Algorithms for Strategyproof Classification

Reshef Meir [*†]
reshef.meir@mail.huji.ac.il

Ariel D. Procaccia[‡]
arielpro@cs.cmu.edu

Jeffrey S. Rosenschein[§]
jeff@cs.huji.ac.il

March 10, 2012

Abstract

The strategyproof classification problem deals with a setting where a decision maker must classify a set of input points with binary labels, while minimizing the expected error. The labels of the input points are reported by self-interested agents, who might lie in order to obtain a classifier that more closely matches their own labels, thereby creating a bias in the data; this motivates the design of *truthful* mechanisms that discourage false reports.

In this paper we give strategyproof mechanisms for the classification problem in two restricted settings: (i) there are only two classifiers, and (ii) all agents are interested in a *shared* set of input points. We show that these plausible assumptions lead to strong positive results. In particular, we demonstrate that variations of a random dictator mechanism, that are truthful, can guarantee approximately optimal outcomes with respect to *any* family of classifiers. Moreover, these results are tight in the sense that they match the best possible approximation ratio that can be guaranteed by any truthful mechanism.

We further show how our mechanisms can be used for learning classifiers from sampled data, and provide PAC-style generalization bounds on their expected error. Interestingly, our results can be applied to problems in the context of various fields beyond classification, including facility location and judgment aggregation.

Keywords: Mechanism design, Classification, Game theory, Approximation.

1 Introduction

Consider a learning algorithm, which takes a labeled set of samples ("training data") as input, and outputs a binary classifier. The training data, typically hand-constructed

[*]Corresponding author.

[†]School of Engineering and Computer Science, Hebrew University, Jerusalem, Israel 91904. Tel. 972-2-6585188.

[‡]Computer Science Department, Carnegie Mellon University, 5000 Forbes, Pittsburgh, PA 15213.

[§]School of Engineering and Computer Science, Hebrew University, Jerusalem, Israel 91904. Tel. 972-2-6585353.

by human experts, is supposed to reflect the knowledge of the experts on the current domain. The basic requirement from such an algorithm is to guarantee that the output classifier minimizes the number of classification errors with respect to the 'truth' (according to the domain experts). Standard machine-learning literature studies the performance of such algorithms given various distributions and concept classes (e.g., linear classifiers), sparse or noisy data, etc.

However in many real-life situations, the experts have a personal interest in the outcome of the algorithm, and therefore they cannot be assumed to be truthful. If an expert can bias the learned classifier in her favor by lying, then the reported training data will no longer reflect the properties of the domain (or even the properties of the real training data). Optimizing a classifier based on such corrupted data may result in a very poor classifier, regardless of the guarantees supplied by learning theory (which assumes truthfulness).

We consider two interrelated settings. The first setting is *decision-theoretic*; a decision must be made based on data reported by multiple self-interested agents. The agents are concerned with the binary labels of a set of input points. Put another way, the agents may disagree on the labels of the points of the input space, and we do not assume any underlying distribution. The utility of an agent with respect to a given decision (i.e., a given classifier) is the number of points on which the label provided by the classifier agrees with the agent's own label. The goal of the decision maker is to choose a classifier that maximizes the social welfare—the sum of utilities. As we will see, results in this setting can also be applied to problems in the context of various other fields, including facility location and judgment aggregation.

The second setting is *learning-theoretic*, a variation of the standard Supervised Classification problem. Samples are drawn from some distribution over the input space, and are then labeled by experts. A classification mechanism receives the sampled data as input, and outputs a classifier. Unlike the standard setting in machine learning (but similarly to our first setting), the experts are assumed to be self-interested agents, and may lie in order to increase their utility. This setting may seem far more involved than the first, as it deals with generalization from partial data (the dataset) to the underlying distribution. However, we show that under the standard assumptions of learning theory, the learning problem effectively reduces to finding a classifier that best fits the available data (i.e., to the first setting, above).

In both settings the decision maker (or mechanism, or learning algorithm) aims to find a classifier that classifies the available data as well as possible. However, the agents may misreport their labels in an attempt to influence the final decision in their favor. The result of a decision making process based on such biased data may be completely unexpected and difficult to analyze. A *truthful* learning mechanism eliminates any such bias and allows the decision maker to select a classifier that best fits the reported data, without having to take into account the hidden interests of the agents. In other words, once we guarantee that agents are telling the truth, we may concentrate on the more standard goal of minimizing the error. In order to obtain truthfulness, however, we may need to trade off optimality. Our goal is to provide mechanisms that are both truthful and approximately optimal in terms of social welfare.

1.1 Restrictions on the domain

In recent work [29] we showed that in an unrestricted domain, it is effectively impossible to design truthful mechanisms that are close to optimal. This motivates the investigation of restricted domains. In this paper we consider several such restrictions, described below.

1.1.1 Restricting the concept class: two functions

A seemingly simple case is when the concept class contains only two functions. This is equivalent to a (binary) decision that has to be made based on data points that are controlled by multiple (possibly) selfish agents, where the decision affects all the agents. The decision maker would like to make a decision which is consistent, as much as possible, with all the available data. However, in our strategic setting the agents might misreport their data in an attempt to influence the final decision in their favor.

As a motivating example, consider a decision that has to be made by the Workers' committee of the TAs in the Hebrew university, regarding an ongoing strike. Each member of the committee (who represents one department) announces how many TAs in his/her department are supporting the strike, and how many oppose it. A final decision is being made based on the total support of the strike. Suppose that 60% of the economics department is opposing the strike. However, the representative of the economics department is majoring in game theory. Therefore she knows that for the benefit of the majority of TAs *in her department*, it would be better to state that everybody objects to the strike.[1]

1.1.2 Restricting the dataset: shared inputs

Our main conceptual contribution in this paper, which leads to strong positive results, is the assumption of *shared inputs*. In the decision-theoretic setting, this means that the agents share the same set of input points, and only disagree on the labels of these points. In the learning-theoretic setting, the shared inputs assumption implies that the agents are interested in a common distribution over the input space, but, once again, differ with respect to the labels.

The first restriction we described did not address the issue of shared inputs. However, as the two possible classifiers are constant, the identity of the input points (i.e., their location) is irrelevant—only their labels matter. Hence, the first restriction is in fact a very special case of the latter (see also footnote 16).

As the shared inputs assumption is a weaker restriction than assuming two functions, the guarantees are also somewhat weaker. Nevertheless, they hold with respect to *any concept class*. We believe that in many environments the requirement of shared inputs is satisfied. As an example, consider a large organization that is trying to fight congestion in an internal email system by designing a smart spam filter. In order to train the system, managers are asked to review the last 1000 emails sent to the "all employees" mailing list (hence, shared inputs) and classify them as either "work-related"

[1] In an attempt to avoid such misrepresentation, major decisions usually require to gather all TAs and hold a standard voting procedure. However most decisions are taken in a much narrower quorum.

(positive label) or "spam" (negative label). Whereas the managers will likely agree on the classification of some of the messages (e.g., "Buy Viagra now!!!" or "Christmas Bonus for all employees"), it is likely that others (e.g., "Joe from the Sales department goes on a lunch break") would not be unanimously classified. Moreover, as each manager is interested in filtering most of what he sees as spam, a manager might try to compensate for the "mistakes" of his colleagues by misreporting his real opinion with respect to some cases. For example, the manager of the R&D department, believing that about 90% of the Sales messages are utterly unimportant, might classify *all* of them as spam in order to reduce the congestion. The manager of Sales, suspecting the general opinion on her department, might do the exact opposite to prevent her e-mails from being filtered. The fact that some users may not have a full understanding of the learning algorithm, does not necessarily prevent them from trying to bias it anyway. Even if their strategy is not optimal for them, it still contaminates the data.

Interestingly, our model for binary classification with shared inputs is equivalent to models that have been suggested in the literature for problems in seemingly unrelated domains, including judgment aggregation, partition aggregation, facility location, and voting (for a more detailed comparison, see Section 1.3 and discussion).

Such a common classification/partition problem is deciding on the operation hours of a shared resource. As a concrete example, consider a building with a central heating system (such buildings are common in Jerusalem and many cities in Europe). Every tenant has certain hours in which he wants the heat to be on (e.g. always **on** when he is home and **off** otherwise, since the cost is shared by all tenants). The household fee is the same for all tenants, and thus there is no transfer of payoffs. A "classifier" is a partition of the day (or week) to **on** and **off** intervals. Further, there are constraints on the final partition. For example, **on** intervals must be at least 3 hours long to achieve better efficiency.

1.1.3 Realizable datasets

In some cases, learning is facilitated if we know that there is at least one "perfect" classifier in our concept class (that is, a classifier that separates all positive data points in the dataset from the negative ones). Such datasets are called *realizable*. It is therefore possible that the labels of each agent will be realizable, even if there is no single classifier that is perfect for all agents. We study how realizability, which can be seen as another restriction on the dataset, affects the optimality of the proposed mechanisms in the context of shared input.

1.2 Overview of our results

We wish to design classification mechanisms that achieve a good outcome in the face of strategic behavior. By "good outcome" we mean that the output of the mechanism provides an approximation of the optimal solution.[2] We would also like our mechanisms to be *strategyproof* (SP), that is, the agents must not be able to benefit from

[2]Approximation algorithms are frequently used in various domains in computer science in order to overcome computational barriers. While we largely ignore issues of computational complexity, optimal algorithms are typically not strategyproof; hence, the need for approximation.

lying. These two key requirements are formalized and demonstrated with examples in Section 2.

We begin by presenting mechanisms for the two-function problem in Section 3. The results of this section serve two purposes. First, the tight worst-case analysis of SP mechanisms provides a full picture of their power and limitations in the binary decision-making setting. Second, the focus on a simple setting allows us to explain in detail subtle issues that are also important for the next, more general, setting.

We put forward a simple deterministic decision-making mechanism which is group strategyproof (i.e., even coalitions of agents do not gain from lying) and gives a 3-approximation of the optimal global risk; in other words, the number of mislabeled points is at most 3 times the minimum number. Moreover, we show that no deterministic strategyproof mechanism can do better. Interestingly, we circumvent this result by designing a strategyproof *randomized* mechanism that gives a 2-approximation, and further demonstrate that this is as far as randomization can take us.

In Section 4, we turn to study the more general case, under the shared inputs assumption. We first show that SP deterministic mechanisms cannot guarantee a sublinear approximation ratio. We show that choosing a dictator at random provides an approximation ratio of 3 in expectation, even if agents have weights, i.e., the decision mechanism values some agents more than others (in that case we randomly select a dictator according to the weights). We then drive the approximation even lower by using a non-trivial selection of the dictator, matching it with the known lower bound of $3 - \frac{2}{n}$; it is quite striking that these results hold with respect to any concept class. In addition, we show that when datasets are realizable, an even better approximation ratio (of $2 - \frac{2}{n}$) can be guaranteed.

In each section we further show how the suggested mechanisms for the decision-theoretic setting can be further exploited to attain similar approximation results in the learning-theoretic setting. We observe that in the learning-theoretic setting, designing strategyproof mechanisms is virtually impossible, since there is an additional element of randomness introduced by sampling the input space. We therefore relax the strategyproof requirements, and instead investigate each of two incomparable strategic assumptions: that agents do not lie if they cannot gain more than ϵ; and that agents always use a dominant strategy if one exists with respect to a specific sample. We show that under either assumption our randomized mechanisms can be run directly on sampled data while maintaining a bounded expected error. Our theorems give a connection between the number of samples and the expected error of the mechanism in each case, in the spirit of PAC-learning algorithms [40].

1.2.1 Mechanisms with payments

An important remark is that in the strategyproof classification setting, standard economic money-based mechanisms such as the Vickrey-Clarke-Groves (VCG) mechanism (see, e.g., [32]) can be used to obtain good results. However, our setting admits strategyproof mechanisms that do well *even without assuming that money is available*. Achieving our goals without resorting to payments is highly desirable, since often payments cannot be made due to legal or ethical considerations. Moreover, in internet environments VCG style payments are notoriously difficult to implement, due

to banking and security issues. Hence, we follow the example set by previous work on strategyproof learning models (e.g., [10], see below) by considering approximation mechanisms that do not require payments.

1.3 Related work

This paper lies at the intersection of several areas, including mechanism design, judgment aggregation, and learning. We cluster the related work by areas.

1.3.1 Approximate mechanism design without money

Mechanisms that deal with strategic behavior of agents have been proposed recently for a large range of applications. While certain restrictions may allow the design of optimal SP mechanisms [39], often this is not the case, and approximation is required. This observation gave rise to the agenda of approximate mechanism design without money (AMDw/oM).

Below, we overview some SP mechanisms for machine learning problems in detail, and compare them to our work. These, however, constitute just one facet of the large variety of problems to which AMDw/oM can be applied. Approximate mechanisms without payments have been proposed for facility location, matching [3, 15], resource allocation [18, 19, 33], scheduling [23], and even auctions [20].

1.3.2 Strategyproof learning algorithms

The work most closely related to ours is a paper by Dekel et al. [10] Their work focused on regression learning, where the labels are real numbers and one is interested in the *distances* between the mechanism's outputs and the labels. Except for this very significant difference, the settings that we study and our goals are very similar to theirs. Dekel et al. provided upper and lower bounds on the approximation ratio achieved by supervised regression mechanisms in this model. Notably, some of our bounds resemble the bounds in their regression setting. Moreover, similar intuitions sometimes apply to both settings, although it seems the results of one setting cannot be analytically mapped to the other. Dekel et al. also concentrate on mechanisms without payments, but their results hold only with respect to very specific function classes (as they do not assume shared inputs; see, e.g., Theorems 4.1 and 4.2 of [10]). We also demand weaker assumptions for some of our generalization theorems, thereby allowing for stronger results.

Strategyproof regression has also been studied by Perote-Peña and Perote [34]. They suggested several mechanisms and compared them to naive learning algorithms in a strategic setting. Unlike Dekel et al., they evaluated their mechanisms empirically rather than analytically, with respect to some specific assumptions on the strategic behavior of the agents.

