Incentive Compatible Regression Learning

Ofer Dekel* Felix Fischer† Ariel D. Procaccia‡

Abstract

We initiate the study of incentives in a general machine learning framework. We focus on a game-theoretic regression learning setting where private information is elicited from multiple agents with different, possibly conflicting, views on how to label the points of an input space. This conflict potentially gives rise to untruthfulness on the part of the agents. In the restricted but important case when every agent cares about a single point, and under mild assumptions, we show that agents are motivated to tell the truth. In a more general setting, we study the power and limitations of mechanisms without payments. We finally establish that, in the general setting, the VCG mechanism goes a long way in guaranteeing truthfulness and economic efficiency.

1 Introduction

Machine learning is the area of computer science concerned with the design and analysis of algorithms that can learn from experience. A supervised learning algorithm observes a training set of labeled examples, and attempts to learn a rule that accurately predicts the labels of new examples. Following the rise of the Internet as a computational platform, machine learning problems have become increasingly dispersed, in the sense that different parts of the training set may be controlled by different computational or economic entities.

Motivation Consider an Internet search company trying to improve the performance of their search engine by learning a ranking function from examples. The ranking function is the heart of a modern search engine, and can be thought of as a mapping that assigns a real-valued score to every pair of a query and a URL. Some of the large Internet search companies currently hire Internet users (which we hereinafter refer to as "experts") to manually rank such pairs. These rankings may then be pooled and used to train a ranking function. Moreover, the experts are chosen in a way such that averaging over the experts' opinions and interests presumably pleases the average Internet user.

However, different experts may have different interests and a different idea of the results a good search engine should return. For instance, take the ambiguous query "Jaguar", which has become folklore in search engine designer circles. The top answer given by most search engines for this query is the website of the luxury car manufacturer. Knowing this, an animal-loving expert may decide to give this pair a disproportionately low score, hoping to improve the relative rank of websites dedicated to the Panthera Onca. An expert who is an automobile enthusiast may counter this measure by giving automotive websites a much higher score than is appropriate. From the search company's perspective, this type of strategic manipulation introduces an undesired bias in the training set.

Setting and Goals Our problem setting falls within the general boundaries of statistical regression learning. *Regression learning* is the task of constructing a real-valued function f based on a training set of examples, where each example consists of an input to the function and its corresponding output. In particular, the example $(\mathbf{x}, y)$ suggests that $f(\mathbf{x})$ should be equal to y. The accuracy of a function f on a given input-output pair $(\mathbf{x}, y)$ is defined using a loss function ℓ. Popular choices of the loss function are the squared loss, $\ell(f(\mathbf{x}), y) = (f(\mathbf{x}) - y)^2$, and the absolute loss, $\ell(f(\mathbf{x}), y) = |f(\mathbf{x}) - y|$. We typically assume that the training set is obtained by sampling i.i.d. from an underlying distribution over the product space of inputs and outputs. The overall quality of the function constructed by the learning algorithm is defined to be its expected loss, with respect to the same distribution.

We augment this well-studied setting by introducing a set of *strategic agents*. Each agent holds as private information an individual distribution over the input space and values for the points in the support of this distribution, and measures the quality of a regression function with respect to this data. The global goal, on the other hand, is to do well with respect to the average of the individual points of view. A training set is obtained by eliciting private information from the agents, who may reveal this information untruthfully in order to favorably influence the result of the learning process.

*School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem 91904, Israel, email: oferd@cs.huji.ac.il

†Institut für Informatik, Ludwig-Maximilians-Universität München, 80538 München, Germany, email: fischerf@tcs.ifi.lmu.de. The work was done while the author was visiting The Hebrew University of Jerusalem. This visit was supported by the School of Computer Science and Engineering and the Leibniz Center for Research in Computer Science at The Hebrew University of Jerusalem, and by the Deutsche Forschungsgemeinschaft under grant BR 2312/3-1.

‡School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem 91904, Israel, email: arielpro@cs.huji.ac.il. The author is supported by the Adams Fellowship Program of the Israel Academy of Sciences and Humanities.

Mechanism design is a subfield of economics that is concerned with the question of how to incentivize agents to truthfully report their private information, also known as their *type*. Given potentially non-truthful reports from the agents, a mechanism determines a global solution, and possibly additional monetary transfers to and from the agents. A mechanism is said to be *incentive compatible* if it is always in the agents' best interest to report their true types, and *efficient* if the solution maximizes social welfare (*i.e.*, minimizes the overall loss). Our goal in this paper will be to design and analyze incentive compatible and efficient mechanisms for the regression learning setting.

Results We begin our investigation by considering a restricted setting where each agent is only interested in a single point of the input space. Quite surprisingly, it turns out that a specific choice of ℓ, namely the absolute loss function, leads to excellent game-theoretic properties: the algorithm which simply finds an empirical risk minimizer on the training set is group strategyproof, *i.e.*, no coalition of agents is motivated to lie. All of our incentive compatibility results are obtained with respect to dominant strategies: truthfulness holds regardless of the other agents' actions. In a sense, this is the strongest incentive compatibility result that could possibly be obtained. We also show that even much weaker truthfulness results cannot be obtained for a wide range of other loss functions, including the popular squared loss.

In the more general case where agents are interested in non-degenerate distributions, achieving incentive compatibility requires more sophisticated mechanisms. We show that the well-known VCG mechanism does very well: with probability $1 - \delta$, no agent can gain more than ϵ by lying, where both ϵ and δ can be made arbitrarily small by increasing the size of the training set. This result holds for any choice of loss function ℓ.

We also study what happens when payments are disallowed. In this setting, we obtain limited positive results for the absolute loss function and for restricted yet interesting function classes. In particular, we present a mechanism which is approximately group strategyproof as above and 3-efficient in the sense that the solution provides a 3-approximation to optimal social welfare. We complement these results with a matching lower bound and provide strong evidence that no approximately incentive compatible and approximately efficient mechanism exists for more expressive functions classes.

Related Work To the best of our knowledge, this paper is the first to study incentives in a general machine learning framework. Previous work in machine learning has investigated the related problem of learning in the presence of inconsistent and noisy training data, where the noise can be either random [13, 7] or adversarial [11, 4]. Barreno *et al.* [2] consider a specific situation where machine learning is used as a component of a computer security system, and account for the possibility that the training data is subject to a strategic attack intended to infiltrate the secured system. In contrast to these approaches, we do not attempt to design algorithms that can tolerate noise, but instead focus on designing algorithms that discourage the strategic addition of noise.

Closely related to our work is the area of *algorithmic mechanism design*, introduced in the seminal work of Nisan and Ronen [17]. Algorithmic mechanism design studies algorithmic problems in a game-theoretic setting where the different participants cannot be assumed to follow the algorithm but rather act in a selfish way. It has turned out that the main challenge of algorithmic mechanism design is the inherent incompatibility of generic truthful mechanisms with approximation schemes for hard algorithmic problems. As a consequence, most of the current work in algorithmic mechanism design focuses on dedicated mechanisms for hard problems (see, *e.g.*, [12, 6]). What distinguishes our setting from that of algorithmic mechanism design is the need for *generalization* to achieve globally satisfactory results on the basis of a small number of samples. Due to the dynamic and uncertain nature of the domain, inputs are usually assumed to be drawn from some underlying fixed distribution. The goal then is to design algorithms that, with high probability, perform well on samples drawn from the same distribution.

