, 2019.

Montasser, O., Hanneke, S., and Srebro, N. VC classes are adversarially robustly learnable, but only improperly. In *Conference on Learning Theory, COLT*, pp. 2512–2530, 2019.

Montasser, O., Hanneke, S., and Srebro, N. Adversarially robust learning with unknown perturbation sets. In *Conference on Learning Theory, COLT 2021*, pp. 3452–3482, 2021.

Pentina, A. and Ben-David, S. Multi-task and lifelong learning of kernels. In *Algorithmic Learning Theory - 26th International Conference, ALT*, pp. 194–208, 2015.

Pentina, A. and Lampert, C. H. A pac-bayesian bound for lifelong learning. In *Proceedings of the 31th International Conference on Machine Learning, ICML*, pp. 991–999, 2014.

Pentina, A. and Urner, R. Lifelong learning with weighted majority votes. In *Advances in Neural Information Processing Systems 29, NIPS*, pp. 3612–3620, 2016.

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

Shavit, Y., Edelman, B. L., and Axelrod, B. Causal strategic linear regression. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8676–8686. PMLR, 2020.

Sundaram, R., Vullikanti, A., Xu, H., and Yao, F. Pac-learning for strategic classification. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, pp. 9978–9988, 2021.

Tsirtsis, S. and Rodriguez, M. G. Decisions, counterfactual explanations and strategic behavior. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems NeurIPS*, 2020.

Valiant, L. G. A theory of the learnable. *Commun. ACM*, 27 (11):1134–1142, 1984.

Vapnik, V. N. and Chervonenkis, A. Y. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971.

Zhang, H. and Conitzer, V. Incentive-aware PAC learning. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI*, pp. 5797–5804, 2021.

A. Note on the maximally robust ERM paradigm

The proof of Theorem 4.2, employs the notions of a maximal graph within $\mathcal{G}$ and maximal ERM hypothesis within the class $\mathcal{H}$. Recall the definitions from Section 4:

- $\mathrm{MG}_{\mathcal{H},\mathcal{G}}(S) = \max_{\preceq_{\mathcal{H}}}\{g \in \mathcal{G} : \min_{h\in\mathcal{H}} \mathcal{L}_S^g(h) = 0\}$
- $\mathrm{MRERM}_{\mathcal{H},\mathcal{G}}(S) \in \mathrm{argmin}_{h\in\mathcal{H}}\mathcal{L}_S^{MG_{\mathcal{H},\mathcal{G}}(S)}(h)$

However, in general, these do not necessarily exist in these classes. In this section, we show that we can always embed the original classes $\mathcal{G}$ and $\mathcal{H}$ in such a way that the above notions are well-defined.

Consider a fixed sample $S = ((x_1, y_1), \ldots, (x_m, y_m))$ that is realizable with respect to the strategic loss for some $g^{\rightarrow} \in \mathcal{G}$. Then the set

$$\mathcal{G}_S = \{g \in \mathcal{G} : \min_{h\in\mathcal{H}} \mathcal{L}_S^g(h) = 0\} \subseteq \mathcal{G}$$

is not empty, and we can define a new manipulation graph $\mathrm{MG}_{\mathcal{H},\mathcal{G}}(S)$ by taking the union of all edge sets of graphs in $\mathcal{G}_S$ as the new edge set for this maximal graph. Now, if the set

$$\mathrm{argmin}_{h\in\mathcal{H}} L_S^{MG_{\mathcal{H},\mathcal{G}}(S)}(h)$$

is not empty, then we can choose any element in this set to define a maximally robust ERM hypothesis $\mathrm{MRERM}_{\mathcal{H},\mathcal{G}}(S)$. Note that this is always the case if the graph class $\mathcal{G}$ (and therefore the set $\mathcal{G}_S$) is finite.

Thus we now focus on the case where $\mathcal{G}_S$ induced by some S is infinite and the set $\mathrm{argmin}_{h\in\mathcal{H}} L_S^{MG_{\mathcal{H},\mathcal{G}}(S)}(h)$ is empty. We will use the concept of an *ultrafilter* to define a maximal hypothesis $h_S^{\max}$ for the sample S. Finally, we will show that adding all possible (over all labelled samples S) such maximal hypotheses yields a hypothesis class $\mathcal{H}'$ with $\mathcal{H} \subset \mathcal{H}'$ and $\mathrm{VC}(\mathcal{H}') = \mathrm{VC}(\mathcal{H})$.

Definition A.1 (Filter)**.** Let $\mathcal{Z}$ be some set and let $\mathcal{F} \subseteq 2^{\mathcal{Z}}$ be a collection of subsets of $\mathcal{Z}$. We call $\mathcal{F}$ a *filter* if the following conditions are satisfied:

- $\mathcal{F} \neq \emptyset$
- $\emptyset \notin \mathcal{F}$
- $\mathcal{F}$ is upwards inclusion closed: if $A \in \mathcal{F}$ and $A \subseteq B$, then $B \in \mathcal{F}$
- $\mathcal{F}$ is closed under finite intersections: if $A \in \mathcal{F}$ and $B \in \mathcal{F}$, then $(A \cap B) \in \mathcal{F}$.

A filter $\mathcal{F}$ is an *ultrafilter* if for every domain subset $C \subseteq \mathcal{Z}$, either C or its complement $\bar{C}$ is a member of $\mathcal{F}$.

It can be shown (using Zorn's lemma) that every filter $\mathcal{F}$ can be extended to an ultra-filter. That is, there always exists an ultrafilter $\widetilde{\mathcal{F}}$ with $\mathcal{F} \subseteq \widetilde{\mathcal{F}}$.

We will start by defining a filter over $\mathcal{G}_S$. Note that $\mathcal{G}_S$ is totally ordered. A *final segment* $G \subseteq \mathcal{G}_S$ is a subset such that $g \in G$ and $g \preceq g'$ implies $g' \in G$. Now consider the collection $\mathcal{F} \subseteq 2^{\mathcal{G}_S}$ defined by:

$$\mathcal{F} = \{G \subseteq \mathcal{G}_S \mid G \text{ contains a final segment of } \mathcal{G}_S\}.$$

It is not difficult to see that $\mathcal{F}$ is indeed a filter. Let $\widetilde{\mathcal{F}}$ denote an ultra-filter extending if (that is, $\mathcal{F} \subseteq \widetilde{\mathcal{F}}$).