Another rather closely related work by the same authors has results of a negative flavor. Perote and Perote-Peña [35] put forward a model of unsupervised *clustering*, where each agent controls a single point in $\mathbb{R}^2$ (i.e., its reported location). A clustering mechanism aggregates these locations and outputs a partition and a set of centroids.

They show that if every agent wants to be close to some centroid, then under very weak restrictions on the clustering mechanism there *always* exists a beneficial manipulation, that is, there are no reasonable (deterministic) clustering mechanisms that are SP.

1.3.3 Judgment and partition aggregation

While the motivation for our model stems from the *binary classification* problem in machine learning, very similar models have been used to describe various problems of judgment aggregation. In particular, a list of binary issues that must be decided upon is essentially equivalent to a dataset with binary labels. Similarly, a suggestion to split a finite set into two parts can also be replaced with labels for each element in the set.

Properties of mechanisms for judgment/partition aggregation have been discussed extensively in the literature since the 1970's [42, 30, 24, 5, 16]. A recent paper that deals explicitly with manipulations is by Dokow and Holzman [14], which characterizes strategyproof aggregation rules (that can also be interpreted as classification mechanisms in our framework).

Our current work differs in two important ways from the literature on judgment aggregation. First, we explicitly measure the quality of proposed mechanisms (in the spirit of AMDw/oM), which enables us to compare SP mechanisms to one another. Second, we study not only deterministic mechanisms, but also *randomized* ones. We believe that the notion of approximation, and the use of randomization (both a common practice in computer science) can also contribute to the study of more "standard" judgment aggregation settings. The current paper is a demonstration of this approach.

1.3.4 Facility location

In the facility location problem, agents report their location (usually in some metric space), and the mechanism outputs a location for a facility that is close, on average, to all agents. SP location mechanisms for various topologies have been suggested and studied (see, e.g., [1, 26], and [36], which also provides a clear overview of the field).

Consider a dataset labeled by several agents, and a binary cube whose dimensions correspond to the samples in the dataset. It is not hard to verify that classification with shared inputs is equivalent to facility location on the binary cube, where the label vector of each agent corresponds directly to a specific vertex of this cube. Similarly, any concept class (which defines the allowed labellings) corresponds to a set of vertices which constitutes the allowed locations. A classification mechanism then seeks the optimal classification (i.e., the *optimal vertex*) within this restricted set.

Although our main focus in the context of binary classification is the binary cube, all of our mechanisms in this paper can be directly applied to facility location problems in *any metric space*.

An important note is that it is typically assumed that the set of allowed locations for the facility coincides with the possible locations of the agents. This is equivalent to the assumption of *realizability* in our classification model. We study SP mechanisms both with and without this assumption.

1.3.5 Voting

A finite set of classifiers can also be thought of as a class of candidates in a voting scenario, where the experts are casting the votes. While such a perspective is sometimes useful (see, for example, [29]), the preferences in voting are typically much more expressive.

We can, however, model any preference profile with a proper input space. Suppose that we have a set of candidates; consider the binary cube from the last section, where every dimension (i.e., a sample in the dataset) corresponds to a *pair of candidates*. The allowed set of vertices (i.e., the concept class) restricts the outcome to vertices that correspond to a linear order over the candidates. The assumption of realizability in this setting is interpreted as *rationality* of the voters. The optimal classification mechanism, which minimizes the average distance to all voters, is equivalent to the Kemeny-Young voting rule [22]. Therefore, SP classification mechanisms can be interpreted in this setting as strategyproof approximations of the Kemeny-Young rule. It is important to note, however, that strategyproofness in our model *does not* coincide with the similar requirement in voting (as in the typical voting setting only the identity of the winner is considered).

1.3.6 Other related work

There is a significant body of work on learning in the face of noise, where the noise can be either random or adversarial (see, e.g., [6, 25]). Dalvi [9], and Dekel and Shamir [11] study settings more similar to ours, where the learning process is modeled as a game between a classifier and an adversary. However, in these papers the goal is to do well in the face of noisy or biased data, rather than provide incentives in a way that prevents the dataset from being manipulated in the first place.

Further afield, it is worth mentioning several examples from the literature that apply machine learning techniques in order to resolve problems in economics or game theory. Balcan et al. [4] apply SP machine learning algorithms to learn bidders' valuations in auctions. However, the authors achieve truthfulness by learning from agents that are not directly influenced by the outcome that relies on their reported data. This is not possible in our setting, as *all* agents are affected by the selected classifier. Other papers such as Procaccia et al. [37] suggest learning algorithms that enable better preference aggregation, but do not consider strategic behavior of the society. Finally, there has been some recent work on automated mechanism design using techniques from machine learning [7, 8]. Although the designed mechanisms are required to be truthful, the learning algorithm itself does not handle private information, and thus truthfulness is irrelevant.

2 Model and Notations

We start by introducing our model and notations for the decision-theoretic setting; additional definitions for the learning-theoretic setting are given subsequently.

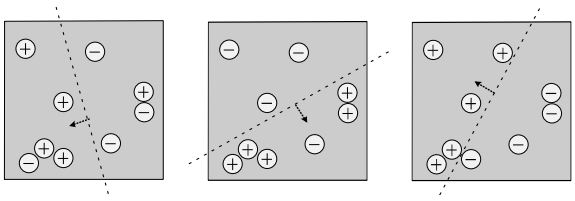

Figure 1: An instance with shared inputs. Here, $\mathcal{X} = \mathbb{R}^2$, $\mathcal{C}$ is the class of linear separators over $\mathbb{R}^2$, and $n = 3$. The data points X of all three agents are identical, but the labels, i.e., their types, are different. The best classifier from $\mathcal{C}$ with respect to each S_i is also shown (the arrow marks the positive halfspace of the separator). Only the rightmost dataset is realizable.

2.1 Binary Classification with Multiple Experts

Let $\mathcal{X}$ be an input space, which we assume to be either a finite set or some subset of $\mathbb{R}^d$. A *classifier* or *concept* c is a function $c : \mathcal{X} \to \{+, -\}$ from the input space to the *labels* $\{+, -\}$. A *concept class* $\mathcal{C}$ is a set of such concepts. For example, the class of linear separators over $\mathbb{R}^d$ is the set of concepts that are defined by the parameters $\mathbf{a} \in \mathbb{R}^d$ and $b \in \mathbb{R}$, and map a point $\mathbf{x} \in \mathbb{R}^d$ to $+$ if and only if $\mathbf{a} \cdot \mathbf{x} + b \geq 0$.

Denote the set of *agents* by $I = \{1, \ldots, n\}$, $n \geq 2$. The agents are interested in a (finite) set of k data points $X \in \mathcal{X}^k$. In this paper we assume that X is *shared* among the agents, that is, all the agents are equally interested in each data point in X. This plausible assumption, as we shall see, allows us to obtain surprisingly strong results. Naturally, the points in X are common knowledge.

Each agent has a private *type*: its labels for the points in X. Specifically, agent $i \in I$ holds a function $Y_i : X \to \{+, -\}$, which maps every point $x \in X$ to the label $Y_i(x)$ that i attributes to x. Each agent $i \in I$ is also assigned a *weight* w_i, which reflects its relative importance; by normalizing the weights we can assume that $\sum_{i \in I} w_i = 1$. Let

$$S_i = \{\langle x, Y_i(x)\rangle \ : \ x \in X\}$$

be the partial *dataset* of agent i, and let $S = \langle S_1, \ldots, S_n\rangle$ denote the complete *dataset*. S_i is said to be *realizable* w.r.t. a concept class $\mathcal{C}$ if there is $c \in \mathcal{C}$ which perfectly separates the positive samples from the negative ones. If S_i is realizable for all $i \in I$, then S is said to be *individually realizable*. Figure 1 shows an example of a dataset with a shared set of points X.

We use the common 0-1 loss function to measure the error. The *risk*,[3] or negative utility, of agent $i \in I$ with respect to a concept c is simply the relative number of errors that c makes on its dataset. Formally,

$$\mathrm{R}_i(c, S) = \frac{1}{k}\sum_{\langle x,y\rangle \in S_i} [\![c(x) \neq y]\!] = \frac{1}{k}\sum_{x \in X} [\![c(x) \neq Y_i(x)]\!] \ , \tag{1}$$

[3]When the dataset S consists of sampled data, the appropriate term is *empirical risk*. This distinction will become significant in Sections 3.3 and 4.3.

where $[\![A]\!]$ denotes the indicator function of the boolean expression A. Note that S_i is realizable if and only if $\min_{c \in \mathcal{C}} \mathrm{R}_i(c, S) = 0$. In contrast to most standard learning scenarios, in our model there is no "ground truth", and the objective is to classify in a way that will be most satisfactory to the agents. Thus the *global risk* is defined as

$$\mathrm{R}_I(c, S) = \sum_{i \in I} w_i \cdot \mathrm{R}_i(c, S) = \frac{1}{k} \sum_{i \in I} \sum_{x \in X} w_i \cdot [\![c(x) \neq Y_i(x)]\!] \ . \tag{2}$$

2.2 Mechanism Properties

A *deterministic mechanism* $\mathbf{M}$ receives as input a dataset S,[4] and outputs a classifier $c \in \mathcal{C}$. Note that since S is finite, there are only finitely many different ways to classify the data; thus, $\mathrm{R}_i(\mathbf{M}(S), S)$ for all $i \in I$ and $\mathrm{R}_I(\mathbf{M}(S), S)$ are well-defined. This will no longer be the case in the learning-theoretic setting, where we will need to slightly modify our definitions.

A *randomized mechanism* is identified with a probability distribution $p_{\mathbf{M}}$ over $\mathcal{S} \times \mathcal{C}$. We restrict our attention to probabilities with a finite support. That is, for every dataset S, the mechanism $\mathbf{M}$ returns $c \in \mathcal{C}$, with a probability of $p_{\mathbf{M}}(c|S)$.

When measuring the risk, we are interested in the *expected* number of errors that the mechanism makes on the given dataset. Formally,

$$\mathrm{R}_i(\mathbf{M}(S), S) = \mathbb{E}_{p_{\mathbf{M}}} \left[\mathrm{R}_i(c, S) \mid S \right] = \sum_{c \in \mathcal{C}} p_{\mathbf{M}}(c \mid S) \cdot \mathrm{R}_i(c, S) \ , \tag{3}$$

and the global risk is defined analogously.

For any (complete or partial) dataset $S' \subseteq S$, the best available classifier with respect to the dataset S' is referred to as the *empirical risk minimizer* (**erm**) – a common term in the machine learning literature. Formally,

$$\mathbf{erm}(S') = \mathrm{argmin}_{c \in \mathcal{C}} \sum_{\langle x, y \rangle \in S'} [\![c(x) \neq y]\!] \ . \tag{4}$$

For the complete dataset, we denote the best classifier by $c^*(S)$, and its risk by $r^*(S)$ (or simply c^*, r^* if S is clear from the context). That is,

$$c^*(S) = \mathbf{erm}(S) = \mathrm{argmin}_{c \in \mathcal{C}} \, \mathrm{R}_I(c, S)$$

and $r^*(S) = \mathrm{R}_I(c^*(S), S)$.

The simple mechanism that always computes and returns $\mathbf{erm}(S)$ is referred to as the **ERM** mechanism (with block letters).[5] If there is more than one optimal classifier, we assume that **ERM** returns one of them arbitrarily. Similarly, a mechanism which returns the best classifier with respect to a partial dataset of a specific agent (e.g., $\mathbf{erm}(S_1)$) is called a *dictator mechanism*.

[4] We implicitly assume that information regarding the weights of the agents is contained in the dataset.

[5] Actual algorithms to compute the **erm** may raise various practical problems that depend on the domain, such as computational complexity. However, such problems are not within the scope of this paper. Since an **erm** always exists and the number of data points is finite, there is an algorithm that computes an **erm** in finite time.

If $r^* = 0$ then c^* is said to be *perfect*. Note that the existence of a perfect classifier in $\mathcal{C}$ implies that all partial datasets are realizable, but the converse does not hold.

We measure the quality of the outcome of a mechanism using the standard notion of multiplicative *approximation*.

Definition 2.1. *A mechanism $\boldsymbol{M}$ is an* α-approximation *mechanism if for any dataset S it holds that $R_I(\boldsymbol{M}(S), S) \leq \alpha \cdot r^*(S)$.*

Note that randomized mechanisms are only required to attain approximation in expectation, and not necessarily with high probability.

We emphasize that the real labels of the input points are private information, and an agent may report different labels than the ones indicated by Y_i. We denote by $\overline{Y}_i : X \to \{+, -\}$ the reported labels of agent i. We also denote by $\overline{S}_i = \{\langle x, \overline{Y}_i(x)\rangle : x \in X\}$ the reported partial dataset of agent i, and by $\overline{S} = \langle \overline{S}_1, \ldots, \overline{S}_n\rangle$ the reported dataset.

Strategyproofness implies that reporting the truthful types is a dominant strategy for all agents. For a dataset S and $i \in I$, let S_{-i} be the complete dataset without the partial dataset of agent i.

Definition 2.2. *A (deterministic or randomized) mechanism $\boldsymbol{M}$ is* strategyproof *(SP) if for every dataset S, for every $i \in I$, and for every $\overline{S}_i$,*

$$R_i(\boldsymbol{M}(S), S) \leq R_i(\boldsymbol{M}(\overline{S}_i, S_{-i}), S) \ . \tag{5}$$

Our goal is to design mechanisms that are both SP and guarantee a low worst-case approximation ratio.

There is an inherent tradeoff between strategyproofness and good approximation. The **ERM** mechanism (which always returns $\mathbf{erm}(S)$), for example, is a 1-approximation mechanism, but is not SP (as we show in the next section). On the other hand, a mechanism that selects agent 1 as a *dictator*, and returns $\mathbf{erm}(S_1)$, is clearly SP but in general may give a very bad approximation (e.g., if all other agents disagree with agent 1).

We remark that for randomized mechanisms, some make a distinction between strategyproofness *in expectation* (as Definition 2.2 implies), and *universal strategyproofness*. The latter, stronger definition requires that an agent cannot gain from lying even after the randomization takes place. Interestingly, the first, weaker notion of strategyproofness is sufficient for our lower bounds, but our upper bounds satisfy universal strategyproofness.