More distantly related to our work is research which applies machine learning techniques in game theory and mechanism design. Balcan *et al.* [1], for instance, use techniques from sample complexity to reduce mechanism design problems to standard algorithmic problems. Another line of research puts forward that machine learning can be used to predict consumer behavior, or find a concise description for collective decision making. Work along this line includes the learnability of choice functions and choice correspondences [10, 20, 19].

Structure of the Paper In the following section, we give a general exposition of regression learning and introduce our model of regression learning with multiple agents. We then examine three settings of increasing generality: in Section 3, we consider the case where the distribution of each agent puts all of the weight on a single point of the input space; in Section 4, we then move to the more general setting where the distribution of each agent is a discrete distribution supported on a finite set of points; we finally investigate arbitrary distributions in Section 5, leveraging the results of the previous sections. In Section 6, we discuss our results and give some directions for future research.

We will informally introduce notions from game theory and mechanism design as needed, and refer the reader to the excellent introduction by Nisan [16] for a more formal treatment.

2 The Model

In this section we formalize the regression learning problem described in the introduction and cast it in the framework of game theory. Some of the definitions are illustrated by relating them to the Internet search example presented in the previous section.

We focus on the task of learning a real-valued function over an *input space* $\mathcal{X}$. In the Internet search example, $\mathcal{X}$ would be the set of all query-URL pairs, and our task would be to learn the ranking function of a search engine. Let $N = \{1, \ldots, n\}$ be a set of agents, which in our running example would be the set of all experts. For each agent $i \in N$, let o_i be a function from $\mathcal{X}$ to $\mathbb{R}$ and let ρ_i be a probability distribution over $\mathcal{X}$. Intuitively, o_i is what agent i thinks to be the correct real-valued function, while ρ_i captures the relative importance that agent i assigns to different parts of $\mathcal{X}$. In the Internet search example, o_i would be the optimal ranking function according to agent i, and ρ_i would be a distribution over query-URL pairs that assigns higher weight to queries from that agent's areas of interest.

Let $\mathcal{F}$ be a class of functions, where every $f \in \mathcal{F}$ is a function from $\mathcal{X}$ to the real line. We call $\mathcal{F}$ the *hypothesis space* of our problem, and restrict the output of the learning algorithm to functions in $\mathcal{F}$. We evaluate the accuracy of each $f \in \mathcal{F}$ using a *loss function* $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$. For a particular input-output pair $(\mathbf{x}, y)$, we interpret $\ell(f(\mathbf{x}), y)$ as the penalty associated with predicting the output value $f(\mathbf{x})$ when the true output is known to be y. As mentioned in the introduction, common choices of ℓ are the squared loss, $\ell(\alpha, \beta) = (\alpha - \beta)^2$, and the absolute loss, $\ell(\alpha, \beta) = |\alpha - \beta|$. The accuracy of a hypothesis $f \in \mathcal{F}$ is defined to be the average loss of f over the entire input space. Formally, define the *risk* associated by agent i with the function f as

$$R_i(f) \;=\; \mathbb{E}_{\mathbf{x} \sim \rho_i}[\ell(f(\mathbf{x}), o_i(x))] \;\; .$$

Clearly, this subjective definition of hypothesis accuracy allows for different agents to have significantly different valuations of different functions in $\mathcal{F}$, and it is quite possible that we will not be able to please all of the agents simultaneously. Instead, our goal is to satisfy the agents in N on average. Define J to be a random variable distributed uniformly over the elements of N. Now define the *global risk* of a function f to be the average risk with respect to all of the agents, namely

$$R_N(f) \;=\; \mathbb{E}[R_J(f)] \;\; .$$

We are now ready to formally define our learning-theoretic goal: we would like to find a hypothesis in $\mathcal{F}$ that attains a global risk as close as possible to $\inf_{f \in \mathcal{F}} R_N(F)$.

Even if N is small, we still have no explicit way of calculating $R_N(f)$. Instead, we use an empirical estimate of the risk as a proxy to the risk itself. For each $i \in N$, we randomly sample m points independently from the distribution ρ_i and request their respective labels from agent i. In this way, we obtain the labeled training set $\tilde{S}_i = \{(\mathbf{x}_{i,j}, \tilde{y}_{i,j}\}_{j=1}^m$. Agent i may label the points in $\tilde{S}_i$ however he sees fit, and we therefore say that agent i *controls* (the labels of) these points. We usually denote agent i's "true" training set by $S_i = \{(\mathbf{x}_{ij}, y_{ij})\}_{j=1}^m$, where $y_{ij} = o_i(x_{ij})$. After receiving labels from all agents in N, we define the *global training set* to be the multiset $\tilde{S} = \biguplus_{i \in N} \tilde{S}_i$.

The elicited training set $\tilde{S}$ is presented to a regression learning algorithm, which in return constructs a *hypothesis* $\tilde{f} \in \mathcal{F}$. Each agent can influence $\tilde{f}$ by modifying the labels he controls. This observation brings us to the game-theoretic aspect of our setting. For all $i \in N$, agent i's private information is a vector of true labels $y_{ij} = o_i(x_{ij})$, $j = 1, \ldots, m$. The sampled points x_{ij}, $j = 1, \ldots, m$, are exogenously given and assumed to be common knowledge. The *strategy space* of each agent then consists of all possible *values* for the labels he controls. In other words, agent i reports a labeled training set $\tilde{S}_i$. We sometimes use $\tilde{S}_{-i}$ as a shorthand for $\tilde{S} \setminus \tilde{S}_i$, the strategy profile of all agents except agent i. The space of possible outcomes is the hypothesis space $\mathcal{F}$, and the utility of agent i for an outcome $\tilde{f}$ is determined by his risk $R_i(\tilde{f})$. More precisely, agent i chooses $\tilde{y}_{i1}, \ldots, \tilde{y}_{im}$ so as to minimize $R_i(f)$. We follow the usual game-theoretic assumption that he does this with full knowledge of the inner workings of our regression learning algorithm, and name the resulting game the *learning game*.

One of the simplest and most popular regression learning techniques is *empirical risk minimization* (ERM). The *empirical risk* associated with a hypothesis f, with respect to a sample S, is denoted by $\hat{R}(f, S)$ and defined to be the average loss attained by f on the examples in S, *i.e.*,

$$\hat{R}(f, S) = \frac{1}{|S|} \sum_{(\mathbf{x}, y) \in S} \ell(f(\mathbf{x}), y) \;\; .$$

An ERM algorithm finds the empirical risk minimizer $\hat{f}$ within $\mathcal{F}$. In other words, the ERM algorithm calculates

$$\hat{f} \;=\; \operatorname*{argmin}_{f \in \mathcal{F}} \hat{R}(f, S) \;\; .$$

For some choices of loss function and hypothesis class, it may occur that the global minimizer of the empirical risk is not unique, and we must define an appropriate tie-breaking mechanism. A large part of this paper will be dedicated to ERM algorithms.

Since our strategy is to use $\hat{R}(f, \tilde{S})$ as a surrogate for $R_N(f)$, we need $\hat{R}(f, \tilde{S})$ to be an unbiased estimator of $R_N(f)$. A particular situation in which this can be achieved is when all agents $i \in N$ truthfully report $\tilde{y}_{ij} = o_i(\mathbf{x}_{ij})$ for all j. Another argument for truthfulness regards the quality of the overall solution and can be obtained by a variation of the well-known revelation principle (see, *e.g.*, [16]). Assume that for a given mechanism and given true inputs there

is an equilibrium in which some agents report their inputs untruthfully, and which leads to an outcome that is strictly better than any outcome achievable by an incentive compatible mechanism. Then we can design a new mechanism that, given the true inputs, simulates the agents' lies and yields the exact same output in equilibrium. To summarize, truthful mechanisms will allow us to obtain an unbiased estimator of the true risk, and this need not come at the expense of the overall solution quality.