Recall that for every $g \in \mathcal{G}_S$ there exists at least one $h \in \mathcal{H}$ such that $\mathcal{L}_S^g(h) = 0$. Let $\tau : \mathcal{G}_S \to \mathcal{H}$ be a function that assigns every manipulation graph $g \in \mathcal{G}_S$ such a hypothesis $h_g = \tau(g)$. The image of τ is thus a subset $\mathcal{H}_S \subseteq \mathcal{H}$ of the hypothesis class of empirical risk minimizers corresponding to the graphs in $\mathcal{G}_S$:

$$\mathcal{H}_S = \{h_g \mid g \in \mathcal{G}_S\}.$$

We can consider the set $\mathcal{H}_S$ as indexed by $\mathcal{G}_S$ and therefore inheriting the order of $\mathcal{G}_S$ (and we can thus refer to final segments of $\mathcal{H}_S$ etc). Note that, if for some $x \in \mathcal{X}$, and $y \in \{0,1\}$, there exists a final segment of $\mathcal{H}_S$, in which all functions assign x the same label y, then the set

$$\{g \in \mathcal{G}_S \mid h_g(x) = y\} \in \mathcal{F}$$

is an element of the filter $\mathcal{F}$ (since the filter is upward inclusion closed). For such cases, the limit function $h_S^{\max}$ will assign this label y to x.

Observe that for every x, the set $\{g \in \mathcal{G}_S \mid h_g(x) = y\}$ is the complement (in $\mathcal{G}_S$) of the set $\{g \in \mathcal{G}_S \mid h_g(x) = 1-y\}$, and thus exactly one of these sets is a member of the ultrafilter $\widetilde{\mathcal{F}}$ that contains $\mathcal{F}$. We can thus define the limit function on the whole domain $\mathcal{X}$ by

$$h_S^{\max}(x) = \begin{cases} 1 \text{ if } \{g \in \mathcal{G}_S \mid h_g(x) = 1\} \in \widetilde{\mathcal{F}} \\ 0 \text{ else.} \end{cases}$$

That is, for every $x \in \mathcal{X}, y \in \{0,1\}$, we have $h_S^{\max}(x) = y$ if and only if $\{g \in \mathcal{G}_S \mid h_g(x) = y\} \in \widetilde{\mathcal{F}}$. The ultra-filter $\widetilde{\mathcal{F}}$ serves to define a tie-breaker label for all domain elements x that are not "eventually" assigned the same label - x's for which every final segment of $\mathcal{G}_S$ contains both g's with $h_g(x) = 0$ and g's with $h_g(x) = 1$.

It remains to show that this so-defined limit function $h_S^{\max}(x)$ has empirical strategic loss 0 on S for all graphs in $\mathcal{G}_S$ and that adding all such limit functions to the hypothesis class $\mathcal{H}$ will not increase its VC-dimension.

For the first property, consider some $g \in \mathcal{G}_S$ and some $(x,y) \in S$. If $y = 0$, then all functions $h_{g'}$ for $g \preceq g'$ assign 0 to all points in $B_g(x)$ (since each $h_{g'}$ is an empirical risk minimizer for the empirical strategic loss $\mathcal{L}_S^{g'}$). If $y = 1$, then all these functions assign label 1 to x. Since the set $\{g' \in \mathcal{G}_S \mid g \preceq g'\}$ is a member of $\mathcal{F}$ and therefore a member of $\widetilde{\mathcal{F}}$, the first case implies that $h_S^{\max}(x') = 0$ for all $x' \in B_g(x)$ and the second case implies that $h_S^{\max}(x) = 1$. In both cases $\ell^g(h_S^{\max}, (x,y)) = 0$, and since this holds for all $(x,y) \in S$ and all $g \in \mathcal{G}_S$, we have

$$\mathcal{L}_S^g(h_S^{\max}(x)) = 0.$$

Now we consider the extended hypothesis class

$$\mathcal{H}' = \mathcal{H} \cup \{h_S^{\max} \in \{0,1\}^{\mathcal{X}} \mid S \in (\mathcal{X} \times \{0,1\})^*\}$$

where we added the limit functions (as defined above) for all possible labeled samples S over $\mathcal{X} \times \{0,1\}$. In order to show that $\mathcal{H}'$ has the same VC-dimension as $\mathcal{H}$, we argue that any finite set that is shattered by $\mathcal{H}'$ is already shattered by $\mathcal{H}$.

Consider domains points $x_1, x_2, \ldots, x_d$, shattered by $\mathcal{H}'$, and some labels $y_1, y_2, \ldots y_d$. Assume this labeling is realized by a limit function $h_S^{\max}$ (that came from some labeled sample S), that is $h_S^{\max}(x_i) = y_i$ for all $i \in [d]$. Note that $h_S^{\max}(x_i) = y_i$ implies that for each i the set

$$\{g \in \mathcal{G}_S \mid h_g(x_i) = y_i\} \in \widetilde{\mathcal{F}}$$

Since ultra-filters are closed under finite intersections, the set

$$\{g \in \mathcal{G}_S \mid h_g(x_i) = y_i \text{ for all } i \in [d]\}$$

is also a member of the ultrafilter $\widetilde{\mathcal{F}}$, and therefore not empty. Thus, there exists a $g \in \mathcal{G}_S$ and corresponding ERM function $h_g \in \mathcal{H}_S \subseteq \mathcal{H}$ with $h_g(x_i) = y_i$ for all $i \in [d]$. Thus, any labeling on a finite set of points that is realized by some limit function, is also already realized by a hypothesis from $\mathcal{H}$. Therefore, $\mathrm{VC}(\mathcal{H}') = \mathrm{VC}(\mathcal{H})$, and we can use the larger class $\mathcal{H}'$ to define the maximally robust ERM hypothesis, we set

- $\mathrm{MG}_{\mathcal{H},\mathcal{G}}(S) = (\mathcal{X}, E)$ with $E = \bigcup_{\{g \in \mathcal{G} : \min_{h \in \mathcal{H}} \mathcal{L}_S^g(h) = 0\}} E_g$
- $\mathrm{MRERM}_{\mathcal{H},\mathcal{G}}(S) = \begin{cases} h \in \operatorname{argmin}_{h \in \mathcal{H}} \mathcal{L}_S^{\mathrm{MG}_{\mathcal{H},\mathcal{G}}(S)}(h) \text{ if this set is not empty} \\ h_S^{\max} \text{ as defined above, otherwise.} \end{cases}$

B. Proofs

Definition B.1. Let $\ell^{\text{str}} : \mathcal{G}_{\text{all}} \times \mathcal{F} \times \mathcal{X} \times \{0,1\}$ be the loss defined by

$$\ell^{\text{str}}(g,h,x,y) = \max\{\ell^{0/1}(h,x,y), \ell^{\text{mani}}(g,h,x)\}.$$

The loss class of $\mathcal{G} \times \mathcal{H}$ with respect to ℓ^{str} is defined by

$$(\mathcal{G} \times \mathcal{H})_{\ell^{\text{str}}} =$$

$$\{\{(x,y) \in \mathcal{X} \times \{0,1\} : \ell^{\text{str}}(g,h,x,y) = 1\}\} : h \in \mathcal{H}, g \in \mathcal{G}\}.$$

Claim B.2. *Let* $d = \text{VC}((\mathcal{G} \times \mathcal{H})_{\ell^{\text{mani}}}) + \text{VC}(\mathcal{H})$. *Then,* $\text{VC}((\mathcal{G} \times \mathcal{H})_{\ell^{\text{str}}}) \leq d\log(d)$.