3 Choosing from Two Classifiers

In this section we consider a very simple concept class, containing only two classifiers. For ease of exposition we assume that there is a positive classifier c_+ and a negative classifier c_-, such that $c_+(x) =$"$+$", $c_-(x) =$"$-$" for any $x \in \mathcal{X}$. Our concept class $\mathcal{C} = \{c_+, c_-\}$ can be thought of as choosing between a global *positive decision* and *negative decision*, respectively.

Remark 1. *Although we define our concept class $\mathcal{C}$ as containing two specific classifiers, our results easily extend to* every *concept class of size 2 (provided that there is at least one datapoint $x \in \mathcal{X}$ on which the two concepts disagree). Indeed, the part of the dataset on which the concepts agree can only improve the approximation ratio, and on the other hand we can always give examples where all data points are in conflict. Thus both upper and lower bounds still hold.*

We start with some observations that will allow us to simplify our model in this setting. Note that the identity of each data point is not important, only the *fraction* of positive and negative labels that each agent attributes to the dataset. We can also think of this setting as if each agent controls a *different set of points* X_i, where the size of each such partial dataset is proportional to the agent's weight. With this interpretation our model becomes even simpler, as both the weight and the type of each agent are completely defined by the *number* of "positive points" and "negative points" it controls.

Consider our TA committee example from the introduction. We can count each TA as a single data point (which is positive if it supports the strike), and the representative of each department reports the opinions of all TAs. The weight of department in this case would be proportional to the number of workers.

We denote the number of points controlled by agent i by $m_i = |X_i| = |S_i|$, and the size of the full dataset by $m = |S| = \sum_{i \in I} m_i$. This notation will be used in this section instead of k. We further denote the number of positive and negative data points by $P_i = |\{\langle x, y\rangle \in S_i : y = +\}|$, and $N_i = m_i - P_i = |\{\langle x, y\rangle \in S_i : y = -\}|$. For convenience we also let $P = \sum_{i \in I} P_i$, $N = \sum_{i \in I} N_i$. We emphasize that $\{P_i, N_i\}_{i \in I}$ contains all the information relevant to our problem and can thus replace S.

With these alternative notations, the private risk of concept c for agent i is the same as in Equation (1), only replacing k with m_i. The risk is further simplified in the two-function case:

$$\mathrm{R}_i(c, S) = \frac{1}{m_i} \sum_{\langle x,y\rangle \in S_i} [\![c(x) \neq y]\!] = \begin{cases} P_i/m_i & \text{, if } c = c_- \\ N_i/m_i & \text{, if } c = c_+ \end{cases} \tag{6}$$

We update the definition of the global risk as follows:

$$\mathrm{R}_I(c, S) = \sum_{i \in I} \frac{m_i}{m} \mathrm{R}_i(c, S) = \frac{1}{m} \sum_{\langle x,y\rangle \in S} [\![c(x) \neq y]\!] \,. \tag{7}$$

Similarly to the private risk, $\mathrm{R}_I(c, S)$ is either P/m (for c_-) or N/m (for c_+). Note that by taking $w_i = \frac{m_i}{m}$, this is a special case of Equation (2).

Unfortunately, if we choose **ERM** as our mechanism, then even in this simple setting the agents may lie in order to decrease their subjective risk.

Example 3.1. *(Illustrated in Figure 2) Agent 1 controls 3 examples: 2 positive and 1 negative. Agent 2 controls 2 examples, both negative. Since there is a majority of negative examples,* ***ERM*** *would return c_-; agent 1 would suffer a subjective risk of $2/3$. On the other hand, if agent 1 reported his negative example to be positive as well,* ***ERM*** *would return c_+, with a subjective risk of only 1/3 for agent 1. Indeed, note that*

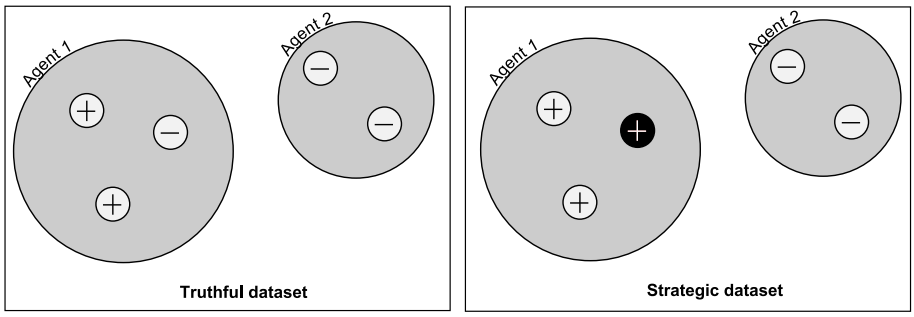

Figure 2: **ERM** is not strategyproof. Agent 1 changes one of its points from negative to positive, thus changing the risk minimizer from c_- to c_+, to agent 1's advantage. In this illustration, $\mathcal{X} = \mathbb{R}^2$.

an agent's utility is measured with respect to its real labels, rather than with respect to the reported labels. $\Diamond$

It is easy to see, however, that an agent cannot gain by lying when it only controls one point. For instance, if an agent has a positive point and **ERM** returns c_-, falsely reporting a negative label will only reinforce the mechanism's decision. This is in striking contrast to the regression learning setting considered in Dekel et al. [10], where the deepest technical results concern the single-point-per-agent scenario.

Despite the fact that **ERM** is not SP, we would still like to use the optimal concept in order to evaluate other concepts and mechanisms. From the definition of the **erm**, we have that

$$r^* = \mathrm{R}_I(c^*, S) = \min\{\mathrm{R}_I(c_+, S), \mathrm{R}_I(c_-, S)\} = \min\left\{\frac{N}{m}, \frac{P}{m}\right\} .$$

3.1 Deterministic Mechanisms

Denote by c_i the **erm** on S_i, i.e., $c_i = c_+$ if $P_i \geq N_i$ and c_- otherwise. Clearly c_i is the best classifier agent i can hope for. Consider the mechanism given as Mechanism 1.

Mechanism 1 THE PROJECTED MAJORITY MECHANISM (**PM**)

Based on the labels of each agent P_i, N_i, calculate c_i. Define each agent as a *negative agent* if $c_i = c_-$, and as a *positive agent* if $c_i = c_+$.
Denote by $P' = \sum_{i:c_i=c_+} m_i$ the number of examples that belong to positive agents, and similarly $N' = \sum_{i:c_i=c_-} m_i = m - P'$.
if $P' \geq N'$ **then return** c_+.
else return c_-.
end if

Remark 2. *Informally we state that in our current setting, we can obtain similar approximation results even under mechanisms that are not SP, assuming agents lie only*

when this is beneficial to them. Nevertheless, strategyproofness gives us a very clean framework to analyze mechanisms in the face of strategic behavior. When we discuss our learning theoretic framework, where obtaining strategyproofness is next to impossible, we shall apply the former, less elegant, type of analysis.

We will show that this mechanism has the excellent game-theoretic property of being *group strategyproof*: no coalition of players can gain by lying. In other words, if some agent in the coalition strictly gains from the joint lie, some other agent in the coalition must strictly lose. While technically simple, this first result demonstrates the key principles of strategyproof mechanisms.

Theorem 3.2. *Mechanism 1 is a 3-approximation group-SP mechanism.*

Proof. We first show group strategyproofness. Let $B \subseteq I$. We can assume without loss of generality that either all agents in B are positive or all of them are negative, since a positive (resp., negative) agent cannot gain from lying if the mechanism returns c_+ (resp., c_-). Again without loss of generality, the agents are all positive. Therefore, if some agent is to benefit from lying, the mechanism has to return c_- on the truthful dataset. However, since the mechanism considers all agents in B to be positive agents when the truthful dataset is given, an agent in B can only hope to influence the outcome by reporting a majority of negative examples. However, this only increases N', reinforcing the mechanism's decision to return c_-.

It remains to demonstrate that the approximation ratio is as claimed. We assume without loss of generality that the mechanism returned c_+, i.e., $P' \geq N'$. We first prove that if the mechanism returned the positive concept, at least $1/4$ of the examples are indeed positive, that is, $P \geq \frac{1}{4}m$.

Indeed, clearly $P' \geq \frac{m}{2} \geq N'$ otherwise we would get $c = c_-$. Now, if an agent is *positive* ($c_i = c_+$), at least half of its examples are also positive. Thus

$$P = \sum_{i \in I} P_i \geq \sum_{i:c_i = c_+} P_i \geq \sum_{i:c_i = c_+} \frac{m_i}{2} = \frac{P'}{2},$$

and hence $P \geq \frac{P'}{2} \geq \frac{m}{4}$.

Now, we know that $P + N = m$, so $N = m - P \leq m - \left(\frac{m}{4}\right) = \frac{3m}{4} \leq 3P$. Clearly if the mechanism decided "correctly", i.e., $P \geq m/2$, then

$$\mathrm{R}_I(c, S) = \mathrm{R}_I(c_+, S) = \frac{N}{m} = r^*.$$

Otherwise, if $P < m/2$, then

$$\mathrm{R}_I(c, S) = \mathrm{R}_I(c_+, S) = \frac{N}{m} \leq 3\frac{P}{m} = 3\mathrm{R}_I(c_-, S) = 3r^*.$$

In any case we have that $\mathrm{R}_I(c, S) \leq 3r^*$, proving that Mechanism 1 is indeed a 3-approximation mechanism. ■

As 3-approximation is achieved by such a trivial mechanism, we would naturally like to know whether it is possible to get a better approximation ratio, without waiving

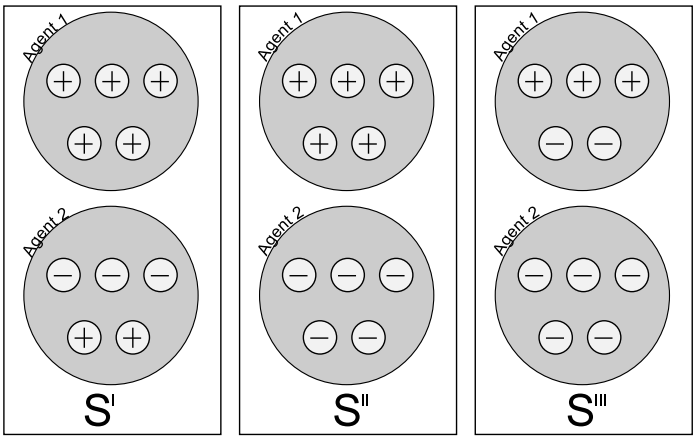

Figure 3: The examples of each agent in the three datasets are shown (for $t = 2$). Agent 1 can make dataset II look like dataset III and vice versa by reporting false labels. The same goes for agent 2 regarding datasets I and II.

the SP property. We show that this is *not* the case by proving a matching lower bound on the best possible approximation ratio achievable by an SP mechanism. Note that the lower bound only requires strategyproofness, not group strategyproofness.

Theorem 3.3. *Let $\epsilon > 0$. There is no $(3-\epsilon)$-approximation strategyproof mechanism.*

Proof. To prove the bound, we present 3 different datasets. We show that any SP mechanism must return the same result on all of them, while neither concept in C yields an approximation ratio of $(3-\epsilon)$ in all three.

Let $\epsilon > 0$. We will use $I = \{1, 2\}$, and an integer $t = t(\epsilon)$ to be defined later. Note that in all three datasets $m_1 = m_2 = 2t + 1$. We define the three datasets as follows (see Figure 3 for an illustration):

- S^I: $P_1 = 2t + 1, N_1 = 0$; $P_2 = t, N_2 = t + 1$
- S^{II}: $P_1 = 2t + 1, N_1 = 0$; $P_2 = 0, N_2 = 2t + 1$
- S^{III}: $P_1 = t + 1, N_1 = t$; $P_2 = 0, N_2 = 2t + 1$

Let $\mathbf{M}$ be some strategyproof mechanism. Then it must hold that $\mathbf{M}(S^I) = \mathbf{M}(S^{II})$. Indeed, otherwise assume first that $\mathbf{M}(S^I) = c_+$ and $\mathbf{M}(S^{II}) = c_-$. Notice that the only difference between the two settings is agent 2's labels. If agent 2's truthful labels are as in S^I, his subjective **erm** is c_-. Therefore, he can report his labels to be as in S^{II} (i.e., all negative) and obtain c_-. Now, if $\mathbf{M}(S^I) = c_-$ and $\mathbf{M}(S^{II}) = c_+$, agent 2 can gain by deviating from S^{II} to S^I. A symmetric argument, with respect to agent 1 (that in all settings prefers c_+) shows that $\mathbf{M}(S^{II}) = \mathbf{M}(S^{III})$.

So, without loss of generality assume that $c = \mathbf{M}(S^I) = \mathbf{M}(S^{II}) = \mathbf{M}(S^{III}) = c_+$ (otherwise, symmetric arguments yield the same result). Therefore:

$$\mathrm{R}_I(c, S^{III}) = \mathrm{R}_I(c_+, S^{III}) = \frac{N_1 + N_2}{m} = \frac{3t+1}{4t+2} \tag{8}$$

On the other hand, the negative concept is much better:

$$r^* = \mathrm{R}_I(c_-, S^{III}) = \frac{t+1}{4t+2}$$

By combining the last two equations:

$$\frac{\mathrm{R}_I(c, S^{III})}{r^*} = \frac{\frac{3t+1}{4t+2}}{\frac{t+1}{4t+2}} = \frac{3t+1}{t+1}$$

Let us set $t > \frac{3}{\epsilon}$; then the last expression is strictly greater than $3 - \epsilon$, and thus $\mathrm{R}_I(c, S^{III}) > (3-\epsilon)r^*$. We conclude that any SP mechanism cannot have an approximation ratio of $3 - \epsilon$. ■

3.2 Randomized mechanisms

What if we let our mechanism flip coins? Can we find an SP randomized mechanism that beats (in expectation) the 3-approximation deterministic lower bound? To answer the question we first recall the definition of the risk of such a mechanism given in (3).