3 Degenerate Distributions

We begin our study by focusing on a special case, where each agent is only interested in a single point of the input space. Even this simple setting has interesting applications. Consider for example the problem of allocating tasks among service providers, *e.g.*, messages to routers, jobs to remote processors, or reservations of bandwidth to Internet providers. Machine learning techniques are used to obtain a global picture of the capacities, which in turn are private information of the respective providers. Regression learning provides an appropriate model in this context, as each provider is interested in an allocation that is as close as possible to its capacity: more tasks mean more revenue, but an overload is clearly undesirable.

More formally, the distribution ρ_i of agent i is now assumed to be degenerate, and the sample S_i becomes a singleton. Let $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ denote the set of true input-output pairs, where now $y_i = o_i(\mathbf{x}_i)$, and $S_i = \{(\mathbf{x}_i, y_i)\}$ is the single example controlled by agent i. Each agent selects an output value $\tilde{y}_i$, and the reported (possibly untruthful) training set $\tilde{S} = \{(\mathbf{x}_i, \tilde{y}_i)\}_{i=1}^n$ is presented to a regression learning algorithm. The algorithm constructs a hypothesis $\tilde{f}$ and agent i's cost is the loss

$$R_i(\tilde{f}) = \mathbb{E}_{\mathbf{x}\sim\rho_i}[\ell(\tilde{f}(\mathbf{x}), o_i(\mathbf{x}))] = \ell(\tilde{f}(\mathbf{x}_i), y_i)$$

on the point he controls, where ℓ is a predefined loss function. Within this setting, we examine the game-theoretic properties of ERM.

As noted above, an ERM algorithm takes as input a loss function ℓ and a training set S, and outputs the hypothesis that minimizes the empirical risk over S according to ℓ. Throughout this section, we write $\hat{f} = \texttt{ERM}(\mathcal{F}, \ell, S)$ as shorthand for $\arg\min_{f\in\mathcal{F}} \hat{R}(f, \ell, S)$. We restrict our discussion to loss functions of the form $\ell(\alpha, \beta) = \mu(|\alpha - \beta|)$, where $\mu : \mathbb{R}_+ \to \mathbb{R}$ is a monotonically increasing convex function, and to the case where $\mathcal{F}$ is a convex set of functions. These assumptions enable us to cast ERM as a convex optimization problem, which are typically tractable. Most choices of ℓ and $\mathcal{F}$ that do not satisfy the above constraints may not allow for computationally efficient learning, and are therefore less interesting.

We prove two main theorems: if μ is a linear function, then ERM is group strategyproof; if on the other hand μ grows faster than any linear function and if $\mathcal{F}$ contains more than one function, then ERM is not incentive compatible.

3.1 ERM with the absolute loss

In this section, we focus on the absolute loss function. Indeed, let ℓ denote the absolute loss, $\ell(a, b) = |a - b|$, and let $\mathcal{F}$ be a convex hypothesis class. Because ℓ is only weakly convex, there may be multiple hypotheses in $\mathcal{F}$ that globally minimize the empirical risk and we must add a tie-breaking step to our ERM algorithm. Concretely, consider the following two-step procedure:

1. Empirical risk minimization: calculate
$$r = \min_{f\in\mathcal{F}} \hat{R}(f, S).$$
2. Tie-breaking: return
$$\tilde{f} = \operatorname*{argmin}_{f\in\mathcal{F} : \hat{R}(f,S)=r} \|f\|,$$
where $\|f\|^2 = \int f^2(\mathbf{x})\, d\mathbf{x}$.

Our assumption that $\mathcal{F}$ is a convex set implies that the set of empirical risk minimizers $\{f \in \mathcal{F} : \hat{R}(f, S) = r\}$ is also convex. The function $\|f\|$ is a strictly convex function and therefore the output of the tie-breaking step is uniquely defined. In our analysis, we only use the fact that $\|f\|$ is a strictly convex function of f. Any other strictly convex function can be used in its place in the tie-breaking step.

The following theorem states that ERM using the absolute loss function has excellent game-theoretic properties. More precisely, it is *group strategyproof*: if a member of an arbitrary coalition of agents strictly gains from a joint deviation by the coalition, then some other member must strictly lose. Group strategyproofness is usually defined in a way that includes individual rationality. In our case, the latter property is satisfied by any mechanism without payments: if some agent does not provide values for his part of the sample, then ERM will simply return the best fit for the points of the other agents, so no agent can gain by not taking part in the mechanism.

Theorem 3.1. *Let N be a set of agents, $S = \uplus_{i\in N} S_i$ a training set such that $S_i = \{\mathbf{x}_i, y_i\}$ for all $i \in N$, and let ρ_i be degenerate at $\mathbf{x}_i$. Let ℓ denote the absolute loss, $\ell(a, b) = |a - b|$, and let $\mathcal{F}$ be a convex hypothesis class. Then, ERM minimizing ℓ over $\mathcal{F}$ with respect to S is group strategyproof.*

We prove this theorem below, as a corollary of the following more explicit result.

Proposition 3.1. *Let $\hat{S} = \{(\mathbf{x}_i, \hat{y}_i)\}_{i=1}^m$ and $\tilde{S} = \{(\mathbf{x}_i, \tilde{y}_i)\}_{i=1}^m$ be two training sets on the same set of points, and let $\hat{f} = \texttt{ERM}(\mathcal{F}, \ell, \hat{S})$ and $\tilde{f} = \texttt{ERM}(\mathcal{F}, \ell, \tilde{S})$. If $\hat{f} \neq \tilde{f}$ then there exists $i \in N$ such that $\hat{y}_i \neq \tilde{y}_i$ and $\ell(\hat{f}(\mathbf{x}_i), \hat{y}_i) < \ell(\tilde{f}(\mathbf{x}_i), \hat{y}_i)$.*

Proof. Let U be the set of indices on which $\hat{S}$ and $\tilde{S}$ disagree, *i.e.*, $U = \{i : \hat{y}_i \neq \tilde{y}_i\}$. We prove the claim by proving its counter-positive, *i.e.*, we assume that $\ell(\tilde{f}(\mathbf{x}_i), \hat{y}_i) \leq \ell(\hat{f}(\mathbf{x}_i), \hat{y}_i)$ for all $i \in U$, and prove that $\hat{f} \equiv \tilde{f}$. We begin by considering functions of the form $f_\alpha(\mathbf{x}) = \alpha\tilde{f}(\mathbf{x}) + (1-\alpha)\hat{f}(\mathbf{x})$ and proving that there exists $\alpha \in (0, 1]$ for which

$$\hat{R}(\hat{f}, \tilde{S}) - \hat{R}(\hat{f}, \hat{S}) \;=\; \hat{R}(f_\alpha, \tilde{S}) - \hat{R}(f_\alpha, \hat{S}) \;. \tag{3.1}$$

For every $i \in U$, our assumption that $\ell(\tilde{f}(\mathbf{x}_i), \hat{y}_i) \leq \ell(\hat{f}(\mathbf{x}_i), \hat{y}_i)$ implies that one of the following four inequalities holds:

$$\tilde{f}(\mathbf{x}_i) \leq \hat{y}_i < \hat{f}(\mathbf{x}_i) \qquad \tilde{f}(\mathbf{x}_i) \geq \hat{y}_i > \hat{f}(\mathbf{x}_i) \tag{3.2}$$

$$\hat{y}_i \leq \tilde{f}(\mathbf{x}_i) \leq \hat{f}(\mathbf{x}_i) \qquad \hat{y}_i \geq \tilde{f}(\mathbf{x}_i) \geq \hat{f}(\mathbf{x}_i) \tag{3.3}$$