Proof. We follow the same argument as in (Lechner & Urner, 2022). Let us denote $(g,h)_{\ell^{\text{mani}}} = \{(x,y) \in \mathcal{X} \times \{0,1\} : \ell^{\text{mani}}(g,h,x) = 1\}$ and $h_{\ell^{0/1}} = \{(x,y) \in \mathcal{X} \times \{0,1\} : \ell^{0/1}(h,x,y) = 1\}$. Lastly, let $(g,h)_{\ell^{\text{str}}} = \{(x,y) \in \mathcal{X} \times \{0,1\} : \ell^{\text{str}}(g,h,x,y) = 1\}$. We can easily see that $(g,h)_{\ell^{\text{str}}}(g,h) = (g,h)_{\ell^{\text{mani}}} \cup h_{\ell^{0/1}}$. Thus, $(\mathcal{G} \times \mathcal{H})_{\ell^{\text{str}}} = \{A \cup B : A \in \mathcal{H}_{\ell^{0/1}}, B \in (\mathcal{G} \times \mathcal{H})_{\ell^{\text{mani}}}\}$. Thus, by standard arguments about VC-classes (e.g. exercises in (Shalev-Shwartz & Ben-David, 2014)), we get the claimed result. □

B.1. Proper Batch Learning

B.1.1. Class of Thresholds

Observation 3.1. The Strategic Batch-Learning for Thresholds Algorithm (Algorithm 3.1) is a proper learner for the strategic-batch learning problem for $\mathcal{H}_{\text{thres}}$ and $\mathcal{G}_{\text{f.r.}}$ in the realizable case with sample complexity $m^{\text{real}}_{\mathcal{H},\mathcal{G}} = O(\frac{\log(\frac{1}{\delta})}{\epsilon^2})$ and round complexity $T^{\text{real}}_{\mathcal{H},\mathcal{G}} = 1$.

Proof. This is a special case of Observation 3.10, as the graphs in $\mathcal{G}_{\text{f.r.}}$ are totally ordered with respect to $\mathcal{F}$ (and thus also with respect to $\mathcal{H}_{\text{thres}}$. □

Theorem 3.2. The Strategic Batch-Learning for Thresholds Algorithm (Algorithm 1) is a proper learner for the strategic-batch learning problem for $\mathcal{H}_{\text{thres}}$ and $\mathcal{G}_{\text{f.r.}}$ in the agnostic case with sample complexity $m_{\mathcal{H},\mathcal{G}}(\epsilon,\delta) = O(\frac{\log(\frac{1}{\delta\epsilon})}{\epsilon^2})$ and round complexity $T_{\mathcal{H},\mathcal{G}}(\epsilon,\delta) = O(\log(\frac{1}{\epsilon}))$.

Proof. We note that the graphs in $\mathcal{G}_{\text{f.r.}}$ are totally ordered with respect to $\mathcal{F}$ (and thus also with respect to $\mathcal{H}_{\text{thres}}$. We also find that $(\mathcal{G}_{\text{f.r.}} \times \mathcal{H}_{\text{thres}})_{\ell^{\text{mani}}})$ is the class of intervals over the real line and thus $C \in \text{VC}((\mathcal{G}_{\text{f.r.}} \times \mathcal{H}_{\text{thres}})_{\ell^{\text{mani}}}) = 2$. Therefore we can treat this Theorem as a special case of Theorem 3.11 (and refer the reader to the proof of that theorem). □

B.1.2. General VC-Classes

Observation 3.4

- The VC-dimension of $(\mathcal{G}_{\text{f.r.}} \times \mathcal{H}_{\text{thres}})_{\ell^{\text{mani}}}$ is 2.
- Let $\mathcal{H}_{half} = \{h_w : w \in \mathbb{R}^d : h_w(x) \text{ iff } x^T w \geq 0\}$ the hypothesis class of linear half spaces and $\mathcal{G}^d_{\text{f.r.}} = \{g_r = (\mathcal{X}, E_{g_r}) : (x,x') \in E_{g_r} \text{ iff } ||x - x'||_2 \leq r\}$ be the class of fixed-radius balls in $\mathbb{R}^d$. Then the VC dimension of $(\mathcal{G}^d_{\text{f.r.}} \times \mathcal{H}_{half})_{\ell^{\text{mani}}}$ is at most $2d$.

Proof.
- We note that the elements of $(\mathcal{G}_{\text{f.r.}} \times \mathcal{H}_{\text{thres}})_{\ell^{\text{mani}}}$ correspond exactly to the class of intervals over the real line. The VC-dimension of that class is known to be 2 (Shalev-Shwartz & Ben-David, 2014).
- We note, that the elements of $(\mathcal{G}^d_{\text{f.r.}} \times \mathcal{H}_{half})_{\ell^{\text{mani}}}$ are sets $C_{w,r} = \{x \in \mathbb{R}^d : -r \leq x^T w \leq 0\}$. Now if we take any set C of $2d+1$ points, then there is one point $x \in C$ such that x is in the convex hull of the remaining points, i.e., $x \in conv(C \setminus \{x\})$. Now it is impossible for any $C_{w,r}$ to achieve $C_{w,r} \cap C = \{x\}$. Thus the set C cannot be shattered by $(\mathcal{G}^d_{\text{f.r.}} \times \mathcal{H}_{half})_{\ell^{\text{mani}}}$. Therefore $\text{VC}((\mathcal{G}^d_{\text{f.r.}} \times \mathcal{H}_{half})_{\ell^{\text{mani}}}) \leq 2d$.

