Incentive-Aware PAC Learning

Hanrui Zhang[1] Vincent Conitzer[1]

Abstract

We study PAC learning in the presence of strategic manipulation, where data points may modify their features in certain predefined ways in order to receive a better outcome. We show that the vanilla ERM principle fails to achieve any nontrivial guarantee in this context. Instead, we propose an incentive-aware version of the ERM principle which has asymptotically optimal sample complexity. We then focus our attention on incentive-compatible classifiers, which provably prevent any kind of strategic manipulation. We give a sample complexity bound that is, curiously, independent of the hypothesis class, for the ERM principle restricted to incentive-compatible classifiers. This suggests that incentive compatibility alone can act as an effective means of regularization. We further show that it is without loss of generality to consider only incentive-compatible classifiers when opportunities for strategic manipulation satisfy a transitivity condition. As a consequence, in such cases, our hypothesis-class-independent sample complexity bound applies even without incentive compatibility. Our results set the foundations of incentive-aware PAC learning.

1. Introduction

Consider the following scenario. A financial institution plans to offer a product to a selected group of customers, and decides that the only thing that should matter in the selection process is the amount of money she currently owns in savings (possibly held at a different financial institution). Each customer may prove her savings balance by submitting a bank statement, and the number shown therein will be used for the selection. However, customers may choose to underreport their balance, by, for example, temporarily transferring their money to another account before the statement date. (We assume for the purpose of this example that overreporting is not possible.) The question is, how does this ability of customers to strategically underreport their balance affect the selection process?

The answer, as one would expect, depends on the specific criteria of the selection, which are presumably determined by the nature of the product. For instance, the product could be a secured loan that requires the customer to use the savings for collateral. In that case, customers will benefit from their balance being as high as possible, which means they would submit a statement with the full balance. As a result, the fact that customers can underreport would not affect the selection at all. Alternatively, the financial institution could be a non-profit entity that aims to make a product available only to those with *low* savings. In that case, every customer who wants access to this product would underreport her balance in order to be approved, and this would make the selection effectively meaningless.

More generally, often the right criteria for the selection may not be clear a priori, but have to be learned from observations. In such cases, the decision maker (e.g., the institution) has access to labeled historical data (e.g., the amount of savings owned by some customer, and whether the product was successful for the same customer). A classifier (e.g., selection criteria) is derived from the data and implemented, to which future data points (e.g., new customers) to be classified respond in a strategic way. The goal, naturally, is to classify future data points as accurately as possible, taking into account that their features may be strategically modified. The key challenge is for the classifier derived to *generalize* from past observations to future strategic data points. Such problems are partially captured by the PAC learning model (discussed in detail below), but also exhibit additional complexity from *strategic manipulation* of the features, *where the nature of this manipulation is affected by the choice of classifier*. In this paper, we investigate novel phenomena in PAC learning introduced by the presence of strategic manipulation.

Our results. Two central questions in PAC learning are the following: statistically, how many labeled data points does one need to observe in order to derive a classifier of a desired quality, and computationally, given this many

[1] Department of Computer Science, Duke University, Durham, USA. Correspondence to: Hanrui Zhang <hrzhang@cs.duke.edu>, Vincent Conitzer <conitzer@cs.duke.edu>.

labeled data points, how can one compute such a classifier? We answer these questions for arbitrary feature spaces and structures of strategic manipulation, by presenting an adapted version of the *empirical risk minimization (ERM)* principle, which roughly says that one should simply pick a classifier that has the best quality on the labeled data points that are observed. We show that our incentive-aware version of the ERM principle requires asymptotically the minimum possible number of labeled data points for any specific problem setup. This can be viewed as a strategic version of the VC theory, one of the central results in traditional PAC learning.

We further consider *incentive-compatible* classifiers, which provably prevent any kind of strategic manipulation by making revealing one's true feature always optimal for any data point. Here, our most remarkable result is a *hypothesis-class-independent* bound on the number of labeled data points required to compute an incentive-compatible classifier of a desired quality. In traditional PAC learning, it is well known that without any prior knowledge about the ground truth (typically modeled by a *hypothesis class* consisting of possible classifiers to be considered), it is impossible to learn anything nontrivial, unless the number of labeled data points observed is trivially large or even infinite. By contrast, we show that in the presence of strategic manipulation, it is possible to learn a nontrivial classifier via our strategic version of the ERM principle without any such prior knowledge, except that the classifier learned has to be incentive-compatible. In other words, incentive-compatibility acts as a means of *regularization*, which provides nontrivial learning guarantees in and of itself.

Moreover, when the structure of strategic manipulation is transitive (i.e., if A can pretend to be B and B can pretend to be C, then A can pretend to be C), considering incentive-compatible classifiers is without loss of generality — this is commonly known as the *revelation principle* in economic theory. This, together with our results for incentive-compatible classifiers, implies that ERM-type learning in the presence of transitive strategic manipulation is *automatically regularized*. That is, the hypothesis-class-independent bound discussed above applies even without any exogenous requirement of incentive compatibility (since requiring this condition is without loss of generality), or any other prior knowledge.

Related work. There is an extremely rich body of research on PAC learning (Valiant, 1984). Since the enlightening result by Vapnik & Chervonenkis (1971), various measures of complexity have been studied (Alon et al., 1997; Bartlett & Mendelson, 2002; Pollard, 2012; Daniely et al., 2015), and tighter sample complexity bounds have been developed (Talagrand, 1994; Hanneke, 2016). See, e.g., (Devroye et al., 2013; Shalev-Shwartz & Ben-David, 2014) for a comprehensive exposition of PAC learning.

A closely related research topic is strategic machine learning, where it is commonly assumed that strategic agents seek to maximize their utility by modifying their features in certain restricted ways, often at some cost (Hardt et al., 2016; Kleinberg & Raghavan, 2019; Haghtalab et al., 2020). Specific topics include strategic aspects of linear regression (Perote & Perote-Pena, 2004; Dekel et al., 2010; Chen et al., 2018) and online learning (Roughgarden & Schrijvers, 2017; Braverman et al., 2019; Feng et al., 2019; Freeman et al., 2020). Our results differ from the above in that we consider a PAC model, where the key challenge is to generalize from observations to future data points to be classified. Also, we consider arbitrary feature spaces and structures of strategic manipulation, while existing results often focus on relatively restrictive setups, e.g., linear models.

Another closely related line of research is mechanism design with partial verification (Green & Laffont, 1986; Yu, 2011; Kephart & Conitzer, 2015; 2016), which is often motivated by similar considerations as strategic machine learning. The general problem considered there is to design an (approximately) optimal mechanism assigning outcomes to data points based on their features, where the authenticity of the features can only be partially verified. Particularly relevant is a series of papers by Zhang et al. with a pronounced learning aspect (Zhang et al., 2019b;a). The main difference between our results and theirs is that in our model, the underlying distribution of data is only accessible via samples observed, and we do not restrict the structure of possible strategic manipulation.

2. Preliminaries

In this section, we review relevant definitions and results in PAC learning, and formulate the notion of strategic manipulation in classification.