Furthermore, we assume without loss of generality that $\tilde{y}_i = \tilde{f}(\mathbf{x}_i)$ for all $i \in U$. Otherwise, we could simply change $\tilde{y}_i$ to equal $\tilde{f}(\mathbf{x}_i)$ for all $i \in U$ without changing the output of the learning algorithm. If one of the two inequalities in (3.2) holds, we set

$$\alpha_i \;=\; \frac{\hat{y}_i - \hat{f}(\mathbf{x}_i)}{\tilde{f}(\mathbf{x}_i) - \hat{f}(\mathbf{x}_i)} \;,$$

and note that $\alpha_i \in (0, 1]$ and $f_{\alpha_i}(\mathbf{x}_i) = \hat{y}_i$. Therefore, for every $\alpha \in (0, \alpha_i]$ it holds that either

$$\tilde{y}_i \leq \hat{y}_i \leq f_\alpha(\mathbf{x}_i) < \hat{f}(\mathbf{x}_i) \quad \text{or} \quad \tilde{y}_i \geq \hat{y}_i \geq f_\alpha(\mathbf{x}_i) > \hat{f}(\mathbf{x}_i) \;.$$

Setting $c_i = |\hat{y}_i - \tilde{y}_i|$, we conclude that for all α in $(0, \alpha_i]$,

$$\begin{aligned} &\ell(\hat{f}(\mathbf{x}_i), \tilde{y}_i) - \ell(\hat{f}(\mathbf{x}_i), \hat{y}_i) = c_i \quad \text{and} \\ &\ell(f_\alpha(\mathbf{x}_i), \tilde{y}_i) - \ell(f_\alpha(\mathbf{x}_i), \hat{y}_i) = c_i. \end{aligned} \tag{3.4}$$

Alternatively, if one of the inequalities in (3.3) holds, we have that either

$$\hat{y}_i \leq \tilde{y}_i \leq f_\alpha(\mathbf{x}_i) \leq \hat{f}(\mathbf{x}_i) \quad \text{or} \quad \hat{y}_i \geq \tilde{y}_i \geq f_\alpha(\mathbf{x}_i) \geq \hat{f}(\mathbf{x}_i) \;.$$

Setting $\alpha_i = 1$ and $c_i = -|\tilde{y}_i - \hat{y}_i|$, we once again have that (3.4) holds for all α in $(0, \alpha_i]$. Moreover, if we choose $\alpha = \min_{i \in U} \alpha_i$, (3.4) holds simultaneously for all $i \in U$. (3.4) also holds trivially for all $i \notin U$ with $c_i = 0$. (3.1) can now be obtained by summing both of the equalities in (3.4) over all i.

Next, we recall that $\mathcal{F}$ is a convex set and therefore $f_\alpha \in \mathcal{F}$. Since $\hat{f}$ minimizes the empirical risk with respect to $\hat{S}$ over $\mathcal{F}$, we specifically have that

$$\hat{R}(\hat{f}, \hat{S}) \;\leq\; \hat{R}(f_\alpha, \hat{S}) \;. \tag{3.5}$$

Combining this inequality with (3.1) results in

$$\hat{R}(\hat{f}, \tilde{S}) \;\leq\; \hat{R}(f_\alpha, \tilde{S}) \;. \tag{3.6}$$

Since the empirical risk function is convex in its first argument, we have that

$$\hat{R}(f_\alpha, \tilde{S}) \;\leq\; \alpha\hat{R}(\tilde{f}, \tilde{S}) \;+\; (1-\alpha)\hat{R}(\hat{f}, \tilde{S}) \;. \tag{3.7}$$

Replacing the left-hand side above with its lower bound in (3.6) yields $\hat{R}(\hat{f}, \tilde{S}) \leq \hat{R}(\tilde{f}, \tilde{S})$. On the other hand, we know that $\tilde{f}$ minimizes the empirical risk with respect to $\tilde{S}$, and specifically $\hat{R}(\tilde{f}, \tilde{S}) \leq \hat{R}(\hat{f}, \tilde{S})$. Overall, we have shown that

$$\hat{R}(\hat{f}, \tilde{S}) \;=\; \hat{R}(\tilde{f}, \tilde{S}) \;=\; \min_{f \in \mathcal{F}} \hat{R}(f, \tilde{S}) \;. \tag{3.8}$$

Next, we turn our attention to $\|\hat{f}\|$ and $\|\tilde{f}\|$. We start by combining (3.8) with (3.7) to get $\hat{R}(f_\alpha, \tilde{S}) \leq \hat{R}(\hat{f}, \tilde{S})$. Recalling (3.1), we have that $\hat{R}(f_\alpha, \hat{S}) \leq \hat{R}(\hat{f}, \hat{S})$. Once again using (3.5), we conclude that $\hat{R}(f_\alpha, \hat{S}) = \hat{R}(\hat{f}, \hat{S})$. Although $\hat{f}$ and f_α both minimize the empirical risk with respect to $\hat{S}$, we know that $\hat{f}$ was chosen as the output of the algorithm, and therefore it must hold that

$$\|\hat{f}\| \;\leq\; \|f_\alpha\| \;. \tag{3.9}$$

Using convexity of the norm, we have $\|f_\alpha\| \leq \alpha\|\tilde{f}\| + (1-\alpha)\|\hat{f}\|$. Combining this inequality with (3.9), we get $\|\hat{f}\| \leq \|\tilde{f}\|$. On the other hand, (3.8) tells us that both $\hat{f}$ and $\tilde{f}$ minimize the empirical risk with respect to $\tilde{S}$, whereas $\tilde{f}$ is chosen as the algorithm output, so $\|\tilde{f}\| \leq \|\hat{f}\|$. Overall, we have shown that

$$\|\hat{f}\| \;=\; \|\tilde{f}\| \;=\; \min_{f \in \mathcal{F} \,:\, \hat{R}(f, \tilde{S}) = \hat{R}(\tilde{f}, \tilde{S})} \|f\| \;. \tag{3.10}$$

In summary, in (3.8) we showed that both $\hat{f}$ and $\tilde{f}$ minimize the empirical risk with respect to $\tilde{S}$, and therefore both move on to the tie breaking step of the algorithm. Then, in (3.10) we showed that both functions attain the minimum norm over all empirical risk minimizers. Since the norm is strictly convex, its minimum is unique, and therefore $\hat{f} \equiv \tilde{f}$. □

To understand the intuition behind Proposition 3.1, as well as its relation to Theorem 3.1, assume that $\hat{S}$ represents the true preferences of the agents, and that $\tilde{S}$ represents the values revealed by the agents and used to train the regression function. Moreover, assume that $\hat{S} \neq \tilde{S}$. Proposition 3.1 states that one of two things can happen. Either $\hat{f} \equiv \tilde{f}$, *i.e.*, revealing the values in $\tilde{S}$ instead of the true values in $\hat{S}$ does not affect the result of the learning process. In this case, the agents might as well have told the truth. Or $\hat{f}$ and $\tilde{f}$ are different hypotheses, and Proposition 3.1 tells us that there must exist an agent i who lied about his true value and is strictly worse off due to his lie. Clearly, agent i has no incentive to actually participate in such a lie. This said, we can now proceed to prove the theorem.