□

Lemma 3.6. Let $\mathcal{G}, \mathcal{H}$ be such that $\mathrm{VC}(\mathcal{G}_{\ell^{\mathrm{mani}},\mathcal{H}}) = d$. Then any Empirical Manipulation Estimation learner (such as Algorithm 3.2 below) has the following success guarantee for learning the manipulation graph:

For every marginal distribution $P_{\mathcal{X}}$, every $g^{\rightarrow} \in \mathcal{G}$, every $h \in \mathcal{H}$ and every sequence of $(g^{\rightarrow}, h)$-induced manipulations Π, and every $\epsilon, \delta \in (0,1)$,

if $m \geq C\frac{d+\log(\frac{1}{\delta})}{\epsilon^2}$ (for some universal constant C) with probability $\geq (1-\delta)$, over samples $S_1 \sim P_{\mathcal{X}}^m, S_2 \sim P_{\mathcal{X}}^m$, for every $\hat{g} \in A_{\mathrm{graph}}(S_1, S_2^{\Pi}, h)$,

$$Dis_{(P_{\mathcal{X}},h)}(g^{\rightarrow}, \hat{g}) \leq \epsilon.$$

and

$$g^{\rightarrow} \in A_{\mathrm{graph}}(S_1, S_2^{\Pi}, h).$$

Proof. We note it is sufficient to show that there is a constant C, such that for i.i.d P samples S_1 and S_2 of size $m \geq C\frac{d+\log(\frac{1}{\delta})}{\epsilon^2}$ with probability $1-\delta$, we have $\mathcal{L}_{S_1}^{\mathrm{mani}} \in [L_{max} - \frac{\epsilon}{2}, L_{\max}]$, to guarantee that if we run A_{graph} with parameter $\frac{\epsilon}{2}$, we get:

- $Dis_{(P_{\mathcal{X}},h)}(g^{\rightarrow}, g) \leq \epsilon$ for all $g \in A_{\mathrm{graph}}(S_1, S_2^{\Pi}, h)$.
- and $g^{\rightarrow} \in A_{\mathrm{graph}}(S_1, S_2^{\Pi}, h)$.

First we note, that for any samples S_1 and S_2 and the true manipulation graph $g^{\rightarrow}$, we have $\mathcal{L}_{S_2^{\Pi}}^{\mathrm{mani}}(g^{\rightarrow}, h) = 0$, as the sequence of mappings Π replaces every sample point which can be manipulated in S_2 (i.e., all sample points with $\ell(g^{\rightarrow}, h) = 1$), with a sample point x' with $h(x') = 1$ (i.e., a sample point with $\ell^{\mathrm{mani}}(g^{\rightarrow}, h, x') = 0$). From $g^{\rightarrow} \in \mathcal{G}$, it follows that we have $g^{\rightarrow} \in \{g \in \mathcal{G} : \mathcal{L}_{S_2^{\Pi}}^{\mathrm{mani}}(g, h) = 0\}$. Then this implies

$$L_{\max} = \max\{\mathcal{L}_{S_1}^{\mathrm{mani}}(g, h) : g \in \mathcal{G}, \mathcal{L}_{S_2^{\Pi}}^{\mathrm{mani}}(g, h) = 0\}$$

$$\geq \mathcal{L}_{S_1}^{\mathrm{mani}}(g^{\rightarrow}, h).$$

Now from VC-theory, we know that there exists a universal constant C, such that a sample size of $m \geq \frac{d+\log(\frac{1}{\delta})}{\epsilon^2}$ guarantees that with probability $1 - \frac{\delta}{2}$ we have that a sample $S \sim P^m$ is $\frac{\epsilon}{8}$-representative with respect to $\mathcal{G}_{\ell^{\mathrm{mani}},h}$.

Now let S_1 and S_2 be $\frac{\epsilon}{8}$-representative with respect to $(\mathcal{G} \times \{h\})_{\ell^{\mathrm{mani}}})$. We want to show that $L_{\max} - \frac{\epsilon}{2} \leq \mathcal{L}_{S_1}^{\mathrm{mani}}(g^{\rightarrow}, h)$. We note that there exists $g_{\max} \in \mathcal{G}$, such that $L_{\max} = \mathcal{L}_{S_1}^{\mathrm{mani}}(g_{\max}, h)$ and $\mathcal{L}_{S_2^{\Pi}}^{\mathrm{mani}}(g_{\max}, h) = 0$.

Because of the $\frac{\epsilon}{8}$-representativeness of both samples, we get for every $g \in \mathcal{G}$, we have

$$|\mathcal{L}_{S_1}^{\mathrm{mani}}(g, h) - \mathcal{L}_{S_2}^{\mathrm{mani}}(g, h)| \leq$$

$$|\mathcal{L}_{S_1}^{\mathrm{mani}}(g, h) - \mathcal{L}_{P_{\mathcal{X}}}^{\mathrm{mani}}(g, h)| + |\mathcal{L}_{S_2}^{\mathrm{mani}}(g, h) - \mathcal{L}_{P_{\mathcal{X}}}^{\mathrm{mani}}(g, h)| \leq \frac{\epsilon}{4}.$$

Furthermore we note, that for any manipulation graph g, $|\mathcal{L}_{S_2^{\Pi}}^{\mathrm{mani}}(g, h) - \mathcal{L}_{S_2}^{\mathrm{mani}}(g, h)| \leq \frac{1}{|S_2|}|\{x \in S_2 : x \notin S_2^{\Pi}\}| = \mathcal{L}_{S_2}^{\mathrm{mani}}(g^{\rightarrow}, h)$.

It now follows that:

$$L_{max} = \mathcal{L}_{S_1}^{\mathrm{mani}}(g_{\max}, h)$$

$$\leq |\mathcal{L}_{S_1}^{\mathrm{mani}}(g_{\max}, h) - \mathcal{L}_{S_2}^{\mathrm{mani}}(g_{\max}, h)|$$

$$+|\mathcal{L}^{\text{mani}}_{S_2}(g_{\max}, h) - \mathcal{L}^{\text{mani}}_{S_2^{\Pi}}(g_{\max}, h)|$$

$$\leq \frac{\epsilon}{4} + \mathcal{L}^{\text{mani}}_{S_2}(g^{\rightarrow}, h)$$

$$\leq \frac{\epsilon}{2} + \mathcal{L}^{\text{mani}}_{S_1}(g^{\rightarrow}, h).$$

Putting everything together, we have $\mathcal{L}^{\text{mani}}_{S_1}(g^{\rightarrow}, h) \in [L_{\max} - \frac{1}{\epsilon}, L_{\max}]$. This concludes our proof.

□

Observation 3.7
If an un-manipulated sample S_1 is ϵ-representative with respect to $(\mathcal{G} \times \mathcal{H})_{\ell^{\text{mani}}}$, then it can be indefinitely re-used by A_{graph} for any hypothesis $h \in \mathcal{H}$ and any manipulated ϵ-representative samples S_2^{Π}. Thus if $\text{VC}(\mathcal{G} \times \mathcal{H})_{\ell^{\text{mani}}} = d$, then $m \geq C\frac{d+\log(\frac{1}{\delta})}{\epsilon^2}$ (for some universal constant C) , implies that with probability $1-\delta$ any $S_1 \sim P^m$ is repeatedly reusable by A_{graph} to guarantee ϵ-success as in the Lemma above, This allows us to reuse the initial unmanipulated sample in all subsequent steps.