For our simple concept class $C = \{c_+, c_-\}$, a randomized mechanism is defined only by the probability of returning a positive or negative concept, given S. Accordingly, the risk (both private and global) is

$$\mathrm{R}(\mathbf{M}(S), S) = p_+ \cdot \mathrm{R}(c_+, S) + p_- \cdot \mathrm{R}(c_-, S)\ ,$$

where p_+, p_- stand for $p_{\mathbf{M}}(c_+ \mid S)$ and $p_{\mathbf{M}}(c_- \mid S)$.

We start our investigation of SP randomized mechanisms by establishing a lower bound of 2 on their approximation ratio.

Theorem 3.4. *Let $\epsilon > 0$. There is no $(2-\epsilon)$-approximation strategyproof randomized mechanism.*

The proof, along with all the remaining proofs of this section, appears in Appendix A.

We presently put forward a randomized SP 2-approximation mechanism, thereby matching the lower bound with an upper bound. However we first propose a simpler mechanism and analyze where it fails: The natural thing to do would be to calculate P' and N' as in our deterministic Projected Majority Mechanism and then simply to select c_+ with probability P'/m and c_- with probability N'/m. We refer to this simple mechanism as the *weighted random dictator* mechanism (**WRD**), for reasons that will become apparent in Section 4.1.[6] Unfortunately, this simple randomization (which is clearly SP) cannot even beat the deterministic bound of $3 - \epsilon$, as demonstrated by the following example.

[6]This procedure is equivalent to randomly selecting an agent with probability proportional to its weight, and using its preferred classifier to classify the entire dataset — hence *Weighted Random Dictator*.

Example 3.5. *Consider the dataset S of n agents with the following examples: one agent with $P_1 = t+1$, $N_1 = t$, and $n-1$ additional agents each holding $2t+1$ negative examples. Thus $P = t+1; N = (n-1)(2t+1)$ but $P' = 2t+1; N' = (n-1)(2t+1)$. The optimal classifier makes $|P| = t+1$ mistakes, thus $r^* = \frac{t+1}{m}$. On the other hand, the expected number of mistakes made by the mechanism is*

$$\begin{aligned} m \cdot R_I(\textbf{\textit{WRD}}(S), S) =& p_- \cdot |P| + p_+ \cdot |N| = \frac{N'}{m} \cdot (t+1) + \frac{P'}{m} \cdot ((n-1)(2t+1) + t) \\ =& \frac{(n-1)(2t+1)}{n(2t+1)}(t+1) + \frac{2t+1}{n(2t+1)}(2nt + n - t - 1) \\ =& \frac{(n-1)(t+1)}{n} + \frac{2nt + n - t - 1}{n} = \\ =& \frac{nt + n - t - 1 + 2nt + n - t - 1}{n} = \frac{3nt + 2n - 2t - 2}{n} \,. \end{aligned}$$

We have that the approximation ratio of this mechanism is at least

$$\frac{R_I(\textbf{\textit{WRD}}(S), S)}{r^*} = \frac{3nt + 2n - 2t - 2}{n(t+1)} \overset{t \to \infty}{\to} 3 - \frac{2}{n}. \tag{9}$$

Thus, for every $\epsilon > 0$, there is a large-enough t such that the approximation ratio is worse than $3 - \frac{2}{n} - \epsilon$. ♢

Note that in this example all agents control datasets of the same size ($2t + 1$). A similar example can be crafted with two weighted agents, by merging the datasets of agents $2, \ldots, n$ to a single, heavier, agent. This example will provide us with a lower bound of $3 - 2w_1$, where w_1 is the weight of the lighter agent.

Crucially, an adjusted, less intuitive randomization can do the trick.

Mechanism 2 The Square Weighted Dictator Mechanism (**SRD**)

Compute P' and N' as in Mechanism 1.
Return c_+ or c_- with probability proportional to $(P')^2, (N')^2$, respectively.

Theorem 3.6. *Mechanism 2 is a group-SP 2-approximation randomized mechanism.*

There are, in fact, multiple ways to achieve a 2-approximation using different randomizations on N' and P'. In a previous version of this paper we suggested one such alternative randomization [28]. A third procedure follows as a special case from the **CRD** mechanism described in Section 4.1.

3.3 Binary Decision in a Learning Theoretic Setting

In this section we extend our simple setting to a more general machine learning framework. Our previous results will be leveraged to obtain powerful learning theoretic results.

Instead of looking at a fixed set of examples and selecting the concept that fits them best, we now turn to look at *sampled datasets*. That is, we assume that there is some fixed and known distribution $\mathcal{D}_X \in \Delta(\mathcal{X})$ (where $\Delta(A)$ is the set of probability distributions over a set A), which represents the *interest* that agents have in different parts of the input space. According to our shared input assumption, the distribution of interest is the same for all agents.

In addition, each agent $i \in I$ now has a private function $Y_i : \mathcal{X} \to \{+, -\}$, which assigns a label to every point in the input space. Observe that Y_i, along with the distribution $\mathcal{D}_X$, induces a (private) distribution $\mathcal{D}_i$ over inputs and labels, i.e., $\mathcal{D}_i \in \Delta(\mathcal{X} \times \{+, -\})$. This distribution determines the type of agent i.

The new definition of the subjective risk naturally extends the previous setting by expressing the errors a concept makes with respect to the distribution $\mathcal{D}_i$:

$$\mathrm{R}_i(c) = \mathbb{E}_{(x,y)\sim\mathcal{D}_i}[[\![c(x) \neq y]\!]] = \mathbb{E}_{x\sim\mathcal{D}_X}[[\![c(x) \neq Y_i(x)]\!]] \,. \tag{10}$$

The global risk is calculated similarly to how it was previously defined, as the weighted average of the private risk, i.e.,

$$\mathrm{R}_I(c) = \sum_{i\in I} w_i \cdot \mathrm{R}_i(c) \,. \tag{11}$$

For ease of exposition, we will assume in this section that all agents have equal weight. Thus, $\mathrm{R}_I(c) = \frac{1}{n}\sum_{i\in I} \mathrm{R}_i(c)$. In Section 4.3, when discussing the more general problem, we will not use this assumption.[7]

Similarly, we can no longer compare the outcome of our mechanism to $r^*(S)$, as this notion of the optimal risk assumes a fixed dataset, whereas an instance of the learning-theoretic setting consists of a set of *distributions*. We therefore define the minimal risk as

$$r_{\min} = \inf_{c\in\mathcal{C}} \mathrm{R}_I(c) \,. \tag{12}$$

Although in the general case $\mathcal{C}$ might be an open set, in our simple two-function setting $\mathcal{C}$ is finite, and $r_{\min} = \min\{\mathrm{R}_I(c_-), \mathrm{R}_I(c_+)\}$.

Note that we cannot directly evaluate the risk in this learning theoretic framework; we may only sample points from the agents' distributions and ask the agents to label them. We then try to minimize the *real* global risk, using the *empirical risk* as a proxy.[8] The empirical risk is the risk on the sampled dataset, as defined in the previous section.

Remark 3. *A subtle point is that the mechanism we present is not strategyproof, and in fact* no mechanism *that gets sampled data points as input is strategyproof. Indeed, even if there is only a single agent, which gives greater weight to negative points (according to $\mathcal{D}_1$), it might be the case that, by miserable chance, the agent's sampled dataset only contains positive points. Thus there is some non-zero probability that the agent will have an incentive to "lie" by reporting negative labels.*

[7] The results in this section can also be generalized to varying weights by sampling for each agent a number of points proportional to its weight, yet still large enough.

[8] This is similar to an oracle model, where we have no direct access to the distribution, but we can ask yes/no questions about it. The major difference is that in our model the "oracle" may lie! (Perhaps the *Sphinx* model would be a better name)

We note that even allowing payments would not guarantee strategyproofness in our example, as it contains only one agent. This fact may seem contradictory to the revelation principle (see, e.g., [32]), but not if we recall that truthful mechanisms are only guaranteed to exist under direct revelation*. In our domain, direct revelation means that the agents must be asked to explicitly select the classifier they prefer. However in the learning-theoretic setting the agents only reveal their preferences indirectly, by submitting their preferred labels on the sampled data points.*

3.3.1 Three Game-Theoretic Assumptions

While full strategyproofness is too much to ask for, we can still make assumptions on the behavior of agents that will allow us to formally analyze the outcome of our mechanisms. We exploit this very simple setting to clarify the distinction between three alternative game-theoretic assumptions on agents' behavior.

The ϵ-truthfulness assumption. The first assumption is that agents will not lie unless their expected gain from this lie is *at least* ϵ. This assumption is stronger than the rationality assumption in the decision-making setting, where we demanded this only for $\epsilon = 0$. In Section 4.3 we refer to this assumption as the "Truthful Approach". This is the approach taken for example by Dekel et al. [10].

The pure rationality assumption. A second assumption is that agents will *always* play a dominant strategy, if one is available to them. The existence of dominant strategies depends on the mechanism, as well as on the dataset, and we allow arbitrary behavior when such a strategy does not exist. This assumption is also stronger than the standard rationality assumption (which does not assume anything about agents' behavior when truth-telling is suboptimal), but it is incomparable with the first assumption. In Section 4.3 we refer to this assumption as the "rational approach". It is important to note that the rational approach entails that agents must have complete knowledge of their own distribution. This implicit assumption is not necessary under the truthful approach.

The weak truthfulness assumption. The third assumption, which is also the weakest, requires that an agent is truthful if this is a weakly dominant strategy, i.e., if it cannot gain by lying.

An agent that always obeys the first, second or third assumption is called *ϵ-truthful*, *purely rational*, or *weakly truthful*, respectively. Note that both ϵ-truthful agents (for any $\epsilon \geq 0$) and purely rational agents are always weakly truthful, which means that the third assumption is indeed the weakest.

In this section we employ the third assumption as it supplies us with the strongest results. Thus the results in this section are "stronger" in a way than the results of Dekel et al. [10] (regression) and the results in Section 4 (classification).[9]

[9] In fact, a simple variant of the proofs in Section 4.3 could be directly applied to the binary decision problem (as it is a special case of shared inputs, and has a bounded VC dimension), yielding an approximation ratio that is close to 2. However, this bound would only hold under either of the first two strategic

Remark 4. *We offer a simple scenario that will highlight the substantial difference between the different assumptions. Suppose we employ the pure rationality assumption, and consider the following simple mechanism: sample one point from $\mathcal{D}_X$, and let all agents label this single point. If an agent labels the point positively, the agent is positive; otherwise it is negative. Now apply either Mechanism 1 or Mechanism 2. This clearly gives us approximation upper bounds of 3 and 2 respectively, using only one sampled data point. In contrast, the ϵ-truthfulness assumption will not guarantee anything in this case. This suggests that the difference between the assumptions is non-trivial. Compare also with the analysis of the two first approaches in Section 4.3.*

Mechanism 3 The Binary Learning Mechanism ($\widetilde{\textbf{SRD}}$)

---

for each agent $i \in I$ **do**
  Sample $m' = m_i$ points i.i.d. from $\mathcal{D}_X$.
  Denote i's set of data points as $X_i = \{x_{i,1}, \ldots, x_{i,m'}\}$.
  Ask agent i to label X_i.
  Denote $\overline{S}_i = \{\langle x_{i,j}, \overline{Y}_i(x_{i,j})\rangle\}_{j=1}^{m'}$.
end for
Use Mechanism 2 on $\overline{S} = \{\overline{S}_1, \ldots, \overline{S}_n\}$, **return SRD**$(\overline{S})$.

The risk of the mechanism is computed as the expectation of the risk of the outcome classifier, where the expectation is taken over both randomizations: the sampling of the data points, and the randomization performed by **SRD**. Formally (for both private and global risk),

$$\mathrm{R}(\widetilde{\textbf{SRD}}) = \mathbb{E}_{X\sim(\mathcal{D}_X)^m}\left[\mathrm{R}(\textbf{SRD}(\overline{S}))\right] \ , \tag{13}$$

where the labels of X in $\overline{S}$ are set according to our strategic assumptions.

We presently establish a theorem that explicitly states the number of examples we need to sample in order to properly estimate the real risk. We will get that, in expectation (taken over the randomness of the sampling procedure and Mechanism 2's randomization), Mechanism 3 yields close to a 2-approximation with relatively few examples, even in the face of strategic behavior.

Theorem 3.7. *Given sampled datasets, assume weak truthfulness. For any $\epsilon > 0$, there is m' (polynomial in $\ln(n)$ and $\frac{1}{\epsilon}$) such that by sampling m' points for each agent, it holds that*

$$R_I(\widetilde{\textbf{SRD}}) \leq 2r_{\min} + \epsilon.$$

Specifically, sampling $m' > 50\frac{1}{\epsilon^2}\ln(\frac{10n}{\epsilon})$ will suffice.

While the proof is quite technical, it can be sketched as follows. Mechanism 2 is SP with respect to the (already sampled) dataset $\overline{S}$. Thus if an agent's sampled dataset faithfully represents its true distribution, and the agent is strongly inclined towards c_+ or c_-, the agent still cannot benefit by lying (by the weak truthfulness assumption). If an agent is almost indifferent between c_+ and c_-, it might wish to lie—but crucially, such an agent contributes little to the global risk.

---

assumptions.

4 Classification with Shared Inputs

We begin with an analysis of the decision-theoretic setting. As in Section 3, these results will later be applied to the learning-theoretic setting.

In this section, we assume that all agents control the same set of data points. The size of this dataset is denoted by k. The total number of labeled data points from all agents is thus $m = n \cdot k$. However, as our mechanisms in this section use only a single agent, k is effectively the size of the input being used.

4.1 Deterministic Mechanisms

We start by examining an extremely simple deterministic mechanism. Recall that **erm**(S') is the concept $c \in \mathcal{C}$ that minimizes the risk w.r.t. $S' \subseteq S$ (see Equation (4)). Our mechanism simply lets the heaviest agent dictate which concept is chosen.

Mechanism 4 The Heaviest Dictator Mechanism (**HD**)

$h \leftarrow \text{argmax}_{i \in I}\, w_i$. // (Let $h \in I$ be an agent with maximal weight)
return erm(S_h).

If more than one **erm** exists, return one of them arbitrarily. The mechanism is clearly SP: the heaviest dictator h has no interest to lie, since its best concept is selected; all other agents are simply ignored, and therefore have no reason to lie either. We have the following result.