2.1. Background on PAC Learning

Our problems of interest fit well with the *probably approximately correct (PAC)* learning model (Valiant, 1984). In PAC learning, there is a feature space $\mathcal{X}$, a label space $\mathcal{Y}$, and a joint population distribution $\mathcal{D} \in \Delta(\mathcal{X} \times \mathcal{Y})$ over the feature space and the label space. For a classifier $f : \mathcal{X} \to \mathcal{Y}$, the population loss $\ell_{\mathcal{D}}(f)$ is defined as the probability that f assigns a wrong label to a random point with respect to $\mathcal{D}$, i.e., $\ell_{\mathcal{D}}(f) = \Pr_{(x,y)\sim\mathcal{D}}[f(x) \neq y]$. The goal of PAC learning, with target relative loss $\varepsilon > 0$ and failure probability $\delta > 0$, is to find a classifier $f \in \mathcal{H}$ from a predetermined hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, by observing $m = m(\varepsilon, \delta)$ iid samples from $\mathcal{D}$, such that with probability at least $1 - \delta$, the population loss of f, $\ell_{\mathcal{D}}(f)$, is at most $\ell_{\mathcal{D}}(\mathcal{H}) + \varepsilon$, where $\ell_{\mathcal{D}}(\mathcal{H})$ is the population loss of the best classifier in $\mathcal{H}$, i.e., $\ell_{\mathcal{D}}(\mathcal{H}) = \min_{f' \in \mathcal{H}} \ell_{\mathcal{D}}(f')$. $\ell_{\mathcal{D}}(\mathcal{H})$ is sometimes called the *approximation error*, which models how well the hypothesis class $\mathcal{H}$ approximates the ground truth classifier. Throughout this paper, we focus on binary labels, i.e., $\mathcal{Y} = \{0, 1\}$.

The two central questions of PAC learning discussed earlier are partially answered by the ERM principle, formally defined below. Given m iid samples $S = \{(x_i, y_i)\}_i \sim \mathcal{D}^m$,[1] the ERM principle finds a classifier $f \in \mathcal{H}$ which minimizes the empirical loss $\ell_S(f)$ on S, defined as

$$\ell_S(f) = \frac{1}{|S|} \sum_{i \in [|S|]} |f(x_i) - y_i|.$$

That is, the ERM principle finds $f \in \operatorname{argmin}_{f' \in \mathcal{H}} \ell_S(f')$. This gives a concrete, though not always efficient, way of computing a classifier — it is known that with sufficiently many samples, the classifier found by the ERM principle achieves the desired loss and failure probability. Moreover, the number of samples (also known as the *sample complexity*) required by the ERM principle is optimal up to a constant factor.[2] This sample complexity is asymptotically determined by the so-called *VC dimension* of the hypothesis class $\mathcal{H}$, defined below.

Definition 1 (VC Dimension)**.** A subset S of the feature space $\mathcal{X}$ is *shattered* by a hypothesis class $\mathcal{H}$, if for any $T \subseteq S$, there is a classifier $c \in \mathcal{H}$ such that $c \cap S = H$. The *VC dimension* of $\mathcal{H}$, $d_{\mathrm{VC}}(\mathcal{H})$, is the cardinality of the largest subset of $\mathcal{X}$ that is shattered by $\mathcal{H}$, i.e.,

$$d_{\mathrm{VC}}(\mathcal{H}) = \max\left\{ |S| \,\middle|\, |\{f \cap S \mid f \in \mathcal{H}\}| = 2^{|S|} \right\}.$$

The VC dimension captures the *capacity* of the hypothesis class $\mathcal{H}$. The larger the capacity is, the more samples it takes to find an approximately optimal classifier in $\mathcal{H}$. More precisely, the sample complexity of ERM is given by the following theorem (see, e.g., (Shalev-Shwartz & Ben-David, 2014)).

Theorem 1 (Sample Complexity of ERM)**.** *Fix a feature space $\mathcal{X}$, a population distribution $\mathcal{D}$, and a hypothesis class $\mathcal{H}$. For any $\varepsilon > 0$, $\delta > 0$, given*

$$m = O\left((d_{\mathrm{VC}}(\mathcal{H}) + \log(1/\delta))/\varepsilon^2\right)$$

iid samples $S \sim \mathcal{D}^m$, with probability at least $1 - \delta$, any classifier $f \in \mathcal{H}$ minimizing the empirical loss on S, i.e., $f \in \operatorname{argmin}_{f' \in \mathcal{H}} \ell_S(f')$, has population loss at most $\ell_{\mathcal{D}}(f) \leq \ell_{\mathcal{D}}(\mathcal{H}) + \varepsilon$. Moreover, finding any classifier achieving the same relative loss ε and failure probability δ requires $\Omega\left((d_{\mathrm{VC}}(\mathcal{H}) + \log(1/\delta))/\varepsilon^2\right)$ iid samples.

The above theorem says that the ERM principle finds an approximately optimal (up to an additive loss ε) classifier within $\mathcal{H}$, with an asymptotically optimal number of samples. In other words, the low empirical loss of the classifier found by the ERM principle generalizes with respect to the population distribution $\mathcal{D}$.

2.2. Strategic Manipulation in Classification

In the classical PAC learning model, the classifier always observes the real feature of the point being classified. However, as argued above, this is often not the case in real-life scenarios, where the point being classified may strategically modify its feature in order to receive a more desirable outcome. In this paper, we model the data point's ability to modify its feature using a binary relation $\rightarrow$, which captures the *reporting structure* over the feature space. For $x_1, x_2 \in \mathcal{X}$, we say $x_1 \rightarrow x_2$ if a point with feature x_1 can pretend to have (i.e., *report*) feature x_2. For any $x \in \mathcal{X}$, we always have $x \rightarrow x$, corresponding to the fact that the point may choose not to modify its feature.[3] Throughout the paper, we assume that all points being classified prefer label 1 (corresponding to, e.g., acceptance) to 0. As a result, fixing a misreporting structure $\rightarrow$ and a classifier $f : \mathcal{X} \rightarrow \{0, 1\}$, a point with feature x will pretend to have feature x' such that $f(x') = 1$ whenever possible, i.e., it will report $x' \in \operatorname{argmax}_{x'' : x \rightarrow x''} f(x'')$. Note that it is possible that $x' = x$. This can happen when $f(x) = 1$, or $f(x'') = 0$ for any x'' all x'' such that $x \rightarrow x''$, including x itself.[4]

In the rest of the paper, we focus on the following variant of the PAC learning model. The learning algorithm has access to m iid *unmodified* samples $\{(x_i, y_i)\}_i \sim \mathcal{D}^m$, based on which a classifier $f \in \mathcal{H}$ is computed.[5] The classifier f is then deployed, and random points drawn from $\mathcal{D}$ modify their features strategically in response to the classifier. The strategic population loss of f, taking into consideration strategic manipulation, can be computed as

$$\widehat{\ell}_{\mathcal{D}}(f) = \Pr_{(x,y) \sim \mathcal{D}} \left[\max_{x' : x \rightarrow x'} f(x') \neq y \right].$$

Again, the goal is to find a classifier $f \in \mathcal{H}$ with strategic population loss at most $\widehat{\ell}_{\mathcal{D}}(\mathcal{H}) + \varepsilon$, with probability at least $1 - \delta$, where $\widehat{\ell}_{\mathcal{D}}(\mathcal{H}) = \min_{f' \in \mathcal{H}} \widehat{\ell}_{\mathcal{D}}(f')$.

[1] We treat S as an unordered set, since the order does not carry information.

[2] This is true only in the agnostic setting where $\ell_{\mathcal{D}}(\mathcal{H}) > 0$, on which we focus our attention throughout the paper. Similar results can be established for the realizable setting where $\ell_{\mathcal{D}}(\mathcal{H}) = 0$.

[3] Note that this is *not* a technical requirement — all results in this paper still hold without this assumption.

[4] A seemingly more general model is one in which there is a cost $c(x, x') \geq 0$ for a point with feature x to report feature x'. We remark that with binary labels, it is without loss of generality to ignore this cost, since a way of reporting is feasible iff the gain of receiving label 1 is larger than the cost of reporting.

[5] Since each data point is only classified once, data points in the sample set observed by the learner have no incentive to manipulate the learning process, and thus will not modify their features.

3. Incentive-Aware Empirical Risk Minimization

In this section, we investigate the ERM principle in the presence of strategic manipulation. We first give an example, showing that the vanilla ERM principle, which ignores the strategic aspect of the problem, has poor performance in general. We then consider an incentive-aware variant of ERM, and analyze its generalization behavior.