Proof. We note that in order for sample S_1 to guarantee success in Lemma 1 we only required it to be $\frac{\epsilon}{4}$-representative with respect to $\mathcal{G}_{\ell^{\text{mani}},h}$. If $\text{VC}((\mathcal{H} \times \mathcal{G})_{\ell^{\text{mani}}}) = d$, then there exists a sample size $m \geq C\frac{d+\log(\frac{1}{\delta})}{\epsilon^2}$ that with probability $1-\delta$ over the sample generation $S_1 \sim P^m$, we have that S_1 is $\frac{\epsilon}{4}$-representative with respect to $(\mathcal{G} \times \mathcal{H})_{\ell^{\text{mani}}}$. This implies that with probability $1-\delta$, S_1 is $\frac{\epsilon}{4}$-representative w.r.t. $\mathcal{G}_{\ell^{\text{mani}},h}$ for all $h \in \mathcal{H}$ simultaneously, proving the obervation.

□

Observation 3.9.

- A graph class $\mathcal{G}$ is totally ordered with respect to the class of all hypotheses $\mathcal{F}$ if and only if for every distinct $g_1, g_2 \in \mathcal{G}$ either g_1 is a subgraph of g_2 or g_2 is a subgraph of g_1.
- For $\mathcal{H}_1 \subset \mathcal{H}_2$ and two graphs g_1, g_2 $g_1 \preceq_{\mathcal{H}_2} g_2$ implies $g_1 \preceq_{\mathcal{H}_1} g_2$. Thus, if a graph class $\mathcal{G}$ is totally ordered with respect to $\mathcal{H}_2$ it is also totally ordered with respect to $\mathcal{H}_1$.
- If $\mathcal{G}$ is totally ordered with respect to $\mathcal{H}$, then $\text{VC}(\mathcal{G}_{\ell^{\text{mani}},\mathcal{H}}) \leq 1$.
- There are $\mathcal{G}$ and $\mathcal{H}$, such that $\text{VC}(\mathcal{H}) = d$ and $\mathcal{G}$ is totally ordered with respect to $\mathcal{H}$, but $\text{VC}((\mathcal{H} \times \mathcal{G})_{\ell^{\text{mani}}}) = \infty$.

Proof.
- "$\rightarrow$": Let $\mathcal{G}$ be totally ordered with respect to $\mathcal{F}$. This means for any two graphs $g_1, g_2 \in \mathcal{G}$, we have either $g_1 \preceq_{\mathcal{F}} g_2$ or $g_2 \preceq_{\mathcal{F}} g_1$. Without loss of generality, assume $g_1 \preceq_{\mathcal{F}} g_2$. We want to show that this implies that g_1 is a subgraph of g_2. For the purpose of contradiction, assume the opposite. This means that there exists an edge $(x, x') \in V_{g_1}$, such that $(x, x') \notin V_{g_2}$. Now consider any function $f \in \mathcal{F}$ with $f(x) = 0$ and $f(x) = 1$. Then $\ell^{\text{mani}}(g_1, f, x) = 1 \geq 0 \geq \ell^{\text{mani}}(g_2, f, x)$, which contradicts $g_1 \preceq_{\mathcal{F}} g_2$.

  "$\leftarrow$" Let $\mathcal{G}$ such that for every two graphs $g_1, g_2 \in \mathcal{G}$, we have either $E_{g_1} \subset E_{g_2}$ or $E_{g_2} \subset E_{g_1}$. Without loss of generality, assume $E_{g_1} \subset E_{g_2}$. We want to show that this implies $g_1 \preceq_{\mathcal{F}} g_2$. For the sake of contradiction, we assume the opposite. This means that there exists an $f \in \mathcal{F}$ and a $x \in \mathcal{X}$, such that $\ell^{\text{mani}}(g_1, f, x) = 1$ and $\ell^{\text{mani}}(g_2, f, x) = 0$. $\ell^{\text{mani}}(g_1, f, x) = 1$ implies that $f(x) = 0$ and that there is an $x' \in \mathcal{X}$ such that $f(x') = 1$ and such that $(x, x') \in E_{g_1}$. Now this implies that $(x, x') \in E_{g_2}$ as well. From $f(x) = 0$ and $f(x') = 1$ it follows that $\ell^{\text{mani}}(g_2, f, x) = 1$, contradicting our assumption.

- This follows directly from the definitions. If for g_1 and g_2, we have $g_1 \preceq_{\mathcal{H}_2} g_2$ then for all $h \in \mathcal{H}_2$ and all $x \in \mathcal{X}$, we have $\ell^{\text{mani}}(g_1, h, x) = 1$ implies $\ell^{\text{mani}}(g_2, h, x) = 1$. Since $\mathcal{H}_1 \subset \mathcal{H}_2$ for all $h \in \mathcal{H}_1$ we thus have $\ell^{\text{mani}}(g_1, h, x) = 1$ implies $\ell^{\text{mani}}(g_2, h, x) = 1$. Thus $g_1 \preceq_{\mathcal{H}_1} g_2$.

- Take any $h \in \mathcal{H}$ and any $\{x_1, x_2\} = C \subset \mathcal{X}$. Assume that C was shattered by $\mathcal{G}_{\ell^{\text{mani}},h}$. Then there is some g_1, such that $\ell^{\text{mani}}(g_1, h, x_1) = 1$ and $\ell^{\text{mani}}(g_1, h, x_2) = 0$. We know that $\mathcal{G}$ is totally ordered with respect to $\mathcal{H}$. Thus we know that for any $g_2 \in \mathcal{G}$, we either have $g_1 \preceq_H g_2$, which implies that $\ell^{\text{mani}}(g_2, h, x_1) = 1$ or $g_2 \preceq_{\mathcal{H}} g_1$ which implies $\ell^{\text{mani}}(g_2, h, x_2) = 0$. Thus there is no $g_2 \in \mathcal{G}$ with $\ell^{\text{mani}}(g_2, h, x_1) = 0$ and $\ell^{\text{mani}}(g_2, h, x_2) = 1$, contradicting that C is shattered by $\mathcal{G}_{\ell^{\text{mani}},h}$. Thus $\text{VC}(\mathcal{G}_{\ell^{\text{mani}},h}) \leq 1$. Since $h \in \mathcal{H}$ was picked arbitrarily we have $\text{VC}((\mathcal{G})_{\ell^{\text{mani}},\mathcal{H}}) = \sup_{h\in\mathcal{H}}(\text{VC}(\mathcal{G}_{\ell^{\text{mani}},h})) \leq 1$.
- For $\text{VC}((\mathcal{H}\times\mathcal{G})_{\ell^{\text{mani}}})$ to be infinite it is sufficient to find one g and a hypothesis class $\mathcal{H}$ such that $\text{VC}(\mathcal{H}\times\{g\})_{\ell^{\text{mani}}} = \infty$. In Theorem 5 in (Lechner & Urner, 2022), we have seen that there are classes $\mathcal{H}$ with $\text{VC}(\mathcal{H}) = 1$ and manipulation graphs g, such that $\text{VC}(\mathcal{H} \times \{g\})_{\ell^{\text{mani}}} = \infty$. Lastly we note that if we pick $\mathcal{G} = \{g\}$, then $\mathcal{G}$ is trivially totally ordered with respect to $\mathcal{H}$.