Theorem 4.1. *Let $|I| = n$. For every concept class $\mathcal{C}$ and any dataset S, Mechanism 4 is an SP $(2n-1)$-approximation mechanism.*

Recall the central negative result regarding deterministic mechanisms with non-restricted input.

Theorem 4.2 (Meir, Procaccia, and Rosenschein [29])**.** *There exist concept classes for which any deterministic SP mechanism has an approximation ratio of at least $\Omega(m)$, where m is the total size of the full dataset.*

We therefore see that the restriction to shared inputs helps by removing the dependency on the size of the dataset, but nevertheless an approximation ratio that increases linearly with the number of agents is not very appealing. However, it turns out that using deterministic mechanisms we cannot do better with respect to every concept class. Indeed, a slight variation of Theorem 4.2 gives us the following result.

Theorem 4.3. *Suppose there are n agents with shared inputs. There exist concept classes for which any deterministic SP mechanism has an approximation ratio of at least $\Omega(n)$, even if all the weights are equal.*

The proof of the theorem is a minor variation of the proof of Theorem 4.2, which applies the Gibbard-Satterthwaite impossibility theorem [17, 38].

Theorem 4.3 implies that Mechanism 4 is optimal, up to a constant, as a generic mechanism that applies to any concept class. Of course, for specific concept classes

one can do much better, as shown in Section 3. One could hope that imposing further restrictions on the dataset, such as realizability, would enable the design of better SP mechanisms. However, recent results show that the $\Omega(n)$ bound remains even if all datasets are realizable [13].

4.2 Randomized Mechanisms

In order to break the lower bound given by Theorem 4.3, we employ a simple randomization. We will see that this randomization yields a constant approximation ratio *with respect to any concept class* (under our assumption of shared inputs, of course). Moreover, if the agents have uniform weights, then this mechanism cannot be further improved.

Mechanism 5 The Weighted Random Dictator (**WRD**) mechanism

select agent i with probability w_i.
return erm(S_i).

Consider Mechanism 5, which is clearly SP. The following theorem bounds its approximation ratio for different cases.

Theorem 4.4. *For every concept class $\mathcal{C}$ and for any dataset S, Mechanism 5 is an SP $(3 - 2w_{\min})$-approximation mechanism, where $w_{\min} = \min_{i \in I} w_i$. Moreover, if S is individually realizable, then $(2 - 2w_{\min})$-approximation is guaranteed.*

When all agents have the same weight, we have that $w_{\min} = \frac{1}{n}$. We therefore have the following corollary which follows directly from Theorem 4.4.

Corollary 4.5. *Let $|I| = n$, and assume all agents have equal weights. For every concept class $\mathcal{C}$ and for any dataset S, Mechanism 5 is an SP $(3 - \frac{2}{n})$-approximation mechanism ($2 - \frac{2}{n}$ when S is individually realizable).*

The last corollary also follows as a special case from results we will see in Section 4.2.2.

It is possible to show that the analysis of Mechanism 5 is tight. Indeed, consider the outcome of the mechanism for the concept class $\{c_-, c_+\}$. In this case, the mechanism is essentially equivalent to the naive randomized mechanism presented in Section 3.2, and yields the same outcome. Therefore, Example 3.5 gives a tight lower bound on the approximation ratio of the mechanism, matching the upper bound given in Theorems 4.4 and 4.5. A similar example can be easily constructed for every concept class of size at least two.

4.2.1 Is the WRD mechanism optimal?

It is natural to ask whether better (randomized) SP mechanisms exist. For specific concept classes, the answer to this question is positive, as demonstrated by Theorem 3.6. For general concept classes, the following lower bound is known.

Theorem 4.6 (Meir, Almagor, Michaely and Rosenschein [27])**.** *Suppose there are n agents with shared inputs. There exist concept classes for which any randomized SP mechanism has an approximation ratio of at least $3 - \frac{2}{n}$, even if all the weights are equal.*

Theorem 4.6 shows that when weights are uniform, the **WRD** mechanism (i.e., selecting a dictator uniformly at random) is in fact optimal. That is, no SP mechanism can do better. However, the mechanism is suboptimal for weighted datasets, as it only guarantees a 3 approximation in this case.

We next turn to close this gap, presenting new mechanisms that beat the **WRD** mechanism on weighted datasets, matching the lower bound given in Theorem 4.6.

4.2.2 Improving the upper bound for weighted agents

Theorem 4.6 in fact tells us that we must pick a dictator at random to have an SP mechanism. However we are still free to define the probabilities of selecting different agents, and we may take agents' weights into account. The **WRD** mechanism is an example of such a randomization, but we can design others.

Recall that in the two-function scenario, we performed an optimal randomization by using the **SRD** mechanism. As a first attempt to improve the upper bound, we translate the **SRD** mechanism to the current setting.[10] That is, the mechanism would select every dictator $i \in I$ with probability proportional to w_i^2. Unfortunately, while **SRD** does attain some improvement over the **WRD** mechanism, it is still suboptimal, even for $n = 3$.

Proposition 4.7. *There is a dataset S with three agents, such that*

$$R_I(\textbf{\textit{SRD}}(S), S) > 2.4 \cdot r^* > \left(3 - \frac{2}{n}\right) r^*.$$

A similar counterexample exists for individually realizable datasets, where the approximation ratio of **SRD** is above 1.39 (i.e., strictly above $2 - \frac{2}{n}$ for $n = 3$). We therefore must take a somewhat different approach in the selection of the dictator. Consider the mechanisms **CRD** and **RRD**, where the latter is a small variation of the former.

Mechanism 6 The Convex-weight Random Dictator Mechanism (**CRD**)

for each $i \in I$, set $p'_i = \frac{w_i}{2 - 2w_i}$.
compute $\alpha_{\mathbf{w}} = \frac{1}{\sum_{i \in I} p'_i}$.
select agent i with probability $p_i = \alpha_{\mathbf{w}} p'_i$.
return erm(S_i).

The **CRD** and **RRD** mechanisms are clearly SP, as the probabilities are unaffected by the reported labels.

Theorem 4.8. *The following hold for Mechanism 6:*

[10] We slightly abuse notation here and use the name **SRD**, although it is no longer equivalent to Mechanism 2.

Mechanism 7 The Realizable-weight Random Dictator Mechanism (**RRD**)

$h \leftarrow \text{argmax}_{i \in I} w_i$.
if $w_h \geq \frac{1}{2}$ **then**
  return erm(S_h).
end if
for each $i \in I$, set $p'_i = \frac{w_i}{1-2w_i}$.
compute $\beta_{\mathbf{w}} = \frac{1}{\sum_{i \in I} p'_i}$.
select agent i with probability $p_i = \beta_{\mathbf{w}} p'_i$.
return erm(S_i).

- $\alpha_{\mathbf{w}} \leq 2 - \frac{2}{n}$.
- ***CRD** has an approximation ratio of* $1 + \alpha_{\mathbf{w}}$*, i.e., at most* $\mathbf{3 - \frac{2}{n}}$.
- *if S is individually realizable, then the approximation ratio is* $\frac{\alpha_{\mathbf{w}}}{2} + 1$*, i.e., at most* $2 - \frac{1}{n}$.

By Theorem 4.6, no SP mechanism can do better on a general dataset in the worst case, thus **CRD** is optimal. However, if the dataset is known to be individually realizable, **CRD** is suboptimal, and **RRD** is strictly better (in the worst case).

Theorem 4.9. *The following hold for Mechanism 7:*

- $\beta_{\mathbf{w}} \leq 1 - \frac{2}{n}$.
- ***RRD** has an approximation ratio of at most* 4*, and at least* 3 *(in the worst case).*
- *if S is individually realizable, then the approximation ratio is* $1 + \beta_{\mathbf{w}}$*, i.e., at most* $\mathbf{2 - \frac{2}{n}}$.

Observe that for two agents the **RRD** simply selects the heavier dictator. Thus if the dataset is not realizable, the approximation ratio can be as high as 3, which accounts for the lower bound in the non-realizable case.

The **CRD** mechanism matches the lower bounds for *any* set of weighted agents, thereby showing that the uniform weight case is, in fact, the hardest. The situation with the **RRD** mechanism is similar—no randomization of dictators can do better. However, it is still an open question whether there are better, more sophisticated, randomized mechanisms for the realizable case. The natural conjecture would be that there are none, as Dokow et al. proved for deterministic mechanisms [13].

Note that when weights are uniform, then the **CRD**, **RRD**, **SRD** and **WRD** mechanisms all coincide.[11] Thus Theorem 4.5 also follows as a special case from Theorems 4.8, 4.9.

Curiously, **RRD** is better than **CRD** when the dataset in known to be realizable, whereas in the general case the converse is true. Therefore, a *different* mechanism should be used, depending on our assumptions on the dataset. However, the mechanism

[11] There is a tiny exception here: when $n = 2$, $w_1 = w_2 = \frac{1}{2}$, then **RRD** returns an arbitrary dictator, rather than random. However in this case any outcome is a 1-approximation.

must be decided on *a-priori*—we cannot select between **CRD** and **RRD** after observing the labels, as this would not be strategyproof!

4.2.3 Applying the mechanisms to the two-function setting

Suppose that $\mathcal{C} = \{+, -\}$. We can join together all positive agents, and all negative agents, and construct an instance with two meta-agents, whose weights are proportional to P', N' (as defined in Section 3.1). The **RRD** mechanism then simply selects the heavier meta-agent (equivalently to the **PM** mechanism), and thus guarantees an approximation ratio of 3. The **CRD** mechanism, applied to this setting, guarantees an approximation ratio of $3 - \frac{2}{n} = 3 - \frac{2}{2} = 2$. It therefore supplies us with an alternative 2-approximation SP mechanism for the two-function setting.

4.3 The Learning-Theoretic Setting

In this section we leverage the upper bounds which were attained in the decision-theoretic setting to obtain results in a machine-learning framework. That is, we present a learning mechanism that guarantees a constant approximation of the optimal risk in expectation, even in the face of strategic behavior.

We use the notations and definitions introduced in Section 3.3, where the preferences of each agent are represented by a function $Y_i : \mathcal{X} \to \{+, -\}$.[12] Reinterpreting our shared input assumption in the learning-theoretic setting, we assume that all agents have *the same* probability distribution $\mathcal{D}_X$ over $\mathcal{X}$, which reflects the relative importance that the agents attribute to different input points; the distribution $\mathcal{D}_X$ is common knowledge.

The private risk of a classifier $c \in \mathcal{C}$ is computed according to Equation (10):

$$\mathrm{R}_i(c) = \mathbb{E}_{x \sim \mathcal{D}_X} \left[[\![c(x) \neq Y_i(x)]\!] \right] .$$

That is, according to the *expected* number of errors that c makes w.r.t. the distribution $\mathcal{D}_X$. As for the global risk, it is computed according to Equation (11), i.e.

$$\mathrm{R}_I(c) = \sum_{i \in I} w_i \mathrm{R}_i(c) .$$

The goal of our mechanisms is to find classifiers with low risk. We therefore compare them to the best risk that is attainable by concepts in $\mathcal{C}$, and thus $r_{\min} = \inf_{c \in \mathcal{C}} \mathrm{R}_I(c)$. Equation (12) is a special case of this definition for $\mathcal{C} = \{c_-, c_+\}$.

Our goal is, once again, to design mechanisms with risk close to optimal. However, constructing an SP mechanism that learns from sampled data is nearly impossible (as explained in Remark 3). Hence, we weaken the strategyproofness requirement, and analyze the performance of our mechanisms under each of the first two strategic assumptions described in Section 3.3: the ϵ-truthfulness assumption, which states that agents do not lie unless they gain at least ϵ; and the pure rationality assumption, under which agents always play a weakly dominant strategy if one exists.

[12] As with the theorems in Section 4.1, our results in this section will follow as a special case from the more general model, where agents have distributions over the labels.

4.3.1 The ϵ-Truthfulness Assumption

An ϵ-strategyproof mechanism is one where agents cannot gain more than ϵ by lying. We show below that, similarly to Dekel et al. [10], the results of Section 4.2 can be employed to obtain a mechanism that is "usually" ϵ-strategyproof. We focus on the following mechanism.

Mechanism 8 The Generic Learning Mechanism ($\widetilde{\textbf{CRD}}$)

Sample k data points i.i.d. from $\mathcal{D}_X$ (denote the sampled points by X).
for each agent $i \in I$ **do**
  Ask agent i to label X_i.
  Denote $\overline{S}_i = \{\langle x_j, \overline{Y}_i(x_j)\rangle\}_{j=1}^{k}$.
end for
Use Mechanism 5 on $\overline{S} = \{\overline{S}_1, \ldots, \overline{S}_n\}$, **return CRD**$(\overline{S})$.

We denote by $\mathrm{R}_I(\widetilde{\textbf{CRD}})$ the expected risk of Mechanism 8, where the expectation is taken over the randomness of the sampling and the randomness of Mechanism 5, just as in Equation (13) in the two-function setting:

$$\mathrm{R}(\widetilde{\textbf{CRD}}) = \mathbb{E}_{X\sim(\mathcal{D}_X)^k}\left[\mathrm{R}(\textbf{CRD}(\overline{S}))\right] \ ,$$

where the labels of X in $\overline{S}$ are set according to our varying strategic assumptions.

We wish to formulate a theorem that asserts that, given enough samples, the expected risk of Mechanism 8 is relatively small under the ϵ-truthfulness assumption. The exact number of samples needed depends on the combinatorial richness of the function class; this is usually measured using some notion of class complexity, such as the VC dimension (see, e.g., [21]). For instance, the VC dimension of the class of linear separators over $\mathbb{R}^d$ is $d+1$. We do not dwell on this point too much, and instead assume that the dimension is bounded.

Theorem 4.10. *Assume all agents are ϵ-truthful, and let $\mathcal{C}$ be any concept class with a bounded dimension. For any $\epsilon > 0$, there is k (polynomial in $\frac{1}{\epsilon}$ and $\ln(n)$) s.t. if at least k datapoints are sampled, then the expected risk of Mechanism 8 is at most $\left(3 - \frac{2}{n}\right) \cdot r_{\min} + \epsilon$.*

The proof sketch is as follows:

(a) There is a high probability that the random sample is "good", i.e., close to the actual interest of the agents.