3.1. How Vanilla ERM Fails

Consider the following example. The feature space $\mathcal{X} = \{x_1, x_2, x_3\}$, the hypothesis class $\mathcal{H} = 2^{\mathcal{X}}$, and the population distribution $\mathcal{D}$ assigns probability 0.5 to feature-label pair $(x_1, 0)$, and probability 0.5 to $(x_2, 1)$. The reporting structure allows (besides $x \to x$ for any $x \in \mathcal{X}$) $x_1 \to x_2$ and $x_2 \to x_3$. Note that since the feature space is extremely simple, even the power set $\mathcal{H} = 2^{\mathcal{X}}$ (which has VC dimension $d_{\mathrm{VC}}(\mathcal{H}) = 3$) exhibits decent generalization behavior *without* strategic manipulation. Recall that the vanilla ERM principle finds an arbitrary classifier which minimizes the empirical loss based on the unmodifed sample set. Suppose the number of samples $m = \omega(1)$. Then with high probability, the sample set consists of copies of $(x_1, 0)$ and $(x_2, 1)$ and nothing else. Any classifier f satisfying $f(x_1) = 0$ and $f(x_2) = 1$ achieves 0 empirical loss. However, with strategic manipulation, such an f ends up assigning label 1 to all (new) points whose (true) feature is x_1, since those points can report x_2 to fool the classifier. As a result, the strategic population loss is

$$\begin{aligned}\widehat{\ell}_{\mathcal{D}}(f) &= \Pr_{(x,y)\sim\mathcal{D}}\left[\max_{x':x\to x'} f(x') \neq y\right] \\ &\geq 0.5 \cdot \left|\max_{x':x_1\to x'} f(x') - 0\right| \\ &= 0.5 \cdot |f(x_2) - 0| = 0.5.\end{aligned}$$

In other words, with high probability, any classifier found using the vanilla ERM principle is no better than random guessing. But in fact, the classifier f' with $f'(x_1) = f'(x_2) = 0$ and $f'(x_3) = 1$ would have 0 loss with strategic behavior, because $\max_{x':x_1\to x'} f'(x') = 0$ and $\max_{x':x_2\to x'} f'(x') = f'(x_3) = 1$. Hence it is possible to do much better than vanilla ERM.

The above example shows that with strategic manipulation, the vanilla ERM principle fails spectacularly even in extremely simple cases that would be trivial without strategic manipulation. This observation aligns with the intuition that any learning procedure achieving nontrivial performance must exploit the reporting structure.

3.2. The Incentive-Aware ERM Principle

Below we present an incentive-aware version of the ERM principle (henceforth IA ERM), and show that (1) IA ERM finds a classifier with the desired properties, and (2) the sample complexity of IA ERM is asymptotically optimal. These properties of IA ERM closely resemble those of its classical counterpart without strategic manipulation, and suggests that IA ERM is the right version of the ERM principle in the strategic setting that we consider.

Incentive-aware ERM. The idea is to minimize the *strategic* empirical loss $\widehat{\ell}_S(f)$ on the sample set $S = \{(x_i, y_i)\}_i$, computed by replacing the true feature of every sample point with the most beneficial feature that the point can report, i.e.,

$$\widehat{\ell}_S(f) = \frac{1}{|S|}\sum_{i\in[|S|]}\left|\max_{x':x_i\to x'} f(x') - y_i\right|.$$

The strategic empirical loss is a natural quantity to consider, since the expectation of $\widehat{\ell}_S(f)$ over S is precisely the strategic population loss $\widehat{\ell}_{\mathcal{D}}(f)$, which we aim to minimize. Given the notion of strategic empirical loss, IA ERM simply finds any classifier f in the hypothesis class $\mathcal{H}$ which minimizes $\widehat{\ell}_S(f)$, i.e., $f \in \mathrm{argmin}_{f'\in\mathcal{H}} \widehat{\ell}_S(f')$.

We now analyze the generalization behavior of IA ERM. Recall that in the classical PAC setting without strategic manipulation, the sample complexity of ERM depends on the VC dimension of the hypothesis class $\mathcal{H}$. Here, with strategic manipulation, it appears that the VC dimension of $\mathcal{H}$ is no longer the right capacity measure to consider. Instead of $\mathcal{H}$, we consider the *effective* hypothesis class $\widehat{\mathcal{H}}$, defined below.

Definition 2 (Effective Classifier / Hypothesis Class)**.** Fixing a feature space $\mathcal{X}$ and a reporting structure $\to$, for each classifier $f \subseteq \mathcal{X}$, the *effective classifier* $\widehat{f}$ consists of the set of features to which f effectively assigns label 1, i.e.,

$$\widehat{f} = \{x \in \mathcal{X} \mid \exists x' : x \to x', f(x') = 1\}.$$

Fixing a hypothesis class $\mathcal{H}$, the effective hypothesis class $\widehat{\mathcal{H}}$ consists of the effective classifier of every classifier in $\mathcal{H}$, i.e., $\widehat{\mathcal{H}} = \{\widehat{f} \mid f \in \mathcal{H}\}$.

One may intuitively interpret the effective hypothesis class in the following way. Any classifier f in the presence of strategic manipulation is equivalent to the effective classifier $\widehat{f}$ without strategic manipulation. Therefore, by considering the effective hypothesis class $\widehat{\mathcal{H}}$, we can effectively ignore strategic manipulation, and reduce the generalization analysis to that in the classical PAC setting, i.e., Theorem 1. This gives us the following theorem.

Theorem 2 (Sample Complexity of IA ERM)**.** *Fix a feature space $\mathcal{X}$, a population distribution $\mathcal{D}$, a reporting structure $\to$, and a hypothesis class $\mathcal{H}$. For any $\varepsilon > 0$, $\delta > 0$, given*

$$m = O\left((d_{\mathrm{VC}}(\widehat{\mathcal{H}}) + \log(1/\delta))/\varepsilon^2\right)$$

iid samples $S \sim \mathcal{D}^m$, with probability at least $1 - \delta$, any classifier $f \in \mathcal{H}$ minimizing the strategic empirical loss on S, i.e., $f \in \operatorname{argmin}_{f' \in \mathcal{H}} \widehat{\ell}_S(f')$, has strategic population loss at most $\widehat{\ell}_{\mathcal{D}}(f) \le \widehat{\ell}_{\mathcal{D}}(\mathcal{H}) + \varepsilon$. Moreover, finding any classifier achieving the same relative loss ε and failure probability δ requires $\Omega\left((d_{\mathrm{VC}}(\widehat{\mathcal{H}}) + \log(1/\delta))/\varepsilon^2\right)$ iid samples.

The proof of Theorem 2, as well as all other proofs, are in Appendix A. In words, the above theorem says that IA ERM finds a classifier whose strategic population loss is close to the best classifier in the hypothesis class $\mathcal{H}$, and the sample complexity of finding such a classifier is determined by the VC dimension of the effective hypothesis class $\widehat{\mathcal{H}}$, rather than that of $\mathcal{H}$ itself.

We make a few remarks regarding IA ERM. First, when the reporting structure $\to$ is trivial, i.e., a point with feature x can only report x itself, then the effective class $\widehat{\mathcal{H}} = \mathcal{H}$, and IA ERM reduces to vanilla ERM. In such cases, Theorem 2 reduces precisely to the classical Theorem 1. Second, in general, the VC dimension $d_{\mathrm{VC}}(\widehat{\mathcal{H}})$ of the effective class $\widehat{\mathcal{H}}$ may be smaller or larger than that of $\mathcal{H}$. Indeed, as we show in Appendix B, either of the two quantities can be arbitrarily larger than the other.