Proof of Theorem 3.1. Let $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ be a training set that represents the true private information of a set N of agents and let $\tilde{S} = \{(\mathbf{x}_i, \tilde{y}_i)\}_{i=1}^m$ be the information revealed by the agents and used to train the regression function. Let $C \subseteq N$ be an arbitrary coalition of agents that have conspired

to decrease some of their respective losses by lying about their values. Now define the hybrid set of values where

$$\text{for all } i \in N, \quad \hat{y}_i = \begin{cases} y_i & \text{if } i \in C \\ \tilde{y}_i & \text{otherwise} \end{cases} ,$$

and let $\hat{S} = \{(\mathbf{x}_i, \hat{y}_i)\}_{i=1}^m$. Finally, let $\hat{f} = \texttt{ERM}(\mathcal{F}, \ell, \hat{S})$ and $\tilde{f} = \texttt{ERM}(\mathcal{F}, \ell, \tilde{S})$.

If $\hat{f} \equiv \tilde{f}$ then the members of C gain nothing from being untruthful. Otherwise, Proposition 3.1 states that there exists an agent $i \in N$ such that $\hat{y}_i \neq \tilde{y}_i$ and $\ell(\hat{f}(\mathbf{x}_i), \hat{y}_i) < \ell(\tilde{f}(\mathbf{x}_i), \hat{y}_i)$. From $\hat{y}_i \neq \tilde{y}_i$ we conclude that this agent is a member of C. Therefore, $\hat{y}_i = y_i$ and $\ell(\hat{f}(\mathbf{x}_i), y_i) < \ell(\tilde{f}(\mathbf{x}_i), y_i)$. This contradicts our assumption that no member of C loses from revealing $\tilde{S}$ instead of $\hat{S}$. We emphasize that the proof holds regardless of the values revealed by the agents that are not members of C, and we therefore have group strategyproofness. □

3.2 ERM with Other Convex Loss Functions We have seen that performing ERM with the absolute loss is incentive compatible. We now show that the same is not true for most other convex loss functions. Specifically, we examine loss functions of the form $\ell(\alpha, \beta) = \mu(|\alpha - \beta|)$, where $\mu : \mathbb{R}_+ \to \mathbb{R}$ is a monotonically increasing strictly convex function with unbounded subderivatives. Unbounded subderivatives mean that μ cannot be bounded from above by any linear function.

For example, μ can be the function $\mu(\alpha) = \alpha^d$, where d is a real number strictly greater than 1. A popular choice is $d = 2$, which induces the squared loss, $\ell(\alpha, \beta) = (\alpha - \beta)^2$. The following example demonstrates that ERM with the squared loss is not incentive compatible.

Example 1. Let ℓ be the squared loss function, $\mathcal{X} = \mathbb{R}$, and $\mathcal{F}$ the class of constant function over $\mathcal{X}$. Further, let $S_1 = \{(x_1, 2)\}$, and $S_2 = \{x_2, 0\}$. On S, ERM outputs the constant function $\hat{f}(x) \equiv 1$, and agent 1 suffers loss 1. However, if agent 1 reports his value to be 4, ERM outputs $\hat{f}(x) \equiv 2$, with loss of 0 for agent 1.

For every $\mathbf{x} \in \mathcal{X}$, let $\mathcal{F}(\mathbf{x})$ denote the *set of feasible values* of $\mathbf{x}$, formally defined as $\mathcal{F}(\mathbf{x}) = \{ f(\mathbf{x}) : f \in \mathcal{F} \}$. Since $\mathcal{F}$ is a convex set, it follows that $\mathcal{F}(\mathbf{x})$ is either an interval on the real line, a ray, or the entire real line. Similarly, for a multiset $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \in \mathcal{X}^n$, denote

$$\mathcal{F}(X) = \{\langle f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n) \rangle : f \in \mathcal{F}\} \subseteq \mathbb{R}^n .$$

We then say that $\mathcal{F}$ is *full* on a multiset $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \in \mathcal{X}^n$ if $\mathcal{F}(X) = \mathcal{F}(\mathbf{x}_1) \times \cdots \times \mathcal{F}(\mathbf{x}_n)$. Clearly, requiring that $\mathcal{F}$ is not full on X is a necessary condition for the existence of a training set with points X where one of the agents gains by lying. Otherwise, ERM will fit any set of values for the points with an error of zero. For an example of a function class that is *not* full, consider any function class $\mathcal{F}$ on $\mathcal{X}$, $|\mathcal{F}| \geq 2$, and observe that there have to exist $f_1, f_2 \in \mathcal{F}$ and a point $\mathbf{x}_0 \in \mathcal{X}$ such that $f_1(\mathbf{x}_0) \neq f_2(\mathbf{x}_0)$. In this case, $\mathcal{F}$ is not full on any multiset X that contains two copies of $\mathbf{x}_0$.

In addition, if $|\mathcal{F}| = 1$, then any algorithm would trivially be incentive compatible irrespective of the loss function. In the following theorem we therefore consider hypothesis classes $\mathcal{F}$ of size at least two which are *not* full on the set X of points of the training set. The proof of the theorem is given in the full version of the paper. As before, we actually prove a slightly stronger and more explicit claim about the behavior of ERM.

Theorem 3.2. *Let $\mu : \mathbb{R}_+ \to \mathbb{R}$ be a monotonically increasing strictly convex function with unbounded subderivatives, and define the loss function $\ell(\alpha, \beta) = \mu(|\alpha - \beta|)$. Let $\mathcal{F}$ be a convex hypothesis class that contains at least two functions, and let $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \in \mathcal{X}^n$ be a multiset such that $\mathcal{F}$ is not full on X. Then there exist $y_1, \ldots, y_n \in \mathbb{R}$ such that, if $S = \uplus_{i \in N} S_i$ with $S_i = \{(\mathbf{x}_i, y_i)\}$, ρ_i is degenerate at $\mathbf{x}_i$, and ERM is used, there is an agent who has an incentive to lie.*

An example for a function not covered by this theorem is given by $\nu(\alpha) = \ln(1 + \exp(\alpha))$, which is both monotonic and strictly convex, but has a derivative bounded from above by 1. Our definition uses the subderivatives of μ, rather than its derivatives, since we do not require μ to be differentiable.

It is natural to ask what happens for loss functions that are *sublinear* in the sense that they cannot be bounded from below by any linear function with strictly positive derivative. A property of such loss functions, and the reason why they are rarely used in practice, is that the set of empirical risk minimizers need no longer be convex. It is thus unclear how tie-breaking should be defined in order to find a unique empirical risk minimizer. Furthermore, the following example provides a negative answer to the question of general incentive compatibility of ERM with sublinear loss.

Example 2. We demonstrate that ERM is not incentive compatible if $\ell(a, b) = \sqrt{|a - b|}$ and $\mathcal{F}$ is the class of constant functions over $\mathbb{R}$. Let $S = \{(x_1, 1), (x_2, 2), (x_3, 4), (x_4, 6)\}$ and $\tilde{S} = \{(x_1, 1), (x_2, 2), (x_3, 4), (x_4, 4)\}$. Clearly, the local minima of $\hat{R}(f, S)$ and $\hat{R}(f, \tilde{S})$ have the form $f(x) \equiv y$ where $(x_i, y) \in S$ or $(x_i, y) \in \tilde{S}$, respectively, for some $i \in \{1, 2, 3, 4\}$. The empirical risk minimizer for S is the constant function $f_1(x) \equiv 2$, while that for $\tilde{S}$ is $f_2(x) \equiv 4$. Thus, agent 4 can declare his value to be 4 instead of 6 to decrease his loss from $1/2$ to $\sqrt{2}/4$.