(b) Whenever the sample is good for some agent, this agent will report truthfully (under the ϵ-truthfulness assumption).

(c) When the sample is good for all agents, the risk of Mechanism 8 is close to the risk of Mechanism 5, and thus we have almost a $3 - \frac{2}{n}$-approximation.

(d) Otherwise the risk can be high, but this has a small effect on the total expected risk, as it occurs with low probability.

We prove Theorem 4.10 along these lines in Appendix B.2, and supply an exact upper bound on the number of samples required for the theorem to hold.

4.3.2 The Pure Rationality Assumption

Recall that under the pure rationality assumption, an agent will always use a dominant strategy, when one exists. We once again consider the performance of Mechanism 8. Note that since our mechanism uses a dictator, each agent i has a weakly dominant strategy. In order to see that, observe that there is some classifier $\hat{c}_i$ that minimizes the risk w.r.t. the whole distribution $\mathcal{D}_i$.[13] The dominant strategy of agent i is to label the sampled dataset X according to $\hat{c}_i$. Note that this does not mean that i is being truthful, as it is possible that $\hat{c}_i(x) \neq Y_i(x)$ (see Remark 3).

Theorem 4.11. *Assume all agents are purely rational, and let $\mathcal{C}$ be any concept class with a bounded dimension. For any $\epsilon > 0$, there is k (polynomial* only *in $\frac{1}{\epsilon}$) s.t. if at least k datapoints are sampled, then the expected risk of Mechanism 8 is at most $\left(3 - \frac{2}{n}\right) \cdot r_{\min} + \epsilon$.*

Interestingly, the alternative assumption improved the sample complexity: the number of required samples no longer depends on n, only on $\frac{1}{\epsilon}$. In a somewhat counter-intuitive way, the rationality assumption provides us with better bounds without using the notion of truthfulness at all. This can be explained by the fact that a *rational* (i.e., self-interested) labeling of the dataset is a better proxy to an agent's real type than a truthful labeling. Indeed, this strange claim is true since the sampling process might produce a set of points X that represents the agent's distribution in an inaccurate way.[14]

5 Discussion

We first review our results in the decision making setting, then in the learning theoretic setting, and finally present some directions for future research.

Decision Making Setting

We started by studying the simple case where there are only two possible decisions. In this setting there is an almost trivial mechanism that is group strategyproof, and guarantees a 3-approximation ratio. While there are no better deterministic mechanisms, we showed how a specific randomization can be used to achieve a 2-approximation ratio, while maintaining the group-SP property.

For the more general case, we showed that a simple randomization of the dictator (the **WRD** mechanism) achieves the best possible approximation ratio when agents have uniform weights, but falls short in the weighted case. We then presented a new mechanism that closes this gap and obtains optimal approximation results in the general case (**CRD**). In the weighted realizable case, we presented a mechanism that matches

[13] There is a fine issue here regarding the finiteness of the concept class, that we deal with in the proof.

[14] As we explained in Remark 3, the revelation principle does not apply here, since the agents do not report their full preferences.

	All Classes (shared inputs)		Binary decision
	general datasets	realizable datasets	
HD	$O(n)$ (Th.4.1)	$\Rightarrow\ O(n)$	$\Rightarrow\ O(n)$
PM	-	-	3 (Th.3.2)
lower bound	$\Omega(n)$ (Th.4.3)	$\Omega(n)$ [13]	3 (Th.3.3)

Table 1: Summary of results (deterministic mechanisms). The corresponding theorem for each result appears in parentheses.

	All Classes (shared inputs)		Binary decision
	general datasets	realizable datasets	
WRD	3 (Th.4.4)	2 (Th.4.4)	$\Rightarrow 3$
SRD	> 2.4 (Prop.4.7)	> 1.39	2 (Th.3.6)
CRD	$3 - \frac{2}{n}$ (Th.4.8)	$2 - \frac{1}{n}$ (Th.4.8)	2
RRD	≥ 3 (Th.4.9)	$2 - \frac{2}{n}$ (Th.4.9)	3
best upper bound	$3 - \frac{2}{n}$ (**CRD**)	$2 - \frac{2}{n}$ (**RRD**)	2 (**SRD**,**CRD**)
lower bound	$3 - \frac{2}{n}$ (Th.4.6 [27])	?	2 (Th.3.4)

Table 2: Summary of results (randomized mechanisms). We conjecture that the upper bound for realizable datasets is tight, but this remains an open question.

the best known results with uniform weights. However it is still an open question whether this bound is tight, as no non-trivial lower bounds are known.

We showed that these approximation results stand in sharp contrast to the deterministic case, where no deterministic mechanism can guarantee a constant approximation ratio. The trivial selection of the heaviest agent as a dictator is the best deterministic SP mechanism at hand. Results also highlight the power of the shared inputs assumption, as they allow us to break the lower bounds that hold in the general case [29].

All these results (summarized in Tables 1 and 2) may help decision makers—both human and automated—in reaching a decision that approximately maximizes social welfare, when data might be biased by conflicting interests.

Implications for Facility location

As we hinted in the introduction, our classification model can be seen as facility location in metric spaces, where the particular space that we use is the binary cube. In fact, the $2 - \frac{2}{n}$ bound in Theorem 4.5 follows directly from a folk result in facility location, and has been employed, for example, by Alon et al. [2]. We will next describe our results in the decision-theoretic setting in the wider context of metric spaces, thereby extending and generalizing the mentioned folk theorem.

Let $\langle \mathcal{F}, d \rangle$ be a metric space.[15] Let $F = \{f_1, \ldots, f_n\}$ be a finite set of points in $\mathcal{F}$, where each point f_i has an attached weight w_i reflecting its importance. Define $d(f, F)$ as the (weighted) average distance from f to F, and let $f^* \in \mathcal{F}$ be the point that minimizes this distance, i.e.,

$$f^* = \text{argmin}_{f \in \mathcal{F}}\, d(f, F) = \text{argmin}_{f \in \mathcal{F}} \sum_{i \leq n} w_i d(f, f_i).$$

We are interested in selecting one of the points in F, that will be as close as possible to all other points. The restriction is that this selection must be "blind". That is, we must select without knowing the actual distances. All we know are the weights of the n points. Clearly, if weights are uniform, one can do no better than simply picking a random point in F. The following inequality, which is a folk theorem, bounds the expected distance achieved in this process.

$$\frac{1}{n} \sum_{i \leq n} d(f_i, F) \leq \left(2 - \frac{2}{n}\right) d(f^*, F). \tag{14}$$

As we informally explained before, the upper bounds on the approximation ratio of the **WRD** mechanism (e.g., the realizable part of Theorem 4.5) can be derived from Eq. (14) by defining a metric over classifiers, reflecting the fraction of the data space on which they disagree. In the uniform-weight, realizable case, the **WRD** mechanism picks an agent at random, and thus its risk is exactly the average distance between each agent's optimal classifier and the other agents. The formal details appear in Appendix B, where we also supply an analog for the non-realizable case, and extend our bounds to weighted agents.

Moreover, the full proofs show that all mechanisms of Section 4 attain the specified approximation ratios in a more general model, where the private labels are non-deterministic, and datasets are given in the form of a (private) *distribution* over $\mathcal{X} \times \{+, -\}$.[16] The theorems in Section 4, under the standard model we presented (with deterministic labels), follow as a special case.

Implications for Partition and Judgment aggregation

Given a subset X of R^d (and in particular an interval), partitions of X just form another metric space. Informally, the distance between two partitions is exactly the volume they disagree on. The set of all partitions that are allowed constitutes the concept class $\mathcal{C}$.

A similar approach to the Judgment aggregation problem requires some additional assumptions, since issues on the agenda cannot always be directly compared and quantified. We will clarify this using the following simple example (the Doctrinal paradox, see e.g. [14]): The agenda contains the three logical expressions $X = (\mathbf{a}, \mathbf{b}, \mathbf{a} \wedge \mathbf{b})$. Legal assignments are those that are also logically consistent (e.g. $(1, 1, 1)$ is legal, but $(1, 1, 0)$ is not). We can therefore naturally define $\mathcal{C}$ as the set of all legal assignments ($|\mathcal{C}| = 4$ in this case). The "dataset" S then contains the opinion of every judge over

[15] It is in fact sufficient to assume that d is a *pseudo-metric*, i.e., it is possible that $d(f, f') = 0$ for $f \neq f'$.

[16] The datasets in Section 3 can be viewed as a single data point with non-deterministic labels. The probabilities of a positive/negative label for agent i are proportional to P_i and N_i, respectively.

the correct assignment. Consistency of the judges' opinions coincides with the requirement that S is individually realizable. The subtle issue is that a-priori, there is no reason to say that, for example, $(1, 1, 0)$ is closer to $(1, 1, 1)$ than to $(0, 0, 0)$. However if we assign a fixed weight to every issue on the agenda (that all judges can agree on) then we have a natural metric, and we are back at the "shared input" setting of Section 4. Our suggested mechanisms can therefore be used to randomize a legal assignment that is close – on average – to the opinions of the judges. It is important to note however that if the judges disagree on the importance of certain issues, then approximation is not well-defined, and even strategyproofness is no longer guaranteed.

Dokow and Holzman [14] characterized those agendas for which (deterministic) non-dictatorial aggregation rules exist.[17] Our randomizations guarantee a constant bound on the social welfare under *any agenda*, but it is likely that under some families of agendas (such as those characterized by Dokow and Holzman), an even better outcome can be guaranteed.

We should mention in this context a recent paper by Nehama [31], which studies approximate judgment aggregation rules from a different angle, without considering incentives or welfare at all. Rather, the paper characterizes rules whose properties (e.g. consistency) only approximately hold. We hope to explore the applicability of similar relaxations to other domains in our future work.

Learning-Theoretic Setting

In all cases where a constant upper bound on the approximation ratio was available, we showed how to use the SP decision mechanism to implement learning mechanisms with a bounded expected risk. More precisely, our mechanisms sample a finite number of data points from a given distribution, which are thereafter labeled by self-interested agents. The expected risk of the mechanism (where expectation is taken over both sampling procedure and internal randomization) is compared to the expected risk (over the given distribution) of the best classifier in the concept class. This allows us to achieve an approximation ratio that is arbitrarily close to the approximation guaranteed in the decision theoretic setting: 2 when there are only two classifiers, and $3 - \frac{2}{n}$ when there are more (provided that all agents sample from the same distribution). When the optimal risk itself is high (say, above $5 - 10\%$) then such results are not very useful. With low optimal risk, a constant approximation ratio of 2 or 3 is quite good, especially since it applies across all concept classes and all distributions.

We made a distinction between alternative game-theoretic assumptions on agents' behavior, showing how the different assumptions affect the mechanism and the number of required samples.

Our results in the learning theoretic setting contribute to the design of algorithms that can function well in non-cooperative environments. We also promote understanding of the underlying assumptions on agents' behavior in such environments, and how these may affect the learning process.

[17] Dokow and Holzman [14] did not require strategyproofness, but different properties that are closely related.

Future Work

Future research may provide answers to some of the questions we left open, and expand this young hybrid field in new directions. More efficient SP mechanisms may be crafted to handle specific concept classes. Further extensions of the SP classification model we presented may be considered: formalizations other than the PAC-like one we suggested; different loss functions; alternative game-theoretic assumptions as well as restrictions on the structure of the dataset. It is also possible to alter the model by allowing different types of strategic behavior, such as misreporting the *location* of the data points rather than their labels.

All of these directions may reveal new parts of the overall picture and promote a better understanding of the conditions under which SP learning can take place effectively. This, in turn, might supply us with new insights regarding our results and regarding their relationship to other areas.

A Proofs of Section 3

Theorem 3.4. *Let $\epsilon > 0$. There is no $(2-\epsilon)$-approximation strategyproof randomized mechanism.*

Proof. We will use the same datasets used in the proof of Theorem 3.3, and illustrated in Figure 3. Let $\mathbf{M}$ be an SP randomized mechanism, and denote by $p_{\mathbf{M}}(c \mid S)$ its probability of outputting c given S.

We first show that the mechanism chooses the positive hypothesis with the same probability in all three datasets.

Lemma A.1. $p_{\boldsymbol{M}}(c_+ \mid S^I) = p_{\boldsymbol{M}}(c_+ \mid S^{II}) = p_{\boldsymbol{M}}(c_+ \mid S^{III})$.

Proof. As in the proof of Theorem 3.3, the agents can make one dataset look like another dataset. If $p_{\mathbf{M}}(c_+ \mid S^I) \neq p_{\mathbf{M}}(c_+ \mid S^{II})$ then agent 2 will report its labels in a way that guarantees a higher probability of c_-. Similarly, $p_{\mathbf{M}}(c_+ \mid S^{II}) \neq p_{\mathbf{M}}(c_+ \mid S^{III})$ implies that agent 1 can increase the probability of c_+ by lying. □

Denote

$$p_+ = p_{\mathbf{M}}(c_+ \mid S^I) = p_{\mathbf{M}}(c_+ \mid S^{II}) = p_{\mathbf{M}}(c_+ \mid S^{III}) \,,$$

and

$$p_- = p_{\mathbf{M}}(c_- \mid S^I) = p_{\mathbf{M}}(c_- \mid S^{II}) = p_{\mathbf{M}}(c_- \mid S^{III}) \,.$$

Without loss of generality $p_+ \geq \frac{1}{2} \geq p_-$. Then:

$$\begin{aligned} \mathrm{R}_I(\mathbf{M}(S^{III}), S^{III}) &= p_+ \mathrm{R}_I(c_+, S^{III}) + p_- \mathrm{R}_I(c_-, S^{III}) \\ &= p_+ \cdot \frac{3t+1}{4t+2} + p_- \cdot \frac{t+1}{4t+2} \\ &\geq \frac{1}{2} \cdot \frac{3t+1}{4t+2} + \frac{1}{2} \cdot \frac{t+1}{4t+2} = \frac{1}{2}, \end{aligned}$$

whereas

$$r^* = \mathrm{R}_I(c_-, S^{III}) = \frac{t+1}{4t+2} \,.$$

For $t > \frac{1}{\epsilon}$ it holds that

$$\frac{\mathrm{R}_I(\mathbf{M}(S^{III}), S^{III})}{r^*} = \frac{4t+2}{2(t+1)} = 2 - \frac{1}{t+1} > 2 - \epsilon \,.$$

As before, if $p_- > p_+$, a symmetric argument shows that $\mathrm{R}_I(\mathbf{M}(S^I), S^I) > (2-\epsilon)r^*$. Therefore no SP mechanism can achieve a $(2-\epsilon)$-approximation, even through randomization. ■

Theorem 3.6. *Mechanism 2 is a group strategyproof 2-approximation randomized mechanism.*

Proof. Similarly to Mechanism 1, Mechanism 2 is clearly group SP, since declaring a false label may only increase the probability of obtaining a classifier that labels correctly less than half of the agent's examples, thus increasing the subjective expected risk.