4. Incentive-Compatible Empirical Risk Minimization

The IA ERM principle, together with Theorem 2, provides a rather general solution for PAC learning in the presence of strategic manipulation. Still, one may wonder whether it is possible to achieve more desirable properties, e.g., enhanced robustness against strategic manipulation and better generalization guarantees, by further refining/regularizing IA ERM. In this section, we propose *incentive compatibility as a means of regularization*, which leads to the incentive-compatible ERM principle (henceforce IC ERM).

We first present the IC ERM principle and provide a general analysis of its sample complexity in Section 4.1. In particular, we show that IC ERM generalizes no worse than IA ERM when applied to the same hypothesis class. In Section 4.2, we consider a special case of IC ERM, *free* IC ERM, where the hypothesis class $\mathcal{H}$ is implicitly all possible classifiers over the feature space $\mathcal{X}$, i.e., $\mathcal{H} = 2^{\mathcal{X}}$. We give an efficient algorithm for free IC ERM, and more importantly, we show that, somewhat surprisingly, it is still possible to derive nontrivial generalization bounds for free IC ERM. Based on the generalization analysis of free IC ERM, we provide a hypothesis-class-independent sample complexity bound for IC ERM, which further illustrates the power of incentive compatibility as a means of regularization. Finally, in Section 4.3, we discuss an important class of reporting structures, i.e., *transitive* reporting structures, on which IC ERM is equivalent to the more general IA ERM. Based on this equivalence, we give a hypothesis-class-independent sample complexity bound for IA ERM that applies whenever the reporting structure is transitive.

4.1. The Incentive-Compatible ERM Principle

First we introduce the notions of incentive-compatible classifiers and hypothesis classes. Incentive compatibility is a standard concept in mechanism design and straightforwardly applying it in our context results in the following definition.

Definition 3 (Incentive-Compatible Classifiers / Hypothesis Classes)**.** Fixing a feature space $\mathcal{X}$ and a reporting structure $\to$, a classifier $f : \mathcal{X} \to \{0, 1\}$ is *incentive compatible* if for any $x_1, x_2 \in \mathcal{X}$,

$$x_1 \to x_2 \Longrightarrow f(x_1) \ge f(x_2).$$

In other words, no point can receive a better outcome by pretending to have a different feature. A hypothesis class $\mathcal{H}$ is incentive compatible if it contains only incentive-compatible classifiers.

The intuition behind the definitions is simple: when an incentive-compatible classifier is deployed, no point is motivated to strategically modify its feature, since it is impossible to obtain a better outcome by doing so. As a result, the effective classifier induced by any incentive-compatible classifier f is itself, i.e., $\widehat{f} = f$, and the strategic (empirical) loss of an incentive-compatible classifier f is the same as the (empirical) loss of the same classifier without strategic manipulation, i.e., $\widehat{\ell}_{\mathcal{D}}(f) = \ell_{\mathcal{D}}(f)$ and $\widehat{\ell}_S(f) = \ell_S(f)$. Incentive-compatible classifiers therefore provide arguably the strongest robustness one may hope for against strategic manipulation — they eliminate strategic manipulation. The IC ERM principle presented below always finds a classifier that is incentive compatible.

Incentive-compatible ERM. For any hypothesis class $\mathcal{H}$, let the *incentive-compatible subclass* $\mathcal{H}^{\mathrm{IC}(\to)}$ be the largest subset of $\mathcal{H}$ that is incentive compatible under $\to$, i.e.,

$$\mathcal{H}^{\mathrm{IC}(\to)} = \{f \in \mathcal{H} \mid f \text{ is incentive compatible under } \to\}.$$

We omit the parameter $\to$ when it is clear from the context. Given a hypothesis class $\mathcal{H}$, the IC ERM principle finds any classifier f in the incentive-compatible subclass $\mathcal{H}^{\mathrm{IC}}$ minimizing the empirical loss $\ell_S(f)$ on the sample set S, i.e., $f \in \operatorname{argmin}_{f' \in \mathcal{H}^{\mathrm{IC}}} \ell_S(f') = \operatorname{argmin}_{f' \in \mathcal{H}^{\mathrm{IC}}} \widehat{\ell}_S(f')$.

We now analyze the sample complexity of IC ERM. Observe that IC ERM with hypothesis class $\mathcal{H}$ is equivalent to vanilla ERM with hypothesis class $\mathcal{H}^{\mathrm{IC}}$. Therefore, applying Theorem 1 to $\mathcal{H}^{\mathrm{IC}}$, we immediately obtain the following asymptotically optimal sample complexity bound for IC ERM.

Theorem 3 (Sample Complexity of IC ERM)**.** *Fix a feature space $\mathcal{X}$, a population distribution $\mathcal{D}$, a reporting structure $\to$, and a hypothesis class $\mathcal{H}$. For any $\varepsilon > 0$, $\delta > 0$, given*

$$m = O\left((d_{\mathrm{VC}}(\mathcal{H}^{\mathrm{IC}}) + \log(1/\delta))/\varepsilon^2\right)$$

iid samples $S \sim \mathcal{D}^m$, with probability at least $1 - \delta$, any classifier $f \in \mathcal{H}^{\mathrm{IC}}$ minimizing the empirical loss on S, i.e., $f \in \operatorname{argmin}_{f' \in \mathcal{H}^{\mathrm{IC}}} \ell_S(f')$, has strategic population loss at most $\widehat{\ell}_{\mathcal{D}}(f) \le \widehat{\ell}_{\mathcal{D}}(\mathcal{H}^{\mathrm{IC}}) + \varepsilon$. Moreover, finding any incentive-compatible classifier within $\mathcal{H}$ achieving the same relative loss ε and failure probability δ requires $\Omega\left((d_{\mathrm{VC}}(\mathcal{H}^{\mathrm{IC}}) + \log(1/\delta))/\varepsilon^2\right)$ iid samples.

Theorem 3 is but a direct application of the classical Theorem 1. To obtain further insights into the sample complexity of IC ERM, we need to take a closer look at the VC dimension of the incentive-compatible subclass $\mathcal{H}^{\mathrm{IC}}$, which dictates the sample complexity. As discussed above, IC ERM can be viewed as a regularized version of IA ERM. Below we formalize this intuition, by showing that the sample complexity of IC ERM is in fact no larger than that of IA ERM, or that of vanilla ERM, when applied to the same hypothesis class $\mathcal{H}$. This is done in the following claim via upper bounding the VC dimension of $\mathcal{H}^{\mathrm{IC}}$ by that of $\widehat{\mathcal{H}}$, as well as that of $\mathcal{H}$.

Proposition 1. *Fixing a feature space $\mathcal{X}$ and a reporting structure $\to$, for any hypothesis class $\mathcal{H}$ over $\mathcal{X}$, $\mathcal{H}^{\mathrm{IC}} \subseteq \mathcal{H} \cap \widehat{\mathcal{H}}$. As a result, $d_{\mathrm{VC}}(\mathcal{H}^{\mathrm{IC}}) \le \min\{d_{\mathrm{VC}}(\mathcal{H}), d_{\mathrm{VC}}(\widehat{\mathcal{H}})\}$.*

In many natural scenarios, the VC dimension of $\mathcal{H}^{\mathrm{IC}}$ is significantly smaller than $d_{\mathrm{VC}}(\widehat{\mathcal{H}})$ (see Appendix B for examples). So, in addition to the highly desirable property of incentive compatibility, IC ERM also has at most the same, and often much better sample complexity than IA ERM, which translates to better generalization fixing the number of samples. We also remark that IC ERM finds an approximately optimal classifier in the incentive-compatible subclass $\mathcal{H}^{\mathrm{IC}}$, which may have a worse approximation error than $\mathcal{H}$. In other words, to achieve incentive compatibility, one in general has to sacrifice some accuracy in the form of a larger approximation error. We will see more desirable properties of IC ERM in the rest of this section.