□

Observation 3.10. Let $\text{VC}(\mathcal{H} \times \mathcal{G})_{\ell^{\text{mani}}} = d_1$ and $\text{VC}(\mathcal{H}) = d_2$. Furthermore let $\mathcal{G}$ be totally ordered with respect to $\mathcal{H}$. Then Algorithm 3.2 is a successful proper strategic batch learner in the realizable case with sample complexity $m^{\text{real}}_{\mathcal{H},\mathcal{G}}(\epsilon, \delta) = O(\frac{(d_1+d_2)\log(d_1+d_2)+\log(\frac{1}{\delta})}{\epsilon^2})$ and round complexity $T^{\text{real}}_{\mathcal{H},\mathcal{G}}(\epsilon, \delta) = 1$.

Proof. We assume to be in the realizable case, i.e. $\inf_{h\in\mathcal{H}} L_P^{g^\rightarrow}(h) = 0$. Thus with probability 1, the sample S_0 will be realizable under $L^{g^\rightarrow}$ as well. This implies $\inf_{h\in\mathcal{H}} L^{0/1}_{S_0}(h) = 0$. Therefore, in line 3 of Algorithm 3, we set $L_0^{0/1}$ to 0. We then determine g_0 to be the maximal graph according to $\preceq_{\mathcal{H}}$ to yield $\inf_{h\in H} L^{g_0}_{S_0}(h) = 0$ (line 5). From this and the realizability assumption, if follows that $g^\rightarrow \preceq_{\mathcal{H}} g_0$. In line 8, we now define $\hat{h}^0$, to be $\arg\min_{h\in\mathcal{H}} L^{g_0}_{S_0}(h)$. From $g^\rightarrow \preceq_{\mathcal{H}} g_0$, it follows that $L^{g^\rightarrow}_{S_0}(\hat{h}^0) = 0$ as well. Now, we can choose a sample size $m \in O(\frac{(d_1+d_2)\log(d_1+d_2)+\log(\frac{1}{\delta})}{\epsilon^2})$, large enough to guarantee $\frac{\epsilon}{8}$-representativeness with respect to $\mathcal{H}_{\ell g^\rightarrow}$, $(\mathcal{G}\times\mathcal{H})_{\ell^{\text{mani}}}$ as well as $(\mathcal{G}\times\mathcal{H})_{\ell^{\text{str}}}$ simultaneously with probability at least $1-\frac{\delta}{2}$. Now with probability $1-\delta$, $S_0 \sim P^m$ and $S_1' \sim P^m$ are both such $\frac{\epsilon}{8}$-representative samples. First we note, that this guarantees that $L_P^{g^\rightarrow}(\hat{h}_0) \leq \frac{\epsilon}{8}$ with probability $1-\frac{\delta}{2}$.

This implies that an $\frac{\epsilon}{8}$-representative sample S_1' at most a fraction of $\frac{\epsilon}{4}$ can be manipulated by a $(g^\rightarrow, \hat{h}_0)$-induced manipulation. We now assume that the parameter ϵ in the algorithm, is the same as the ϵ in our sample complexity analysis (our algorithm lets us set this parameter. We can assume that we know in advance what the size of the first sample will be.). Now if S_1' is $\frac{\epsilon}{8}$-representative with respect to $\ell^{g^\rightarrow}$, then for g_1 (as chosen in line 5), we have $L^{g_1}_{S_1'}(\hat{h}_0) \geq L^{g_1}_{S_0}(\hat{h}_0) - \frac{\epsilon}{4} \geq \inf_{h\in\mathcal{H}} L^{g_1}_{S_0}(\hat{h}_0) - \frac{\epsilon}{4} = \frac{3\epsilon}{4}$. Since $L^{0/1}_{S_0}(\hat{h}_0) = 0$, we have $L^{0/1}_{S_1'}(\hat{h}_0) \leq \frac{\epsilon}{4}$. Thus $\mathcal{L}^{\text{mani}}_{S_1'}(g_1, \hat{h}^0) \geq \frac{3\epsilon}{4} - \frac{\epsilon}{4} = \frac{\epsilon}{2}$. Now we noticed before that only at most a fraction of $\frac{\epsilon}{4}$ samples in S_1' get replaced in S_1. Thus g_1 does **not** fulfill $\mathcal{L}^{\text{mani}\,g_1}_{S_1'}(\hat{h}_0) = 0$. Therefore $G_1 = \{g_0\}$. Which means that for all consecutive rounds t, we get $\hat{h}^t = \inf_{h\in\mathcal{H}} L^{g_0}_{S_0}(h)$, which yields the guarantee $L_P^{g^\rightarrow}(\hat{h}^t) \leq L_P^{g_0}(\hat{h}^t) \leq \frac{\epsilon}{8}$.

□

Theorem 3.11.

Let $\text{VC}(\mathcal{H} \times \mathcal{G})_{\ell^{\text{mani}}} = d_1$ and $\text{VC}(\mathcal{H}) = d_2$. Furthermore let $\mathcal{G}$ be totally ordered with respect to $\mathcal{H}$. Then Algorithm 3.2 is a successful proper strategic batch learner with sample complexity $m_{\mathcal{H},\mathcal{G}}(\epsilon, \delta) = O(\frac{(d_1+d_2)(\log(d_1+d_2))+\frac{1}{\delta\epsilon}}{\epsilon^2})$ and round complexity $T_{\mathcal{H},\mathcal{G}}(\epsilon, \delta) = O(\log(\frac{1}{\epsilon}))$.

Proof. We will first argue that among the candidate graphs $\mathcal{G}^0$ we pick in round 0, there is a candidate graph g_i, which yields a close-to-optimal hypothesis h, given some amount of ϵ-representativeness and for certain choices of parameters.