Assume without loss of generality that $N \geq P$, i.e., that the negative classifier c_- is better. Denote by $w = \frac{N'}{m}$ the total weight of all agents that support c_-.

Lemma A.2. $1 - r^* \leq \frac{1+w}{1-w} r^*$.

Proof. The largest possible number of negative examples is achieved when all the negative agents control only negative examples, and all the positive agents control only a slight majority of positive labels. Formally, $N \leq N' + \frac{P'}{2}$, and thus:

$$1 - r^* = \mathrm{R}_I(c_+) = \frac{N}{m} \leq \frac{N'}{m} + \frac{P'}{2m} = w + \frac{1-w}{2} = \frac{1+w}{2}.$$

It must follow that $r^* = 1 - (1 - r^*) \geq \frac{1-w}{2}$. By dividing the two inequalities, $\frac{1-r^*}{r^*} \leq \frac{1+w}{1-w}$; thus the lemma is proved. □

$$\begin{aligned}
\mathrm{R}_I(\mathbf{SRD}(S), S) &= \frac{w^2 \mathrm{R}_I(c_-, S) + (1-w)^2 \mathrm{R}_I(c_+, S)}{w^2 + (1-w)^2} \\
&= \frac{w^2 r^* + (1-w)^2 (1-r^*)}{w^2 + (1-w)^2} \leq \frac{w^2 r^* + (1-w)^2 \frac{1+w}{1-w} r^*}{w^2 + (1-w)^2} \quad \text{(from Lemma A.2)} \\
&= \frac{w^2 r^* + (1-w)(1+w) r^*}{w^2 + (1-w)^2} = \frac{1}{2w^2 - 2w + 1} r^* \\
&\leq \frac{1}{1/2} r^* = 2r^*,
\end{aligned}$$

where the last inequality holds since $2w^2 - 2w + 1$ has a minimum in $w = \frac{1}{2}$. ■

Theorem 3.7. *Given sampled datasets, assume weak truthfulness. For any $\epsilon > 0$, there is m' (polynomial in $\ln(n)$ and $\frac{1}{\epsilon}$) such that by sampling m' points for each agent, it holds that*

$$R_I(\widetilde{\boldsymbol{SRD}}) \leq 2r_{\min} + \epsilon.$$

Specifically, sampling $m' > 50 \frac{1}{\epsilon^2} \ln(\frac{10n}{\epsilon})$ will suffice.

Proof. In this proof we will differentiate the real risk, as defined for the learning-theoretic setting, from the *empirical* risk on a given sample, as defined in the simple setting. The empirical risk will be denoted by

$$\widehat{\mathrm{R}}_I(c, S) = \frac{1}{m} \sum_{\langle x, y \rangle \in S} [\![c(x) \neq y]\!].$$

Also, to simplify notation we replace $\widetilde{\mathbf{SRD}}$ with just **M** throughout the proof. Note that **M** can equally stand for any other group strategyproof 2-approximation mechanism (including **CRD**, and the mechanism presented in [28]).

Without loss of generality we assume that $r^* = \mathrm{R}_I(c_-) < \mathrm{R}_I(c_+)$. Notice that if $r^* = \mathrm{R}_I(c^*) = \mathrm{R}_I(c_-) > \frac{1}{2} - 3\epsilon$ then any concept our mechanism returns will trivially attain a risk of at most $\frac{1}{2} + 3\epsilon \leq r^* + 6\epsilon$. Therefore, we can assume for the rest of this proof that

$$\mathrm{R}_I(c_-) + 3\epsilon \leq \frac{1}{2} \leq \mathrm{R}_I(c_+) - 3\epsilon. \tag{15}$$

Let us introduce some new notations and definitions. Denote the data set with the real labels by $S_i = \{\langle x_{i,j}, Y_i(x_{i,j})\rangle\}_{j \leq m'}; S = \{S_1, \ldots, S_n\}$. Note that the mechanism has no direct access to S, but only to the reported labels as they appear in $\overline{S}$.

Define G as the event "the empirical and real risk differ by at most ϵ for all agents"; formally:

$$\forall c \in \{c_+, c_-\}, \forall i \in I,\ |\widehat{\mathrm{R}}_i(c, S_i) - \mathrm{R}_i(c)| < \epsilon. \tag{16}$$

Lemma A.3. *Let $\delta > 0$. If $m' > \frac{1}{2\epsilon^2}\ln(\frac{2n}{\delta})$, then with probability of at least $1 - \delta$, G occurs.*

Proof. Fix $i \in I$. Consider the event $Y_i(x) = +$, and its indicator random variable $[\![Y_i(x) = +]\!]$. We can rewrite the empirical and real risk as the sum and the expectation of this variable:

$$\mathrm{R}_i(c_-) = \mathbb{E}_{x \sim \mathcal{D}_X}[[\![Y_i(x) = +]\!]] = \mathbb{E}_{(x,y) \sim \mathcal{D}_i}[[\![y = +]\!]]$$

$$\widehat{\mathrm{R}}_i(c_-, S_i) = \frac{1}{m'} \sum_{(x,y) \in S_i} [\![Y_i(x) = +]\!] = \frac{1}{m'} \sum_{(x,y) \in S_i} [\![y = +]\!]$$

Since S_i is sampled i.i.d. from $\mathcal{D}_i$, the empirical risk is the sum of independent Bernoulli random variables with expectation $\mathrm{R}_i(c_-)$. We derive from the Chernoff bound that for any data set of size $|S_i| = m'$:

$$\Pr[|\widehat{\mathrm{R}}_i(c_-, S_i) - \mathrm{R}_i(c_-)| > \epsilon] < 2e^{-2\epsilon^2 m'}$$

Taking $m' > \frac{1}{2\epsilon^2}\ln(\frac{2n}{\delta})$, we get:

$$\begin{aligned}
\Pr[\neg G] &= \Pr[\exists i \in I,\ |\widehat{\mathrm{R}}_i(c_-, S_i) - \mathrm{R}_i(c_-)| > \epsilon] \\
&\leq \sum_{i \in I} \Pr[|\widehat{\mathrm{R}}_i(c_-, S_i) - \mathrm{R}_i(c_-)| > \epsilon] \\
&\leq |I| 2e^{-2\epsilon^2 m'} < n\frac{\delta}{n} = \delta,
\end{aligned}$$

where the first inequality is due to the union bound. □

Note that since

$$|\widehat{\mathrm{R}}_i(c_-, S_i) - \mathrm{R}_i(c_-)| = |\widehat{\mathrm{R}}_i(c_+, S_i) - \mathrm{R}_i(c_+)|,$$

it is enough to show the above for c_-.

If G occurs, then from (16) and the triangle inequality it holds that for all $c \in \{c_+, c_-\}$ and $i \in I$,

$$|\mathrm{R}_I(c) - \widehat{\mathrm{R}}_I(c, S)| \leq \sum_{i \in I} \frac{1}{n} |\mathrm{R}_i(c) - \widehat{\mathrm{R}}_i(c, S)| \leq \epsilon. \tag{17}$$

Using (17) we could have bounded the risk of $\mathbf{M}(S)$, but unfortunately this would not do as the mechanism may only access $\overline{S}$ and not S. In order to bound $\mathrm{R}_I(\mathbf{M}(\overline{S}))$, we need to know, or estimate, how the agents label their examples. To handle this problem, we will first analyze which agents may gain by lying, and then define a new data set $\tilde{S}$ with the following two properties: no agent has motivation to lie (thus we can assess the result of running $\mathbf{M}$ on $\tilde{S}$), and $\tilde{S}, S$ are very similar.

We now divide I into two types of agents: $I' = \{i \in I : \ |\mathrm{R}_i(c_-) - \frac{1}{2}| < \epsilon\}$, and $I'' = I \setminus I'$. For each agent $i \in I$, we denote by P_i, N_i the number of positive/negative examples the agent controls in S_i. Note that $P_i = m'\widehat{\mathrm{R}}_i(c_-, S_i)$. Since $\mathrm{R}_I(c_-) < \mathrm{R}_I(c_+)$ we may assume without loss of generality that all agents $i \in I'$ prefer c_+ (otherwise lying only lowers the expected risk of our mechanism). Agents in I'', on the other hand, cannot benefit by lying, since S_i must reflect i's truthful preferences, and Mechanism 2 (which is used by Mechanism 3 in step 3) is SP.

For each agent i define a new set of examples $\tilde{S}_i$ as follows:

- If $i \in I''$, $\tilde{S}_i = S_i$.
- If $i \in I'$, define $\tilde{P}_i = P_i + \lceil \epsilon m' \rceil$ and let $\tilde{S}_i$ contain $\tilde{P}_i$ positive examples and $m' - \tilde{P}_i$ negative ones.

Lemma A.4. *If G occurs, then for all agents in I*

$$\tilde{N}_i \leq \tilde{P}_i \iff R_i(c_-) \geq R_i(c_+)$$

Proof. If $i \in I''$ then w.l.o.g. $\mathrm{R}_i(c_-) \leq \mathrm{R}_i(c_+) - 2\epsilon$, thus from (16)

$$\begin{aligned} \tilde{P}_i = P_i = m'\hat{\mathrm{R}}_i(c_-, S_i) &\leq m'(\mathrm{R}_i(c_-) + \epsilon) \\ &\leq m'(\mathrm{R}_i(c_+) - \epsilon) \leq m'\hat{\mathrm{R}}_i(c_+, S_i) = N_i = \tilde{N}_i. \end{aligned}$$

If $i \in I'$ then according to our assumption

$$\mathrm{R}_i(c_+) \leq \mathrm{R}_i(c_-) \leq \mathrm{R}_i(c_+) + 2\epsilon.$$

Moreover, by the definition of $\tilde{P}_i$,

$$\tilde{P}_i \geq P_i + m'\epsilon \ ; \ \tilde{N}_i \leq N_i - m'\epsilon.$$

Thus

$$\begin{aligned} \tilde{P}_i \geq P_i + m'\epsilon \ &= m'\hat{\mathrm{R}}_i(c_-, S_i) + m'\epsilon \geq m'\mathrm{R}_i(c_-) \\ &\geq m'\mathrm{R}_i(c_+) \geq m'(\hat{\mathrm{R}}_i(c_+, S_i) - \epsilon) \geq N_i - m'\epsilon \geq \tilde{N}_i. \end{aligned}$$

Lemma A.4 implies that, if G occurs, agents cannot do better than report $\tilde{S}$ under Mechanism 3, since $\tilde{S}_i$ reflects the real preferences of agent i. Now, if agent i reports truthfully, then $\overline{P}_i = P_i$. If i decides to lie, it may report more positive labels, but cannot gain from reporting more than $\tilde{P}_i$ such labels, and, crucially, the mechanism's outcome will not change in this case. The immediate result is that we can assume:

$$P \leq \overline{P} = \sum_{i \in I} \frac{1}{n} \overline{P}_i \leq \sum_{i \in I} \frac{1}{n} \tilde{P}_i = \tilde{P},$$

and, since the expected risk of **M** only increases with the number of positive examples (the probability of Mechanism 3 choosing the positive classifier increases),

$$\mathrm{R}_I(\mathbf{M}(S)) \leq \mathrm{R}_I(\mathbf{M}(\overline{S})) \leq \mathrm{R}_I(\mathbf{M}(\tilde{S})). \tag{18}$$

We can now concentrate on bounding the empirical risk on $\tilde{S}$.

Lemma A.5. *If G occurs,*

$$\forall c \in \{c_+, c_-\},\ |R_I(c) - \hat{R}_I(c, \tilde{S})| \leq 3\epsilon. \tag{19}$$

As in Lemma A.3, it will suffice to show this only for c_-.