4.2. Free IC ERM and Generalization from Incentive Compatibility

We now consider free IC ERM, which is a special case of IC ERM where the hypothesis class consists of all possible classifiers over $\mathcal{X}$, i.e., $\mathcal{H} = \mathcal{H}_0 = 2^{\mathcal{X}}$. At first glance this may appear senseless — in the classical PAC setting without strategic manipulation, no (nontrivial) generalization is possible if nothing is known about the ground truth a priori, i.e., when the hypothesis class consists of all possible classifiers. However, as we show below, the reporting structure, together with the requirement of incentive compatibility, in fact induces nontrivial structure over the incentive-compatible subclass $\mathcal{H}_0^{\mathrm{IC}}$, which allows nontrivial sample complexity and generalization bounds. These structures are captured by the following complexity measure, which we term the *intrinsic VC dimension*.

Definition 4 (Intrinsic VC Dimension)**.** Fix a reporting structure $\to$ over a feature space $\mathcal{X}$. For any $x, x' \in \mathcal{X}$, we say x can *reach* x' (denoted $x \Rightarrow x'$) if there exists a sequence of features $x_1, \dots, x_k$ for some integer $k > 0$, such that $x_1 = x$, $x_k = x'$, and for any $i \in [k-1]$, $x_i \to x_{i+1}$. A set of features $S \subseteq \mathcal{X}$ is *independent* if for any $x, x' \in S$ where $x \ne x'$, x cannot reach x'. The intrinsic VC dimension of $\to$ over $\mathcal{X}$, $d_{\mathrm{VC}}(\mathcal{X}, \to)$, is the cardinality of any maximum subset of $\mathcal{X}$ that is independent, i.e.,

$$\begin{aligned} d_{\mathrm{VC}}(\mathcal{X}, \to) = \max\{|S| \mid S \subseteq \mathcal{X} : \nexists\, x, x' \in S \\ \text{such that } x \ne x' \text{ and } x \Rightarrow x'\}. \end{aligned}$$

It turns out that the intrinsic VC dimension of the reporting structure is precisely the VC dimension of the incentive-compatible subclass $\mathcal{H}_0^{\mathrm{IC}}$ of the null hypothesis class $\mathcal{H}_0$, as formalized in the following proposition.

Proposition 2 (VC Dimension of Null Hypothesis Class)**.** *For any feature space $\mathcal{X}$ and reporting structure $\to$ over $\mathcal{X}$, $d_{\mathrm{VC}}(\mathcal{X}, \to) = d_{\mathrm{VC}}(\mathcal{H}_0^{\mathrm{IC}(\to)})$, where $\mathcal{H}_0 = 2^{\mathcal{X}}$ is the null hypothesis class.*

Note that for any hypothesis class $\mathcal{H} \subseteq 2^{\mathcal{X}} = \mathcal{H}_0$ over $\mathcal{X}$, the incentive-compatible subclass of $\mathcal{H}$ is a subclass of that of the null hypothesis class $\mathcal{H}_0$, i.e., $\mathcal{H}^{\mathrm{IC}} \subseteq \mathcal{H}_0^{\mathrm{IC}}$. As a result, we always have

$$d_{\mathrm{VC}}(\mathcal{H}^{\mathrm{IC}}) \le d_{\mathrm{VC}}(\mathcal{H}_0^{\mathrm{IC}}) = d_{\mathrm{VC}}(\mathcal{X}, \to).$$

This, together with Theorem 3, immediately implies the following hypothesis-class-independent sample complexity bound for IC ERM, which, in particular, applies to free IC ERM.

Theorem 4 (Hypothesis-Class-Independent Sample Complexity Bound for IC ERM)**.** *Fix a feature space $\mathcal{X}$, a population distribution $\mathcal{D}$, a reporting structure $\to$, and a hypothesis class $\mathcal{H}$. For any $\varepsilon > 0$, $\delta > 0$, given*

$$m = O\left((d_{\mathrm{VC}}(\mathcal{X}, \to) + \log(1/\delta))/\varepsilon^2\right)$$

iid samples $S \sim \mathcal{D}^m$, with probability at least $1 - \delta$, any classifier $f \in \mathcal{H}^{\mathrm{IC}}$ minimizing the empirical loss *on S, i.e., $f \in \operatorname{argmin}_{f' \in \mathcal{H}^{\mathrm{IC}}} \ell_S(f')$, has strategic population loss $\widehat{\ell}_{\mathcal{D}}(f) \leq \widehat{\ell}_{\mathcal{D}}(\mathcal{H}^{\mathrm{IC}}) + \varepsilon$. Moreover, when the hypothesis class $\mathcal{H}$ is the null hypothesis class $\mathcal{H}_0$, finding any incentive-compatible classifier achieving the same relative loss ε and failure probability δ requires $\Omega\left((d_{\mathrm{VC}}(\mathcal{X}, \to) + \log(1/\delta))/\varepsilon^2\right)$ iid samples.*

We also present an efficient algorithm for free IC ERM in Appendix C.

4.3. Transitive Reporting and the Revelation Principle

Finally, we investigate an important family of reporting structures, transitive reporting structures. It turns out that with transitivity, the so-called revelation principle from mechanism design holds: if one accounts for strategic reporting, then without loss of generality one may focus on incentive-compatible classifiers. That is, IC ERM is as general as IA ERM. As a result, with transitivity, the hypothesis-class-independent sample complexity bound extends to IA ERM applied to any hypothesis class. We first give the formal definition of transitive reporting structures, which is essentially the same as that of transitive binary relations.

Definition 5 (Transitive Reporting Structures)**.** A reporting structure $\to$ over $\mathcal{X}$ is *transitive* if for any $x_1, x_2, x_3 \in \mathcal{X}$, $(x_1 \to x_2$ and $x_2 \to x_3) \implies x_1 \to x_3$.

One example of transitive reporting structures is the example from the introduction, where $\mathcal{X} = \mathbb{R}_+$, and $\to$ is precisely the same as $\geq$, which is clearly transitive (see Appendix B for more details). It turns out that transitivity of reporting structures is equivalent to the revelation principle holding, which roughly says that any classifier is equivalent to a (possibly different) incentive-compatible classifier. Formally:

Proposition 3 (Revelation Principle for PAC Learning)**.** *The following is true if and only if the reporting structure is transitive: for any classifier $f : \mathcal{X} \to \{0, 1\}$, the effective classifier $\widehat{f}$ is incentive compatible, and as a result, for any hypothesis class $\mathcal{H}$, the effective class $\widehat{\mathcal{H}}$ is incentive-compatible (i.e., $(\widehat{\mathcal{H}})^{\mathrm{IC}} = \widehat{\mathcal{H}}$).*

In light of Proposition 3, when the reporting structure is transitive, for any $\mathcal{H}$, IA ERM with hypothesis class $\mathcal{H}$ is equivalent to IC ERM with hypothesis class $\widehat{\mathcal{H}}$, in the sense that they yield the same effective classifier given any sample set, with the same sample complexity. This gives us a way to translate the hypothesis-class-independent sample complexity bound for IC ERM to IA ERM, and obtain the following theorem.