4 Uniform Distributions Over the Sample

We now turn to settings where a single agent holds a (possibly) nondegenerate distribution over the input space. However, we still do not move to the full level of generality. Rather, we concentrate on a setting where for each agent i,

ρ_i is the uniform distribution over the sample points x_{ij}, $j = 1, \ldots, m$. While this setting is equivalent to curve fitting with multiple agents and may be interesting in its own right, we primarily engage in this sort of analysis as a stepping stone in our quest to understand the learning game. The results in this section will function as building blocks for the theorems of Section 5.

Since each agent now holds a uniform distribution over his sample, we can simply assume that each agent's cost is his average empirical loss on the sample, $\hat{R}(\tilde{f}, S_i) = 1/m \sum_{j=1}^{m} \ell(\tilde{f}(\mathbf{x}_{ij}), y_{ij})$. The mechanism's goal is to minimize $\hat{R}(\tilde{f}, S)$. We stress at this point that the results in this section also hold if the agents' samples differ in size (this is of course true for the negative results, but also holds for the positive ones). As we move to this more general setting, truthfulness of ERM immediately becomes a thorny issue even under absolute loss. Indeed, the next example indicates that more sophisticated mechanisms must be used to achieve incentive compatibility.

Example 3. Let $\mathcal{F}$ be the class of constant functions over $\mathbb{R}^k$, $N = \{1, 2\}$, and assume the absolute loss function is used. Let $S_1 = \{(1, 1), (2, 1), (3, 0)\}$ and $S_2 = \{(4, 0), (5, 0), (6, 1)\}$. The global empirical risk minimizer (according to our tie-breaking rule) is the constant function $f_1(x) \equiv 0$ with $\hat{R}(f_1, S_1) = 2/3$. However, if agent 1 declares $\tilde{S}_1 = \{(1, 1), (2, 1), (3, 1)\}$, then the empirical risk minimizer becomes $f_2(x) \equiv 1$, which is the optimal fit for agent 1 since $\hat{R}(f_2, S_1) = 1/3$.

4.1 Mechanisms with Payments One possibility to overcome the issue that became manifest in Example 3 is to consider mechanisms that not only return an allocation, but can also transfer payments to and from the agents based on the inputs they provide. A famous example for such a payment rule is the Vickrey-Clarke-Groves (VCG) mechanism [23, 5, 8]. This mechanism starts from an efficient allocation, and computes each agent's payment according to the utility of the other agents, thus aligning the individual interests of each agent with that of society.

In our setting, where social welfare equals the total empirical risk, ERM generates a function (or outcome) which maximizes social welfare and can therefore be directly augmented with VCG payments. Given an outcome $\hat{f}$, each agent i has to pay an amount of $\hat{R}(\hat{f}, \tilde{S}_{-i})$. In turn, the agent can receive some amount $h_i(\tilde{S}_{-i})$ that does *not* depend on the values he has reported, but possibly on the values reported by the other agents. It is well known [8], and also easily verified, that this family of mechanisms is incentive compatible: no agent is motivated to lie regardless of the other agents' actions. Furthermore, this result holds for any loss function, and may thus be an excellent solution for some settings.

In many other settings, however, especially in the world of the Internet, transferring payments to and from users can pose serious problems, up to the extent that it might become completely infeasible. The practicality of VCG payments in particular has recently been disputed for various other reasons [21]. Perhaps most relevant to our work is the fact that VCG mechanisms are in general susceptible to manipulation by coalitions of agents and thus not group strategyproof. It is therefore worthwhile to explore which results can be obtained when payments are disallowed. This will be the subject of the following section.

4.2 Mechanisms without Payments In this section, we restrict ourselves to the absolute loss function. When ERM is used, and for the special case covered in Section 3, this function was shown to possess properties far superior to any other loss function. This fuels hope that similar incentive compatibility results can be obtained with uniform distributions over the samples, even when payments are disallowed. This does not necessarily mean that good mechanisms without payments cannot be designed for other loss functions, even in the more general setting of this section. We leave the study of such mechanisms for future work.

ERM is *efficient*, *i.e.*, it minimizes the overall loss and maximizes social welfare. In light of Example 3, we shall now sacrifice efficiency for incentive compatibility. More precisely, we seek incentive compatible mechanisms which are *approximately efficient* in the sense that for all samples S, the ratio $\hat{R}(f, S)/\hat{R}(\hat{f}, S)$ between the empirical risk of the solution f returned by the mechanism and that of the optimal solution $\hat{f}$ is bounded. We say that a regression learning mechanism is α-efficient if for all S, $\hat{R}(f, S)/\hat{R}(\hat{f}, S) \leq \alpha$. We should stress that the reason we resort to approximation is not to make the mechanism computationally tractable, but to achieve incentive compatibility without payments, like we had in Section 3.

Example 3, despite its simplicity, is surprisingly robust against many conceivably truthful mechanisms. The reader may have noticed, however, that the values of the agents in this example are not "individually realizable": in particular, there is no constant function which *realizes* agent 1's values, *i.e.*, fits them with a loss of zero. In fact, agent 1 benefits from revealing values which are consistent with his individual empirical risk minimizer. This insight leads us to design the following simple but useful mechanism, which we will term "project-and-fit":

Input: A hypothesis class $\mathcal{F}$ and a sample $S = \uplus S_i$, $S_i \subseteq \mathcal{X} \times \mathbb{R}$
Output: A function $f \in \mathcal{F}$.
Mechanism:

1. For each $i \in N$, let $f_i = \texttt{ERM}(\mathcal{F}, S_i)$.
2. Define $\tilde{S}_i = \{(\mathbf{x}_{i1}, f_i(x_{i1})), \ldots, (\mathbf{x}_{im}, f_i(x_{im}))\}$.
3. Return $f = \texttt{ERM}(\tilde{S})$, where $\tilde{S} = \uplus_{i=1}^{n} \tilde{S}_i$.

In other words, the mechanism calculates the individual empirical risk minimizer for each agent and uses it to relabel the agent's sample. Then, the relabeled samples are combined, and ERM is performed. It should be noted that this mechanism achieves group strategyproofness at least with respect to Example 3.

More generally, it can be shown that the mechanism is group strategyproof when $\mathcal{F}$ is the class of constant functions over $\mathbb{R}^k$. Indeed, it is natural to view our setting through the eyes of social choice theory [14]: agents entertain (weak) preferences over a set of alternatives, *i.e.*, the functions in $\mathcal{F}$. In the case of constant functions, agents' preferences are what is known as *single-plateau* [15]: each agent has an interval of ideal points minimizing his individual empirical risk, and moving away from this plateau in either direction strictly decreases the agent's utility. More formally, let a_1, a_2 be constants such that the constant function $f(x) \equiv a$ minimizes an agent's empirical risk if and only if $a \in [a_1, a_2]$. If a_3, a_4 satisfy $a_3 < a_4 \leq a_1$ or $a_3 > a_4 \geq a_2$, then the agent strictly prefers the constant function a_4 to the constant function a_3. As such, single-plateau preferences generalize the class of single-peaked preferences. For dealing with single-plateau preferences, Moulin [15] defines the class of generalized Condorcet winner choice functions, and shows that these are group strategyproof.

When $\mathcal{F}$ is the class of constant functions and ℓ is the absolute loss, the constant function equal to a median value in a sample S minimizes the empirical risk with respect to S. This is because there must be at least as many values below the median value as are above, and thus moving the fit upward (or downward) must monotonically increase the sum of distances to the values. Via tie-breaking, project-and-fit essentially turns the single-plateau preferences into single-peaked ones, and then chooses the median peak. Once again, group strategyproofness follows from the fact that an agent can only change the mechanism's output by increasing its distance from his own empirical risk minimizer.