Assume we have a sample S_0 which is $\frac{\epsilon}{16}$-representative with respect to $(\mathcal{G}\times\mathcal{H})_{\ell^{\text{str}}}$. Then S_0 is $\frac{\epsilon}{16}$-representative with respect to $\mathcal{H}_{\ell^g}$ for every $g \in \mathcal{G}$. Now let us run the algorithm with the parameter "ϵ" set as $\frac{\epsilon}{4}$. Let $\mathcal{G}^0$ be defined as in the algorithm (line 7). Furthermore for every $g_i \in \mathcal{G}^0$, define $h_i = \arg\min_{h\in\mathcal{H}} L^{g_i}_{S_0}(h)$ and $\mathcal{H}' = \{h_0, \ldots, h_{\frac{4}{\epsilon}}\}$. Now let h^* denote the optimal hypothesis $h^* = \arg\min_{h\in\mathcal{H}} L_P(g^\rightarrow)(h)$. Now let $j = \lceil \frac{4(L^{g^\rightarrow}_{S_0}(h^*) - L_0^{0/1})}{\epsilon} \rceil + 1$. Now we compare h^* to h_j from the candidate set $\mathcal{H}'$. By definition of g_j and h_j, we know that $\frac{j\epsilon}{4} + L_0^{0/1} = L^{g_j}_{S_0}(h_j) \leq L^{g_j}_{S_0}(h^*)$. Furthermore, we

know that $L_P^{g^\rightarrow}(h^*) \leq L_{S_0}^{g^\rightarrow}(h^*) + \frac{\epsilon}{16} \leq (\frac{(j-1)\epsilon}{4} + L_0^{0/1}) + \frac{\epsilon}{16} < \frac{j\epsilon}{4} + L_0^{0/1}$. Thus $g^\rightarrow \preceq_{\mathcal{H}_{\text{thres}}} g^j$. Therefore, $L_P^{g^\rightarrow}(h_j) \leq L_P^{g^j}(h_j) \leq L_{S_0}^{g^j}(h_j) + \frac{\epsilon}{16} \leq L_{S_0}^{g^j}(h^*) + \frac{\epsilon}{16} = \frac{j\epsilon}{4} + L_0^{0/1} + \frac{\epsilon}{16} = \frac{(j-2)\epsilon)}{4} + L_0^{0/1} + \frac{9\epsilon}{16} \leq L_{S_0}^{g^\rightarrow}(h^*) + \frac{9\epsilon}{16} \leq L_P^{g^\rightarrow}(h^*) + \frac{10\epsilon}{16}$. This shows that there is indeed a hypothesis in our candidate-set that is close to optimal.

We now note that the graph elimination that happens in lines 13 to 15 and lines 20 to 21 is equivalent to the estimation of Algorithm 2, which we know to keep the optimal graph if the samples encountered are $\frac{\epsilon'}{8}$-representative. We further note that the update in line 23 only occurs, if the sample $S_t = S_t'^{\Pi}$ was not significantly manipulated according to $\hat{h}^{t-1}$. Thus g^{t-1} already sufficiently accounted for all possible manipulation caused by $g^\rightarrow$, which means that eliminating candidate graphs g with $g^{t-1} \preceq_{\mathcal{H}} g$ will not cause a miss-estimation of $g^\rightarrow$ that causes more that $2\epsilon'$ difference in strategic loss. Thus the elimination of graphs will yield an approximately optimal hypothesis. Furthermore this is a kind of binary search which eliminates half the candidates in each step (as the algorithm always picks the median candidate in line 15 and line 26 respectively and the elimination either eliminates all graphs smaller or all graphs greater to the current candidate.) Therefore the algorithm needs at most $O(\log(\frac{1}{\epsilon}))$ rounds.

Lastly, we note that a sample size of $m = O(\frac{(d_1+d_2)\log(d_1+d_2)+\log(\frac{\log(\frac{1}{\epsilon})}{\delta})}{\epsilon^2})$ is sufficient to guarantee that in each of the $O(\log(\frac{1}{\epsilon}))$ rounds, the probability of receiving an $\frac{\epsilon}{16}$-representative sample is at least $1 - \frac{\delta}{\log(\frac{1}{\epsilon})}$. Thus via union bound the sample size of $m = O(\frac{(d_1+d_2)\log(d_1+d_2)+\log(\frac{\log(\frac{1}{\epsilon})}{\delta})}{\epsilon^2})$ is sufficient to guarantee (ϵ, δ)-learning success. □

B.2. Improper Learning

Observation 4.1. There are classes $\mathcal{H}$ of finite VC-dimension and graph classes $\mathcal{G}$ which are totally ordered with respect to $\mathcal{H}$, such that there is no proper successful batch-learner for $\mathcal{H}$ and $\mathcal{G}$ for any finite sample and round complexity, not even in the realizable case.

Proof. This follows directly from Theorem 4 of (Lechner & Urner, 2022) (which is an adaptation of Theorems 1 and 4 of (Montasser et al., 2019) for the strategic loss). The theorem states that there exists a class $\mathcal{H}$ and a fixed manipulation graph g of VC-dimension 1 which cannot be properly PAC-learned with respect to strategic loss by any proper learner. If we now consider this $\mathcal{H}$ and the graph class $\mathcal{G} = \{g\}$, then it is easy to see that $\mathcal{G}$ is totally ordered with respect to $\mathcal{H}$ (as it only has one element). Furthermore $\mathcal{H}$ was picked to have VC dimension 1. Furthermore if proper batch-learning was possible for $\mathcal{H}$ and $\mathcal{G}$, then proper PAC-learning would be possible for $\mathcal{H}$ with respect to ℓ^g, which we know to be impossible. This proves the theorem. □

Theorem 4.2. Let $\text{VC}(\mathcal{H})$ be finite and let $\mathcal{G}$ be totally ordered with respect to $\mathcal{H}$. Then there is a strategically robust (improper) PAC-learner which is successful for every $g^\rightarrow \in \mathcal{G}$ in the strategically robustly realizable case (i.e. when $\inf_{h\in\mathcal{H}} L_P^{g^\rightarrow}(h) = 0$).

Proof. This is an adaptation of Theorem 4 from (Montasser et al., 2019) (and its adaptation to strategic loss in (Lechner & Urner, 2022)). The main difference here is that those theorems focus on the robustness with respect to a fixed (and known) manipulation graph, whereas in our setting we want to guarantee robustness with respect to an unknown element of a totally ordered graph class $\mathcal{G}$. Thus we cannot use any knowledge of this graph in the learning process, which makes it impossible to use robust empirical risk minimization (RERM) with respect to the true manipulation graph. However, we can use the realizability assumption and define Maximally Robust Empirical Risk Minimization(MRERM) with respect to a totally ordered graph class $\mathcal{G}$. First let us define

$$g_{\max}(S, \mathcal{H}, \mathcal{G}) =$$
$$\max_{\preceq_{\mathcal{H}}}\{g \in \mathcal{G} : \text{ for all } h \in RERM_{g,\mathcal{H}}(S) : L_S^g(h) = 0\}.$$

The set of maximally robust empirical risk minimizers with respect to $\mathcal{G}$ and $\mathcal{H}$ is then defined by $MRERM_{\mathcal{G},\mathcal{H}}(S) = RERM_{g_{max}(S,\mathcal{H}),\mathcal{H}}(S)$. We can now replace all use of the fixed deterministic $RERM_{g,\mathcal{H}}$ algorithm for a fixed manipulation graph (or pertubation sets) in the proof of Theorem 4 of (Montasser et al., 2019) by a fixed deterministic $MRERM_{\mathcal{G},\mathcal{H}}$-algorithm.