Theorem 5 (Hypothesis-Class-Independent Sample Complexity Bound for IA ERM)**.** *Fix a feature space $\mathcal{X}$, a population distribution $\mathcal{D}$, a transitive reporting structure $\to$, and a hypothesis class $\mathcal{H}$. For any $\varepsilon > 0$, $\delta > 0$, given*

$$m = O\left((d_{\mathrm{VC}}(\mathcal{X}, \to) + \log(1/\delta))/\varepsilon^2\right)$$

iid samples $S \sim \mathcal{D}^m$, with probability at least $1 - \delta$, any classifier $f \in \mathcal{H}$ minimizing the strategic empirical loss on S, i.e., $f \in \operatorname{argmin}_{f' \in \mathcal{H}} \widehat{\ell}_S(f')$, has strategic population loss $\widehat{\ell}_{\mathcal{D}}(f) \leq \widehat{\ell}_{\mathcal{D}}(\mathcal{H}) + \varepsilon$. Moreover, when the hypothesis class $\mathcal{H}$ is the null hypothesis class $\mathcal{H}_0$, finding any classifier achieving the same relative loss ε and failure probability δ requires $\Omega\left((d_{\mathrm{VC}}(\mathcal{X}, \to) + \log(1/\delta))/\varepsilon^2\right)$ iid samples.

To develop some intuition, consider again the introductory example. As discussed above, there, $\mathcal{X} = \mathbb{R}_+$ and $\to = \geq$ is transitive, and the intrinsic VC dimension of $\to$ is $d_{\mathrm{VC}}(\mathcal{X}, \to) = 1$. In fact, any classifier $f \in 2^{\mathcal{X}}$ effectively implements a threshold θ_f, where $f(x) = 1$ iff $x \geq \theta_f$ or $x > \theta_f$ (see Appendix B for more details). It is well-known (see, e.g., (Shalev-Shwartz & Ben-David, 2014)) that $\Theta(\log(1/\delta)/\varepsilon^2)$ samples suffice to learn such a threshold with relative loss ε and failure probability δ, which coincides with the above sample complexity bound.

Broader Impact

Potential positive impact:

- Our results can be applied, for example, to prevent fraud.
- Naïve learning processes that do not anticipate strategic behavior will not perform at all as expected in practice, creating substantial risk.
- Our results could make a socially beneficial classification process more accurate.
- By encouraging truthful reporting, our results could allow for more efficient allocation of resources to those in actual need.
- Classifiers that are not incentive compatible result in strategic behavior being advantageous to individuals. Vulnerable communities often are not aware of this or do not know how to play the game strategically, exacerbating inequality. Incentive compatible classifiers remove this concern. (Similar arguments have been made for the use of incentive compatible mechanisms in allocating students to public schools, where wealthier parents often know how to play the game better.)

Potential negative impact:

- Our results could make a socially harmful classification process more accurate.
- Our results encourage entities being classified to report their true features, which, if handled carelessly, could result in privacy issues.

References

Alon, N., Ben-David, S., Cesa-Bianchi, N., and Haussler, D. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM (JACM)*, 44(4):615–631, 1997.

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Braverman, M., Mao, J., Schneider, J., and Weinberg, S. M. Multi-armed bandit problems with strategic arms. In *Conference on Learning Theory*, pp. 383–416, 2019.

Chen, Y., Podimata, C., Procaccia, A. D., and Shah, N. Strategyproof linear regression in high dimensions. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pp. 9–26, 2018.

Daniely, A., Sabato, S., Ben-David, S., and Shalev-Shwartz, S. Multiclass learnability and the erm principle. *The Journal of Machine Learning Research*, 16(1):2377–2404, 2015.

Dekel, O., Fischer, F., and Procaccia, A. D. Incentive compatible regression learning. *Journal of Computer and System Sciences*, 76(8):759–777, 2010.

Devroye, L., Györfi, L., and Lugosi, G. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

Feng, Z., Parkes, D. C., and Xu, H. The intrinsic robustness of stochastic bandits to strategic manipulation. *arXiv preprint arXiv:1906.01528*, 2019.

Freeman, R., Pennock, D. M., Podimata, C., and Vaughan, J. W. No-regret and incentive-compatible prediction with expert advice. *arXiv preprint arXiv:2002.08837*, 2020.

Green, J. R. and Laffont, J.-J. Partially verifiable information and mechanism design. *The Review of Economic Studies*, 53(3):447–456, 1986.

Haghtalab, N., Immorlica, N., Lucier, B., and Wang, J. Maximizing welfare with incentive-aware evaluation mechanisms. In *29th International Joint Conference on Artificial Intelligence*, 2020.

Hanneke, S. The optimal sample complexity of pac learning. *The Journal of Machine Learning Research*, 17(1):1319–1333, 2016.

Hardt, M., Megiddo, N., Papadimitriou, C., and Wootters, M. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pp. 111–122, 2016.

Kephart, A. and Conitzer, V. Complexity of mechanism design with signaling costs. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pp. 357–365, 2015.

Kephart, A. and Conitzer, V. The revelation principle for mechanism design with reporting costs. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, pp. 85–102, 2016.

Kleinberg, J. and Raghavan, M. How do classifiers induce agents to invest effort strategically? In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pp. 825–844, 2019.

Perote, J. and Perote-Pena, J. Strategy-proof estimators for simple regression. *Mathematical Social Sciences*, 47(2): 153–176, 2004.

Pollard, D. *Convergence of stochastic processes*. Springer Science & Business Media, 2012.

Roughgarden, T. and Schrijvers, O. Online prediction with selfish experts. In *Advances in Neural Information Processing Systems*, pp. 1300–1310, 2017.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Talagrand, M. Sharper bounds for gaussian and empirical processes. *The Annals of Probability*, pp. 28–76, 1994.

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

Vapnik, V. 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.

Yu, L. Mechanism design with partial verification and revelation principle. *Autonomous Agents and Multi-Agent Systems*, 22(1):217–223, 2011.

Zhang, H., Cheng, Y., and Conitzer, V. Distinguishing distributions when samples are strategically transformed. In *Advances in Neural Information Processing Systems*, pp. 3187–3195, 2019a.

Zhang, H., Cheng, Y., and Conitzer, V. When samples are strategically selected. In *International Conference on Machine Learning*, pp. 7345–7353, 2019b.

A. Omitted Proofs

Proof of Theorem 2. First observe that for any classifier $f \in 2^{\mathcal{X}}$, the strategic population loss $\widehat{\ell}_{\mathcal{D}}(f) = \ell_{\mathcal{D}}(\widehat{f})$, and the strategic empirical loss $\widehat{\ell}_S(f) = \ell_S(\widehat{f})$. The plan is to apply Theorem 1 to the effective hypothesis class $\widehat{\mathcal{H}}$. Let

$$f \in \operatorname{argmin}_{f' \in \mathcal{H}} \widehat{\ell}_S(f') = \operatorname{argmin}_{f' \in \mathcal{H}} \ell_S(\widehat{f'}).$$

We then have

$$\widehat{f} \in \operatorname{argmin}_{f' \in \widehat{\mathcal{H}}} \ell_S(f').$$

That is, $\widehat{f}$ is a minimizer of ℓ_S in $\widehat{\mathcal{H}}$. By Theorem 1, with m samples, with probability at least $1 - \delta$,

$$\ell_{\mathcal{D}}(\widehat{f}) \le \ell_{\mathcal{D}}(\widehat{\mathcal{H}}) + \varepsilon = \widehat{\ell}_{\mathcal{D}}(\mathcal{H}).$$

On the other hand, Theorem 1 states that finding such an $\widehat{f}$ in $\widehat{\mathcal{H}}$ with relative loss ε and failure probability δ requires asymptotically the same number of samples. This conlcudes the proof of the theorem. □

Proof of Theorem 3. Theorem 3 is a direct corollary of Theorem 1 (or Theorem 2 which is more general). Applying Theorem 1 with hypothesis class $\mathcal{H}^{\mathrm{IC}}$, we immediately obtain that with m samples, for any

$$f \in \operatorname{argmin}_{f' \in \mathcal{H}^{\mathrm{IC}}} \ell_S(f'),$$

with probability $1 - \delta$,

$$\ell_{\mathcal{D}}(f) \le \ell_{\mathcal{D}}(\mathcal{H}^{\mathrm{IC}}) + \varepsilon.$$