Quite surprisingly, project-and-fit is not only truthful but also provides a constant approximation ratio when $\mathcal{F}$ is the class of constant functions or the class of homogeneous linear functions over $\mathbb{R}$, *i.e.*, functions of the form $f(\mathbf{x}) = \mathbf{a} \cdot \mathbf{x}$. The class of homogeneous linear functions, in particular, is important in machine learning, for instance in the context of Support Vector Machines [22]. The proof of the following theorem is given in the full version of the paper.

Theorem 4.1. *Assume that $\mathcal{F}$ is the class of constant functions over $\mathbb{R}^k$, $k \in \mathbb{N}$ or the class of homogeneous linear functions over $\mathbb{R}$. Then project-and-fit is group strategyproof and* 3*-efficient.*

A simple example shows that the 3-efficiency analysis given in the proof is tight. We generalize this observation by proving that, for the class of constant or homogeneous linear functions and irrespective of the dimension of $\mathcal{X}$, no truthful mechanism without payments can achieve an efficiency ratio better than 3. It should be noted that this lower bound holds for any choice of points x_{ij}. The proof is again deferred to the full version of the paper.

Theorem 4.2. *Let $\mathcal{F}$ be the class of constant functions over $\mathbb{R}^k$ or the class of homogeneous linear functions over $\mathbb{R}^k$, $k \in \mathbb{N}$. Then there exists* no *incentive compatible mechanism without payments that is* $(3 - \epsilon)$*-efficient for any* $\epsilon > 0$*, even when* $|N| = 2$.

Let us recapitulate. We have found a group strategyproof and 3-efficient mechanism for the class of constant functions over $\mathbb{R}^k$ and for the class of homogeneous linear functions over $\mathbb{R}$. A matching lower bound, which also applies to multi-dimensional homogeneous linear functions, shows that this result cannot be improved upon for these classes. It is natural to ask at this point if project-and-fit remains incentive compatible when considering more complex hypothesis classes, such as homogeneous linear functions over $\mathbb{R}^k$, $k \geq 2$, or linear functions. An example serves to answer this question in the negative.

Is there some other mechanism which deals with more complex hypothesis classes and provides a truthful approximation? We believe that this is *not* the case even for the class of homogeneous linear functions over $\mathbb{R}^k$ for any $k \geq 2$. The following conjecture formalizes this statement. Arguments supporting the conjecture can be found in the full version of the paper.

Conjecture 4.1. *Let $\mathcal{F}$ be the class of homogeneous linear functions over $\mathbb{R}^k$, $k \geq 2$, and assume that $m = |S_i| \geq 3$. Then any incentive compatible and surjective mechanism is a dictatorship.*

Conceivably, dictatorship would be an acceptable solution if it could guarantee approximate efficiency. A simple example shows that unfortunately this is not the case.

5 Arbitrary Distributions Over the Sample

In Section 4 we established several positive results in the setting where each agent cares about a uniform distribution on his portion of a global training set. In this section we extend these results to the general regression learning setting defined in Section 2. More formally, the extent to which agent $i \in N$ cares about each point in $\mathcal{X}$ will now be defined by the distribution function ρ_i, and agent i controls the labels of a finite set of points sampled according to ρ_i. Our strategy in this section will consist of two steps. First, we want to show that under standard assumptions on the hypothesis class $\mathcal{F}$ and the number m of samples, each agent's empirical risk on the training set S_i estimates his real risk according to ρ_i. Second, we intend to establish that,

as a consequence, our incentive compatibility results are not significantly weakened when moving to the general setting.

Abstractly, let $\mathcal{D}$ be a probability distribution and let $\mathcal{G}$ be a class of real-valued functions, bounded in $[0, C]$. We would like to prove that for any $\epsilon > 0$ and $\delta > 0$ there exists $m \in \mathbb{N}$ such that, if $X_1, \ldots, X_m$ are sampled i.i.d. according to $\mathcal{D}$,

(5.11)
$$\Pr\left(\text{for all } g \in \mathcal{G}, \left|\mathbb{E}_{X\sim\mathcal{D}}[g(X)] - \frac{1}{m}\sum_{i=1}^{m} g(X_i)\right| \le \epsilon\right) \ge 1 - \delta.$$

To establish this bound, we use standard *uniform convergence* arguments. A specific technique is to show that the hypothesis class $\mathcal{G}$ has bounded complexity. The complexity of $\mathcal{G}$ can be measured in various different ways, for example using the pseudo-dimension [18, 9], an extension of the well-known VC-dimension to real-valued hypothesis classes, or the Rademacher complexity [3]. If the pseudo-dimension of $\mathcal{G}$ is bounded by a constant, or if the Rademacher complexity of $\mathcal{G}$ with respect to an m-point sample is $O(\sqrt{m})$, then there indeed exists m such that (5.11) holds.

More formally, assume that the hypothesis class $\mathcal{F}$ has bounded complexity, choose $\epsilon > 0$, $\delta > 0$, and consider a sample S_i of size $m = \Theta(\log(1/\delta)/\epsilon^2)$ drawn i.i.d. from the distribution ρ_i of any agent $i \in N$. Then we have that

(5.12)
$$\Pr\left(\text{for all } f \in \mathcal{F},\ \left|R_i(f) - \hat{R}(f, S_i)\right| \le \epsilon\right) \ge 1 - \delta\ .$$

In particular, we want the events in (5.12) to hold simultaneously for all $i \in N$, *i.e.*,

$$\text{for all } f \in \mathcal{F},\ \left|R_N(f) - \hat{R}(f, S)\right| \le \epsilon\ . \tag{5.13}$$

Using the union bound, this holds with probability at least $1 - n\delta$.

We now turn to incentive compatibility. The following theorem implies that mechanisms which do well in the setting of Section 4 are also good, but slightly less so, when arbitrary distributions are allowed. Specifically, given a training set satisfying (5.12) for all agents, a mechanism that is incentive compatible in the setting of Section 4 becomes ϵ-incentive compatible, *i.e.*, no agent can gain more than ϵ by lying, no matter what the other agents do. Analogously, a group strategyproof mechanism for the setting of Section 4 becomes ϵ-group strategyproof, *i.e.*, there exists an agent in the coalition that gains less than ϵ. Furthermore, efficiency is preserved up to an additive factor of ϵ. We wish to point out that ϵ-equilibrium is a well-established solution concept; the underlying assumption is that agents would not bother to lie if they were to gain an amount as small as ϵ. This concept is particularly appealing when one recalls that ϵ can be chosen to be arbitrarily small. The proof of the following theorem can be found in the full version of the paper.

Theorem 5.1. *Let $\mathcal{F}$ be a hypothesis class, ℓ some loss function, and $S = \uplus S_i$ a training set such that for all $f \in \mathcal{F}$ and $i \in N$, $|R_i(f) - \hat{R}(f, S_i)| \le \epsilon/2$, and $|R_N(f) - \hat{R}(f, S)| \le \epsilon/2$. Let M be a mechanism with or without payments.*

1. *If M is incentive compatible (group strategyproof, respectively) under the assumption that each agent's cost is $\hat{R}(\tilde{f}, S_i)$, then M is ϵ-incentive compatible (ϵ-group strategyproof) in the general regression setting.*
2. *If M is α-efficient under the assumption that the mechanism's goal is to minimize $\hat{R}(\tilde{f}, S)$, $M(S) = \tilde{f}$, then $R_N(\tilde{f}) \le \alpha \cdot \operatorname{argmin}_{f\in\mathcal{F}} R_N(f) + \epsilon$.*

As discussed above, the conditions of Theorem 5.1 are satisfied with probability $1 - \delta$ when $\mathcal{F}$ has bounded dimension and $m = \Theta(\log(1/\delta)/\epsilon^2)$. As the latter expression depends logarithmically on $1/\delta$, the sample size only needs to be increased by an additive factor of $\Theta(\log(n)/\epsilon^2)$ to achieve the stronger requirement of (5.13).