We note that for any $S' \subset S$, we have $g_{\max}(S, \mathcal{H}, \mathcal{G}) \preceq_{\mathcal{H}} g_{\max}(S', \mathcal{H}, \mathcal{G})$. Furthermore, we note that under the realizability assumption, we have $g_{\max}(S, \mathcal{H}, \mathcal{G}) \geq g^\rightarrow$ and thus $L_S^{g^\rightarrow}(MRERM_{\mathcal{G},\mathcal{H}}(S)) = 0$ with probability 1 over the sample

generation. We can now follow the proof of (Montasser et al., 2019) with some modified definitions, mainly replacing RERM with MRERM and replacing the inflated sample according to the true manipulation graph (which we don't know) with an inflated sample according to the $g_{max}(S, \mathcal{H}, \mathcal{G})$: For a training sample $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$, we can now define the inflated sample according to the maximal graph that still allows realizability according to $\mathcal{H}$: Let the inflated (potentially infinite) sample $S_{g_{\max}}$ be defined as $S_{g_{\max}} = S \cup (\bigcup_{i \in \{1,\ldots,m\}: y_i = 0} \{(x, 0) : x \in B_{g_{\max}(S,\mathcal{H},\mathcal{G})}(x_i)\})$. We now want to define a discretized version of this inflated sample. From standard PAC-learning theory we know that there is a positive integer $n \in O(\mathrm{VC}(\mathcal{H}))$ that guarantees for any distribution D over $\mathcal{X} \times \{0, 1\}$ with $\inf_{h \in \mathcal{H}} L_D^{0/1}(h) = 0$, for n i.i.d. D-distributed samples $S' = \{(x'_1, y'_1), \ldots, (x'_n, y'_n)\}$, with nonzero probability, every $h \in \mathcal{H}$ satisfying $L_{S'}^{0/1}(h) = 0$ also $L_D(h) \leq \frac{1}{3}$. Now let $\hat{\mathcal{H}} = \{MRERM_{\mathcal{G},\mathcal{H}}(L) : L \subset S \text{ and } |L| = n\}$. We note that $|\hat{\mathcal{H}}| \leq |\{MRERM_{\mathcal{G},\mathcal{H}}(L) : L \subset S \text{ and } |L| = n\}| \leq (\frac{em}{n})^n$. Now consider the dual space $\mathcal{W}$ of function $w_{(x,y)} : \mathcal{H} \to \{0, 1\}$ defined by $w_{(x,y)}(h) = \mathbf{1}[[] h(x) \neq y]$ and every $(x, y) \in S_{g_{\max}}$. The VC-dimension of $\mathcal{W}$ is now at most the dual VC-dimension of $\mathcal{H}$, which is known to be upper-bounded by $\mathrm{VC}^* \leq 2^{\mathrm{VC}(\mathcal{H})+1}$. We now define $\hat{S}_{g_{\max}}$ to be a subset of $S_{g_{\max}}$ which includes exactly one element $(x, y) \in \S_{g_{\max}}$ for each distinct classification $\{w_{(x,y)}(h)\}_{hin\hat{\mathcal{H}}}$ of $\hat{\mathcal{H}}$ realized function of $w_{(x,y)} \in \mathcal{W}$. By the Sauer lemma we have $|\hat{S}_{g_{\max}}| \leq (\frac{e|\hat{\mathcal{H}}|}{\mathrm{VC}^*(\mathcal{H})})^{\mathrm{VC}^*(\mathcal{H})}$, which for $m > 2\mathrm{VC}(\mathcal{H})$ is at most $(\frac{e^2 m}{\mathrm{VC}(\mathcal{H})})^{\mathrm{VC}(\mathcal{H})\mathrm{VC}^*(\mathcal{H})}$. We can now note that for any $T \in \mathbb{N}$ and any $h_1, \ldots, h_T \in \hat{\mathcal{H}}$ if $\frac{1}{T}\sum_{t=1}^{T} \mathbf{1}[h_t(x) = y] > \frac{1}{2}$ for every $(x, y) \in \hat{S}_{g_{\max}}$, then $\frac{1}{T}\sum_{t=1}^{T} \mathbf{1}[h_t(x) = y] > \frac{1}{2}$ for every $(x, y) \in S_{g_{\max}}$ as well, which would then imply $L_S^{g_{\max}(S,\mathcal{G},\mathcal{H})}(\text{Majority}(h_1, \ldots, h_T)) = 0$, which implies $L_S^{g^{\rightarrow}}(\text{Majority}(h_1, \ldots, h_T)) = 0$. We can now find these functions $h_1, \ldots, h_T$ in exactly the same way as in (Montasser et al., 2019)(via using the α-Boost algorithm). The resulting classifier $\hat{h} = \text{Majority}(h_1, \ldots, h_T)$ satisfies $L_S^{g^{\rightarrow}}(\hat{h}) = 0$. Furthermore we note that each of the classifiers h_t is the result of $MRERM_{\mathcal{G},\mathcal{H}}(L_t)$ for some $L_t \subset S$ with $|L_t| = n$. Thus, the classifier $\hat{h}$ is representable as the value of an (order-dependent) reconstruction function of set size

$$nT = O(\mathrm{VC}(\mathcal{H}) \log(|\hat{S}_{g_{\max}}|))$$

$$= O(\mathrm{VC}(\mathcal{H})^2 \mathrm{VC}^*(\mathcal{H}) \log(\frac{m}{\mathrm{VC}(\mathcal{H})}).$$

Thus invoking Lemma 11 of (Montasser et al., 2019)[1] with respect to $L^{g^{\rightarrow}}$ if $m > c\mathrm{VC}(\mathcal{H})^2 \mathrm{VC}^*(\mathcal{H}) \log(\mathrm{VC}(\mathcal{H})\mathrm{VC}^*(\mathcal{H}))$ (for a sufficiently large numerical constant c), we have that with probability at least $1 - \delta$,

$$L_P^{g^{\rightarrow}}(\hat{h}) \leq$$

$$O(\mathrm{VC}(H)^2 \mathrm{VC}^*(\mathcal{H})) \log(\frac{m}{\mathrm{VC}(\mathcal{H})}) \log(m) + \frac{1}{m} \log(\frac{1}{\delta}).$$

This concludes our proof. □

[1]We need a slight adaptation from adversarial loss to strategic loss here, but the proof for this goes through exactly as is for strategic manipulation loss as argued in (Lechner & Urner, 2022).