And moreover, the number of samples is asymptotically tight. On the other hand, since $\mathcal{H}^{\mathrm{IC}}$ is incentive-compatible,

$$\operatorname{argmin}_{f' \in \mathcal{H}^{\mathrm{IC}}} \ell_S(f') = \operatorname{argmin}_{f' \in \mathcal{H}^{\mathrm{IC}}} \widehat{\ell}_S(f'),$$

and

$$\widehat{\ell}_{\mathcal{D}}(f) \le \widehat{\ell}_{\mathcal{D}}(\mathcal{H}^{\mathrm{IC}}) + \varepsilon.$$

This concludes the proof. □

Proof of Proposition 1. For any $f \in \mathcal{H}^{\mathrm{IC}}$, since f is incentive-compatible, we have

$$f = \widehat{f} \in \widehat{\mathcal{H}},$$

and therefore $\mathcal{H}^{\mathrm{IC}} \subseteq \widehat{\mathcal{H}}$. On the other hand, by definition, $\mathcal{H}^{\mathrm{IC}} \subseteq \mathcal{H}$. This implies that $\mathcal{H}^{\mathrm{IC}} \subseteq \mathcal{H} \cap \widehat{\mathcal{H}}$. □

Proof of Proposition 2. First we show $d_{\mathrm{VC}}(\mathcal{H}_0^{\mathrm{IC}}) \ge d_{\mathrm{VC}}(\mathcal{X}, \to)$. Let $S \subseteq \mathcal{X}$ be an independent subset of $\mathcal{X}$ with cardinality $d_{\mathrm{VC}}(\mathcal{X}, \to)$. Such a subset exists by the definition of $d_{\mathrm{VC}}(\mathcal{X}, \to)$. We argue that S can be shattered by $\mathcal{H}_0^{\mathrm{IC}}$. For any $T \subseteq S$, we construct a classifier $f_T \in \mathcal{H}_0^{\mathrm{IC}}$ such that for any $x \in S$, $f_T(x) = 1 \iff x \in T$. Let f_T be such that

$$f_T(x) = \begin{cases} 1, & \text{if there exists } x' \in T : x \Rightarrow x' \\ 0, & \text{otherwise.} \end{cases}$$

We only need to check that f_T is incentive-compatible. Suppose otherwise, i.e., there exist $x_1, x_2 \in \mathcal{X}$, such that $x_1 \to x_2$, $f_T(x_1) = 0$, and $f_T(x_2) = 1$. It must be the case that for some $x_3 \in T$, such that $x_2 \Rightarrow x_3$. Then, by definition, we have $x_1 \Rightarrow x_3$, and therefore it should be the case that $f_T(x_1) = 1$, a contradiction.

Now we show $d_{\mathrm{VC}}(\mathcal{H}_0^{\mathrm{IC}}) \le d_{\mathrm{VC}}(\mathcal{X}, \to)$. That is, for any subset $S \subseteq \mathcal{X}$ where $|S| > d_{\mathrm{VC}}(\mathcal{X}, \to)$, S cannot be shattered by $\mathcal{H}_0^{\mathrm{IC}}$. By the definition of $d_{\mathrm{VC}}(\mathcal{X}, \to)$, S cannot be independent. Let $x, x' \in S$ be such that $x \Rightarrow x'$. Furthermore, let

$$x = x_1, x_2, \ldots, x_{k-1}, x_k = x' \in \mathcal{X}$$

be a sequence through which x can reach x', i.e., for any $i \in [k-1]$, $x_i \to x_{i+1}$. We show that for any incentive-compatible f, it cannot be the case that $f(x) = 0$ and $f(x') = 1$. Suppose otherwise. Let $t \in [k-1]$ be the largest integer such that $f(x_t) = 0$. t exists since $f(x_1) = 0$ and $f(x_k) = 1$. Then we have $x_t \to x_{t+1}$, but $f(x_t) = 0 < 1 = f(x_{t+1})$, a contradiction. This concludes the proof of the proposition. □

Proof of Theorem 4. The theorem is a direct corollary of Theorem 3 and Proposition 2. Observe that $\mathcal{H}^{\mathrm{IC}} \subseteq \mathcal{H}_0^{\mathrm{IC}}$, and $d_{\mathrm{VC}}(\mathcal{H}^{\mathrm{IC}}) \le d_{\mathrm{VC}}(\mathcal{H}_0^{\mathrm{IC}}) = d_{\mathrm{VC}}(\mathcal{X}, \to)$. Applying Theorem 3, the number of samples required for IC ERM on $\mathcal{H}$ is

$$O\left(\frac{d_{\mathrm{VC}}(\mathcal{H}^{\mathrm{IC}}) + \log(1/\delta)}{\varepsilon^2}\right) = O\left(\frac{d_{\mathrm{VC}}(\mathcal{X}, \to) + \log(1/\delta)}{\varepsilon^2}\right).$$

This concludes the proof. □

Proof of Proposition 3. Fix any $f \in \mathcal{H}$, we show $\widehat{f}$ is incentive-compatible. First observe that since → is transitive,

$$\widehat{f}(x) = 1 \iff \exists x' : x \to x' \text{ and } f(x') = 1 \iff \exists x' : x \Rightarrow x' \text{ and } f(x') = 1.$$

For any $x_1, x_2 \in \mathcal{X}$ where $x_1 \to x_2$, if $\widehat{f}(x_2) = 1$, then there exists x' such that $x_2 \Rightarrow x'$ and $f(x') = 1$. Then since $x_1 \to x_2$, we have $x_1 \Rightarrow x'$, and as a result, $\widehat{f}(x_1) = 1$. This immediately implies the proposition. □

Proof of Theorem 5. The theorem is a corollary of Theorem 4 and Proposition 3. By Proposition 3, when → is transitive, IA ERM with hypothesis $\mathcal{H}$ is equivalent to IC ERM with hypothesis $\widehat{\mathcal{H}}$. By Theorem 4, the number of samples required for the latter is

$$O\left(\frac{d_{\mathrm{VC}}(\mathcal{X}, \to) + \log(1/\delta)}{\varepsilon^2}\right).$$

This concludes the proof. □

B. Useful Examples

Relation between $d_{\mathrm{VC}}(\mathcal{H})$ and $d_{\mathrm{VC}}(\widehat{\mathcal{H}})$. We first give an example where $d_{\mathrm{VC}}(\mathcal{H}) = \infty$ and $d_{\mathrm{VC}}(\widehat{\mathcal{H}}) = 1$. Consider the example discussed in the introduction. There, the feature space $\mathcal{X} = \mathbb{R}_+$, and the reporting structure → satisfies for any $x_1, x_2 \in \mathbb{R}_+$, $x_1 \to x_2 \iff x_1 \ge x_2$. Consider $\mathcal{H} = 2^{\mathbb{R}_+}$. Clearly $d_{\mathrm{VC}}(\mathcal{H}) = \infty$, since any $S \subseteq \mathbb{R}_+$ is shattered by $\mathcal{H}$. On the other hand, for any $f \in \mathcal{H}$, $\widehat{f}$ is essentially determined by a threshold θ_f, where

$$\theta_f = \inf\{x \mid f(x) = 1\}.$$

$\widehat{f}$ is then defined by

$$\widehat{f}(x) = \begin{cases} 0, & x < \theta_f \\ 1, & x > \theta_f \\ f(x), & x = \theta_f. \end{cases}$$

In other words, $\widehat{\mathcal{H}}$ is the class of threshold classifiers over $\mathbb{R}_+$. It is well-known that $d_{\mathrm{VC}}(\widehat{\mathcal{H}}) = 1$.