Let us examine how Theorem 5.1 applies to our positive results. Since ERM with VCG payments is incentive compatible and efficient under uniform distributions over the samples, we obtain ϵ-incentive compatibility and efficiency up to an additive factor of ϵ when it is used in the general learning game, *i.e.*, with arbitrary distributions. This holds for any loss function ℓ. The project-and-fit mechanism is ϵ-group strategyproof in the learning game when $\mathcal{F}$ is the class of constant functions or of homogeneous linear functions over $\mathbb{R}$, and is 3-efficient up to an additive factor of ϵ. This is true only for the absolute loss function.

6 Discussion

In this paper, we have studied mechanisms for a general regression learning framework involving multiple strategic agents. In the case where each agent controls one point, we have obtained a strong and surprising characterization of the truthfulness of ERM. When the absolute loss function is used, ERM is group strategyproof. On the other hand, ERM is not even individually incentive compatible for any loss function that is superlinear in a certain well-defined way. This particularly holds for the popular squared loss function. In the general learning setting, we have established the following result: For any $\epsilon, \delta > 0$, given a large enough training set, and with probability $1 - \delta$, ERM with VCG payments is efficient up to an additive factor of ϵ, and ϵ-incentive compatible. We have also obtained limited positive results for the case when payments are disallowed, namely an algorithm that is ϵ-group strategyproof and 3-efficient up to an additive factor of ϵ for constant functions over $\mathbb{R}^k$, $k \in \mathbb{N}$, and for homogeneous linear functions over $\mathbb{R}$. We also gave a matching lower bound, which also applies to multidimensional homogeneous linear functions. The number of samples required by the aforementioned

algorithms depends on the combinatorial richness of the hypothesis space $\mathcal{F}$, but differs only by an additive factor of $\Theta(\log(n)/\epsilon^2)$ from that in the traditional regression learning setting without strategic agents. Since $\mathcal{F}$ can be assumed to be learnable in general, this factor is not very significant.

Since at the moment there is virtually no other work on incentives in machine learning, many exciting directions for future work exist. While regression learning constitutes an important area of machine learning with numerous applications, adapting our framework for studying incentives in classification or in unsupervised settings will certainly prove interesting as well. In classification, each point of the input space is assigned one of two labels, either $+1$ or -1. ERM is trivially incentive compatible in classification when each agent controls only a single point. The situation again becomes complicated when agents control multiple points. In addition, we have not considered settings where ERM is computationally intractable. Just like in general algorithmic mechanism design, VCG is bound to fail in this case. It is an open question whether one can simultaneously achieve tractability, approximate efficiency, and (approximate) incentive compatibility. Several interesting questions follow directly from our work. The one we are most interested in is settling Conjecture 4.1: are there incentive compatible and approximately efficient mechanisms without payments for homogeneous linear functions? Do such mechanisms exist for other interesting hypothesis classes? These questions are closely related to general questions about the existence of incentive compatible and non-dictatorial mechanisms, and are interesting well beyond the scope of machine learning and computer science.

Acknowledgements

We thank Jeff Rosenschein for helpful discussions, and Bezalel Peleg for sharing his views on Conjecture 4.1. We are also grateful to an anonymous referee for valuable comments.

References

[1] M.-F. Balcan, A. Blum, J. D. Hartline, and Y. Mansour. Mechanism design via machine learning. In *Proceedings of the 46th Symposium on Foundations of Computer Science (FOCS)*, pages 605–614. IEEE Press, 2005.

[2] M. Barreno, B. Nelson, R. Sears, A. D. Joseph, and J. D. Tygar. Can machine learning be secure? In *Proceedings of the 1st ACM Symposium on Information, Computer and Communications Security (ASIACCS)*, pages 16–25. ACM Press, 2006.

[3] P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2003.

[4] N. H. Bshouty, N. Eiron, and E. Kushilevitz. PAC learning with nasty noise. *Theoretical Computer Science*, 288(2):255–275, 2002.

[5] E. H. Clarke. Multipart pricing of public goods. *Public Choice*, 11:17–33, 1971.

[6] S. Dobzinski, N. Nisan, and M. Schapira. Truthful randomized mechanisms for combinatorial auctions. In *Proceedings of the 38th Annual ACM Symposium on the Theory of Computing (STOC)*, pages 644–652, 2006.

[7] S. A. Goldman and R. H. Sloan. Can PAC learning algorithms tolerate random attribute noise? *Algorithmica*, 14(1):70–84, 1995.

[8] T. Groves. Incentives in teams. *Econometrica*, 41:617–631, 1973.

[9] D. Haussler. Decision theoretic generalization of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992.

[10] G. Kalai. Learnability and rationality of choice. *Journal of Economic Theory*, 113(1):104–117, 2003.

[11] M. Kearns and M. Li. Learning in the presence of malicious errors. *SIAM Journal on Computing*, 22(4):807–837, 1993.

[12] D. Lehmann, L. I. O'Callaghan, and Y. Shoham. Truth revelation in rapid, approximately efficient combinatorial auctions. *Journal of the ACM*, 49(5):577–602, 2002.

[13] N. Littlestone. Redundant noisy attributes, attribute errors, and linear-threshold learning using Winnow. In *Proceedings of the 4th Annual Workshop on Computational Learning Theory (COLT)*, pages 147–156, 1991.

[14] A. Mas-Colell, M. D. Whinston, and J. R. Green. *Microeconomic Theory*. Oxford University Press, 1995.

[15] H. Moulin. Generalized Condorcet-winners for single peaked and single-plateau preferences. *Social Choice and Welfare*, 1(2):127–147, 1984.

[16] N. Nisan. Introduction to mechanism design (for computer scientists). In N. Nisan, T. Roughgarden, E. Tardos, and V. Vazirani, editors, *Algorithmic Game Theory*, chapter 9. Cambridge University Press, 2007. To Appear.

[17] N. Nisan and A. Ronen. Algorithmic mechanism design. In *Proceedings of the 31st Annual ACM Symposium on the Theory of Computing (STOC)*, pages 129–140. ACM Press, 1999.

[18] D. Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, 1984.

[19] A. D. Procaccia, A. Zohar, Y. Peleg, and J. S. Rosenschein. Learning voting trees. In *Proceedings of the 22nd Conference on Artificial Intelligence (AAAI)*, pages 110–115, 2007.

[20] A. D. Procaccia, A. Zohar, and J. S. Rosenschein. Automated design of voting rules by learning from examples. In *Proceedings of the 1st International Workshop on Computational Social Choice (COMSOC)*, pages 436–449, 2006.

[21] M. Rothkopf. Thirteen reasons the Vickrey-Clarke-Groves process is not practical. *Operations Research*, 55(2):191–197, 2007.

[22] J. Shawe-Taylor and N. Cristianini. *Support Vector Machines and other Kernel Based Learning Methods*. Cambridge University Press, 2000.

[23] W. Vickrey. Counter speculation, auctions, and competitive sealed tenders. *Journal of Finance*, 16(1):8–37, 1961.