Now we show that $d_{\mathrm{VC}}(\widehat{\mathcal{H}})$ can be arbitrarily larger than $d_{\mathrm{VC}}(\mathcal{H})$. Let $\mathcal{X} = \mathbb{N}$, and $\mathcal{H} = \{\{i\} \mid i \in \mathbb{N}\}$, i.e., the set of all singletons. It is clear that $d_{\mathrm{VC}}(\mathcal{H}) = 1$. Now for any $d \in \mathbb{N}$, we construct → such that $d_{\mathrm{VC}}(\widehat{\mathcal{H}}) = d$.[6] Let → be such that (1) $i \to i$ for any $i \in \mathbb{N}$, and (2) for any $i \in \{d, \dots, d + 2^d - 1\}$, $j \to i$ if the j-th digit in the binary representation of $i - d$ is 1. Now we argue that $\widehat{\mathcal{H}}$ shatters $\{0, \dots, d-1\}$. In fact, any subset $S \subseteq \{0, \dots, d-1\}$ can be viewed as the binary representation of an integer i_S in $\{0, \dots, 2^d - 1\}$. Consider the classifier $f_S = \{i_S + d\} \in \mathcal{H}$. Clearly $\widehat{f_S} \cap \{0, d-1\} = S$, since precisely the points in S can report $i_S + d$, which is the only point assigned label 1 by f_S.

Relation between $d_{\mathrm{VC}}(\mathcal{H})$, $d_{\mathrm{VC}}(\mathcal{H}^{\mathrm{IC}})$, and $d_{\mathrm{VC}}(\mathcal{X}, \to)$. We give an example where $d_{\mathrm{VC}}(\mathcal{H}) = \infty$ and $d_{\mathrm{VC}}(\mathcal{H}^{\mathrm{IC}}) = 1$. Consider again the introductory example, where $\mathcal{X} = \mathbb{R}_+$. As we argue above, when $\mathcal{H} = 2^{\mathbb{R}_+}$, $\widehat{\mathcal{H}}$ is all threshold classifiers, and $d_{\mathrm{VC}}(\widehat{\mathcal{H}}) = 1$. On the other hand, by Proposition 1, $\mathcal{H}^{\mathrm{IC}} \subseteq \widehat{\mathcal{H}}$ (and in fact when → is transitive, $\mathcal{H}^{\mathrm{IC}} = \mathcal{H} \cap \widehat{\mathcal{H}}$). One may show that every classifier in $\widehat{\mathcal{H}}$ is incentive-compatible, and $\mathcal{H}^{\mathrm{IC}} = \widehat{\mathcal{H}}$. As a result, $d_{\mathrm{VC}}(\mathcal{H}^{\mathrm{IC}}) = 1$.

One may alternatively bound $d_{\mathrm{VC}}(\mathcal{H}^{\mathrm{IC}})$ using proposition 2, since the intrinsic VC dimension of → over $\mathcal{X}$ is $d_{\mathrm{VC}}(\mathcal{X}, \to) = 1$. This is almost trivial, since any singleton set is independent, and for any $x_1 \ne x_2$, either $x_1 < x_2$ or $x_2 < x_1$, so $\{x_1, x_2\}$ cannot be independent. As a result, $d_{\mathrm{VC}}(\mathcal{X}, \to) = 1$.

[6] Similar constructions could also give $d_{\mathrm{VC}}(\widehat{\mathcal{H}}) = \infty$.

Algorithm 1 Algorithm for Free IC ERM

```
Input: A reporting structure $\to$ over feature space $\mathcal{X}$, a sample set $S = \{(x_i, y_i)\}_i \sim \mathcal{D}^m$ of size $m$.
Output: An incentive-compatible classifier $f$ minimizing the empirical loss $\ell_S(f)$ on $S$.

Let $G = (V, E)$ be a capacitated directed graph, where $V = \{x_i\}_{i \in [m]} \cup \{s, t\}$ and $E = \emptyset$.
for $1 \le i < j \le m$ do
  if $x_i \to x_j$ then
    Let $E \leftarrow E \cup \{(x_j, x_i, \infty)\}$, i.e., add an edge from $x_j$ to $x_i$ with capacity $\infty$.
  end if
end for
for $i \in [m]$ do
  if $y_i = 0$ then
    Let $E \leftarrow E \cup \{(x_i, t, 1)\}$, i.e., add an edge from $x_i$ to $t$ with capacity $1$.
  else
    Let $E \leftarrow E \cup \{(s, x_i, 1)\}$, i.e., add an edge from $s$ to $x_i$ with capacity $1$.
  end if
end for
Compute an $s$-$t$ mincut $(C, \overline{C})$ on $G$, where $C$ is the set of vertices on the $s$ side of the cut.
Let $f$ be such that for any $x \in \mathcal{X}$, $f(x) = 1$ iff there exists $x' \in C \setminus \{s\}$ where $x \Rightarrow x'$; return $f$.
```

C. Algorithm for Free IC ERM

In this section we present an efficient algorithm, Algorithm 1, for free IC ERM.

We show below that Algorithm 1 does compute an empirical risk minimizer among all incentive-compatible classifiers.

Theorem 6. *Algorithm 1 finds a classifier f which satisfies*

$$f \in \text{argmin}_{f' \in \mathcal{H}_0^i c} \ell_S(f').$$

Proof. First observe that given an incentive-compatible classifier $f|_S$ restricted to $\{x_i\}_i$, one can always extend the classifier by assigning label 1 to $x \in \mathcal{X}$ iff there exists $x' \in \{x_i\}_i$ where $f|_S(x') = 1$ and $x \Rightarrow x'$. Such an extension assigns label one only if incentive-compatibility is violated otherwise.

Given the above observation, we only need to show that f minimizes the empirical loss among incentive-compatible classifiers restricted to the sample set S. We argue that each incentive-compatible classifier $f' : \{x_i\}_i \to \{0, 1\}$ corresponds bijjectively to a finite capacity s-t cut in the graph G constructed in Algorithm 1. Recall that a classifier f' is incentive-compatible (restricted to $\{x_i\}_i$) iff for any $i, j \in [m]$,

$$x_i \to x_j \implies f(x_i) \ge f(x_j).$$

Consider the cut $(C', \overline{C'})$ corresponding to f' defined such that $x_i \in C' \iff f'(x_i) = 1$. Per the construction in Algorithm 1, $x_i \to x_j$ iff there is an edge from x_j to x_i with infinite capacity, and $f'(x_i) < f'(x_j)$ iff $x_i \notin C'$ and $x_j \in C'$. The condition for f' being incentive-compatible is therefore equivalent to: no infinite capacity edge is cut by C'. In other words, C' has finite capacity.

Now since f found by the algorithm corresponds to a min-cut, it has to be incentive-compatible. We show below that f also minimizes the empirical loss on S. We rewrite the empirical loss of f' in the following way.

$$\begin{aligned}
\ell_S(f') &= \frac{1}{m} \sum_{i \in [m]} |f'(x_i) - y_i| \\
&= \frac{1}{m} \sum_{i \in [m] : y_i = 0} f'(x_i) + \frac{1}{m} \sum_{i \in [m] : y_i = 1} (1 - f'(x_i)) \\
&= \frac{1}{m} \left(\sum_{i \in [m] : y_i = 0} \mathbb{I}[x_i \in C'] + \sum_{i \in [m] : y_i = 1} \mathbb{I}[x_i \notin C'] \right).
\end{aligned}$$

Observe that the last line multiplied by m is exactly the capacity of C', since for each i where $y_i = 0$, $x_i \in C'$ iff the edge from s to x_i with capacity 1 is cut, and for each i where $y_i = 1$, $x_i \notin C'$ iff the edge from x_i to t with capacity 1 is cut. Therefore, minimizing the capacity of the cut is equivalent to minimizing the empirical loss of f' on S. We conclude that f found by Algorithm 1 is in fact an incentive-compatible classifier with minimum empirical loss on S. □