Proof. From (16), for $m' > \frac{1}{\epsilon}$,

$$\begin{aligned}\hat{\mathrm{R}}_I(c_-, \tilde{S}) = \frac{\tilde{P}_i}{m'} = \frac{P_i + \lceil m'\epsilon \rceil}{m'} \leq \frac{P_i + m'\epsilon + 1}{m'} \\ \leq \frac{{P_i}'}{m} + 2\epsilon = \hat{\mathrm{R}}_I(c_-, S) + 2\epsilon \leq \mathrm{R}_I(c_-) + \epsilon + 2\epsilon = \mathrm{R}_I(c_-) + 3\epsilon.\end{aligned}$$

From (15) and (19)

$$\widehat{\mathrm{R}}_I(c_-, \tilde{S}) \leq \mathrm{R}_I(c_-) + 3\epsilon \leq \mathrm{R}_I(c_+) - 3\epsilon \leq \widehat{\mathrm{R}}_I(c_+, \tilde{S}) \tag{20}$$

So c_- is also empirically the best concept for $\tilde{S}$; Mechanism 2 guarantees:

$$\widehat{\mathrm{R}}_I(\mathbf{M}(\tilde{S}), \tilde{S}) \leq 2\widehat{\mathrm{R}}_I(c_-, \tilde{S}) \tag{21}$$

Furthermore, since the risk of Mechanism 3 is a convex combination of the risk of c_+, c_-, we get from (19),

$$\mathrm{R}_I(\mathbf{M}(\tilde{S})) \leq \widehat{\mathrm{R}}_I(\mathbf{M}(\tilde{S}), \tilde{S}) + 3\epsilon \tag{22}$$

Finally, by using (18), (22), (21) and (20) in this order, we get that if G occurs:

$$\begin{aligned}\mathrm{R}_I(\mathbf{M}(\overline{S})) \leq \mathrm{R}_I(\mathbf{M}(\tilde{S})) \leq \hat{\mathrm{R}}_I(\mathbf{M}(\tilde{S}), \tilde{S}) + 3\epsilon \leq 2\hat{\mathrm{R}}_I(c_-, \tilde{S}) + 3\epsilon \\ \leq 2(\mathrm{R}_I(c_-) + 3\epsilon) + 3\epsilon = 2r^* + 9\epsilon\end{aligned}$$

If G does not occur, the risk cannot exceed 1. Thus by applying Lemma A.3 with $\delta = \epsilon = \frac{\epsilon'}{10}$ we find that for $m' > 50\frac{1}{\epsilon'^2}\ln(\frac{10n}{\epsilon'})$:

$$\mathrm{R}_I(\widetilde{\mathbf{SRD}}) \leq \Pr[G](2r^* + 9\epsilon) + \Pr[\neg G)]1 \leq 2r^* + 9\epsilon + \epsilon \leq 2r^* + \epsilon',$$

as required. ■

B Proofs of Section 4

B.1 Proofs of Upper Bounds under Shared Inputs (Sections 4.1, 4.2)

We formulate and prove our results in a somewhat more general model, in which the preferences of each agent are encoded by a distribution, rather than a deterministic function. The new model extends the one presented in Section 4 with two components: (a) some data points may receive more attention than others; (b) the preferences of each agent can reflect uncertainty, or indeterminism, regarding the label of a specific data point. The theorems in Section 4 follow easily as a special case. In addition, the use of distributions makes the proofs in the generalization section (Section 4.3) easier and more natural.

For that purpose we replace the profile of finite datasets $S = \langle S_1, \ldots, S_n \rangle$ with a profile of distributions $F = \langle F_1, \ldots, F_n \rangle$ over $\mathcal{X} \times \{-, +\}$. The marginal of all distributions over $\mathcal{X}$ is the same. We denote this marginal by $F_{\mathcal{X}}$, and take it as a measure of the interest that the agents have in different parts of the input space. Let $\mathcal{H}$ be the set of all deterministic functions $h : \mathcal{X} \to \{-, +\}$. In particular, $\mathcal{C} \subseteq \mathcal{H}$.

We adjust the definition of the private and global risk to handle distributions.

The private risk of $h \in \mathcal{H}$ to agent i w.r.t. the profile F is thus defined as

$$\mathrm{R}_i(h, F) = \mathbb{E}_{\langle x,y \rangle \sim F_i} \left[[\![h(x) \neq y]\!] \right].$$

As usual, the global risk is defined as

$$\mathrm{R}_I(h, F) = \sum_{i \in I} w_i \mathrm{R}_i(h, F).$$

As with discrete datasets, F_i is said to be realizable w.r.t. a concept class $\mathcal{C} \subseteq \mathcal{H}$ if there is a concept $c \in \mathcal{C}$ such that $\mathrm{R}_i(c, F_i) = 0$.

Every distribution p on $\mathcal{X} \times \{-, +\}$ induces a non-deterministic function f_p from $\mathcal{X}$ to labels. Formally, $\Pr(f_p(x) = + | x) = \mathbb{E}_{\langle x,y \rangle \sim p} [[\![y = +]\!] | x]$, and for convenience we denote this probability by $\overline{f}_p(x) \in [0, 1]$. Similarly,

$$\underline{f}_p(x) = 1 - \overline{f}_p(x) = \Pr(f_p(x) = - | x) = \mathbb{E}_{\langle x,y \rangle \sim p} \left[[\![y = -]\!] \mid x \right].$$

We denote by $\mathcal{F}$ the set of all such non-deterministic functions. Note that $\mathcal{H} \subset \mathcal{F}$, and thus every concept class $\mathcal{C}$ is also a subset of $\mathcal{F}$.

A special case is when $p = F_i$, in which case $f_i \equiv f_p$ conveys the preferences of agent i. We assume that agents' preferences are independent; thus for every two agents $i \neq j$, for every $x \in \mathcal{X}$ and every $y, y' \in \{-, +\}$,

$$\Pr(f_i(x) = y, f_j(x) = y' \mid x) = \Pr(f_i(x) = y \mid x) \Pr(f_j(x) = y' \mid x). \tag{23}$$

Definition B.1. *We define the* distance *between two classifiers (w.r.t. a fixed distribution $F_X \in \Delta(\mathcal{X})$), as the part of space they label differently. Formally:*

$$d(f, f') = d_{F_X}(f, f') = \mathbb{E}_{x \sim F_X} \left[\Pr(f(x) \neq f'(x) \mid x) \right]. \tag{24}$$

Let $\mathcal{C} \subseteq \mathcal{H}$ any concept class, then the following holds.

$$\forall c \in \mathcal{C}, \forall j \in I, \;\; d(f_j, c) = \mathrm{R}_j(c, F). \tag{25}$$

The proof of Equation (25) is as follows.

$$\begin{aligned}
\mathrm{R}_j(c,F) \equiv& \mathbb{E}_{\langle x,y\rangle \sim F_j}\left[[\![c(x) \neq y]\!]\right] = \mathbb{E}_{x\sim F_X}\left[\sum_{y\in\{-,+\}} \Pr_{y\sim F_j}(y \mid x)[\![c(x)\neq y]\!]\right] \\
=& \mathbb{E}_{F_X}\left[\underline{f}_j(x)[\![c(x)\neq -]\!] + \overline{f}_j(x)[\![c(x)\neq +]\!]\right] \\
=& \mathbb{E}_{F_X}\left[\Pr(f_j(x) = - \mid x)[\![c(x)\neq -]\!] + \Pr(f_j(x)=+\mid x)[\![c(x)\neq +]\!]\right] \\
=& \mathbb{E}_{F_X}\left[\Pr(f_j(x)=-, c(x)=+\mid x) + \Pr(f_j(x)=+, c(x)=-\mid x)\right] \\
=& \mathbb{E}_{F_X}\left[\Pr(f_j(x)\neq c(x)\mid x)\right] = d(c,f_j) \qquad \text{(from (24).)}
\end{aligned}$$

Recall that $c_i = \operatorname{argmin}_{c\in\mathcal{C}} \mathrm{R}_i(c,F)$ and $c^* = \operatorname{argmin}_{c\in\mathcal{C}} \mathrm{R}_I(c,F)$.
As a special case of Equation (25), we get that

$$\forall i, j \; \left(d(c_i, f_j) = \mathrm{R}_j(c_i, F)\right). \tag{26}$$

$$\forall i \in I \; \left(c_i = \operatorname{argmin}_{c\in\mathcal{C}} d(c, f_i)\right). \tag{27}$$

The following lemma can be seen as a formalization of the statement that our decision-making setting is equivalent to facility location in some metric space (the binary cube).

Lemma B.2. *d is reflexive, non-negative, symmetric and satisfies the triangle inequality.*

Proof. Non-negativity and symmetry are trivial.
$d(f,f) = \mathbb{E}_{x\sim F_X}\left[\Pr(f(x)\neq f(x)\mid x)\right] = \mathbb{E}_{x\sim F_X}[0] = 0$, thus it is reflexive as well.
We prove the triangle inequality. Let $f, f', f'' \in \mathcal{F}$. Note that disagreement of f and f'' requires that at least one of them disagrees with f'; thus for all $x \in \mathcal{X}$

$$\begin{aligned}
\Pr(f(x)\neq f''(x)\mid x) =& \Pr(f(x)\neq f'(x), f'(x) = f''(x)\mid x) \\
&+ \Pr(f(x)=f'(x), f'(x)\neq f''(x)\mid x) \\
\leq& \Pr(f(x)\neq f'(x)\mid x) + \Pr(f'(x)\neq f''(x)\mid x),
\end{aligned}$$

and therefore

$$\begin{aligned}
d(f,f'') =& \mathbb{E}_{x\sim F_X}\left[\Pr(f(x)\neq f''(x)\mid x)\right] \\
\leq& \mathbb{E}_{x\sim F_X}\left[\Pr(f(x)\neq f'(x)\mid x) + \Pr(f'(x)\neq f''(x)\mid x)\right] \\
=& \mathbb{E}_{x\sim F_X}\left[\Pr(f(x)\neq f'(x)\mid x)\right] + \mathbb{E}_{x\sim F_X}\left[\Pr(f'(x)\neq f''(x)\mid x)\right] \\
=& d(f,f') + d(f',f'').
\end{aligned}$$

Thus the triangle inequality holds. □

Lemma B.3.

$$\sum_{i\in I} w_i R_I(c_i, F) = \sum_i \sum_j w_i w_j d(c_i, f_j)$$

Proof.

$$\begin{aligned}\sum_{i\in I} w_i \mathrm{R}_I(c_i, F) =& \sum_i w_i \mathrm{R}_I(c_i, F) \\ =& \sum_i w_i \left(\sum_j w_j \mathrm{R}_j(c_i, F) \right) = \sum_i \sum_j w_i w_j d(c_i, f_j). \qquad \square\end{aligned}$$

Lemma B.4.

$$\sum_i \sum_j w_i w_j d(f_i, f_j) \leq (2 - 2w_{\min})\, r^*$$

Proof.

$$\begin{aligned}\sum_i \sum_j w_i w_j d(f_i, f_j) &= \sum_i \sum_{j\neq i} w_i w_j d(f_i, f_j) && \text{(since } d(f_i, f_i) = 0\text{)} \\ &\leq \sum_i \sum_{j\neq i} w_i w_j (d(f_i, c^*) + d(c^*, f_j)) + && \text{(Triangle Inequality)} \\ &= \sum_i w_i d(f_i, c^*) \sum_{j\neq i} w_j + \sum_i w_i \sum_{j\neq i} w_j d(f_j, c^*) \\ &= \sum_i w_i d(f_i, c^*)(1 - w_i) + \sum_i w_i \left(\sum_j w_j d(f_j, c^*) - w_i d(f_i, c^*) \right) \\ &= \sum_i w_i \left(d(f_i, c^*)(1 - w_i) + r^* - w_i d(f_i, c^*) \right) \\ &\leq \sum_i w_i \left(d(f_i, c^*)(1 - w_{\min}) + r^* - w_{\min} d(f_i, c^*) \right) \\ &= (1 - w_{\min}) \sum_i w_i \left(d(f_i, c^*) + r^* \sum_i w_i - w_{\min} \sum_i w_i d(f_i, c^*) \right) \\ &= (1 - w_{\min}) r^* + r^* - w_{\min} r^* = (2 - 2w_{\min}) r^*. && \square\end{aligned}$$

Note that Equation (14) is derived as a special case of the lemma when weights are uniform.

We can now use these lemmas to bound the approximation ratio of our mechanism in this extended setting. We begin with the simpler, deterministic mechanism.

Theorem 4.1'. *Let $|I| = n$. For every concept class $\mathcal{C}$ and any* profile *F, Mechanism 4 is an SP $(2n-1)$-approximation mechanism.*

Proof. We first find a lower bound on r^*:

$$\begin{aligned} r^* =& \mathrm{R}_I(c^*, F) = \sum_{i \in I} w_i \mathrm{R}_i(c^*, F) = \sum_{i \in I} w_i d(c^*, f_i) && (28) \\ \geq& w_j d(c^*, f_j) \geq \frac{1}{n} d(c^*, f_j) && \text{(since } j \text{ is heaviest)} \end{aligned}$$

Then we upper bound the risk of c_j:

$$\begin{aligned} \mathrm{R}_I(\mathbf{HD}(F), F) =& \mathrm{R}_I(c_j, F) = \sum_{i \in I} w_i d(c_j, f_i) = w_j d(c_j, f_j) + \sum_{i \neq j} w_i d(c_j, f_i) \\ \leq& w_j d(c^*, f_j) + \sum_{i \neq j} w_i (d(c_j, c^*) + d(c^*, f_i)) \\ & && \text{(from the triangle inequality)} \\ =& d(c_j, c^*) \sum_{i \neq j} w_i + \sum_{i \in I} w_i d(c^*, f_i) = d(c_j, c^*) \sum_{i \neq j} w_i + r^* \\ \leq& d(c_j, c^*) \frac{n-1}{n} + r^* && (w_j \geq \tfrac{1}{n}) \\ \leq& r^* + \frac{n-1}{n} (d(c_j, f_j) + d(f_j, c^*)) && \text{(triangle inequality)} \\ \leq& r^* + \frac{n-1}{n} 2 d(c^*, f_j) && \text{(from (27))} \\ \leq& r^* + \frac{n-1}{n} 2n \cdot r^* && \text{(from (28))} \\ =& r^* + (n-1) 2 r^* = (2n-1) r^* && \blacksquare \end{aligned}$$

Theorem 4.4'. *For every concept class $\mathcal{C}$ and for any dataset S, Mechanism 5 is an SP $(3 - 2w_{\min})$-approximation mechanism, where $w_{\min} = \min_{i \in I} w_i$. Moreover, if S is individually realizable, then $(2 - 2w_{\min})$-approximation is guaranteed.*

Proof. Using the lemmas above,

$$\begin{aligned}
\mathrm{R}_I(\mathbf{WRD}(F), F) =& \sum_{i\in I} w_i \mathrm{R}_I(c_i, F) = \sum_i \sum_j w_i w_j d(f_i, c_j) \\
\leq& \sum_i \sum_j w_i w_j (d(f_i, f_j) + d(f_j, c_j)) \qquad \text{(Triangle Inequality)} \\
\leq& \sum_i \sum_j w_i w_j (d(f_i, f_j) + d(f_j, c^*)) \qquad \text{(from (27))} \\
=& \sum_i \sum_j w_i w_j d(f_i, f_j) + \sum_j w_j d(f_j, c^*) \sum_i w_i \\
\leq& (2 - 2w_{\min}) r^* + \sum_j w_j d(f_j, c^*) \qquad \text{(from Lemma B.4)} \\
=& (2 - 2w_{\min}) r^* + \sum_j w_j \mathrm{R}_j(c^*, F) \qquad \text{(from (25))} \\
=& (2 - 2w_{\min}) r^* + \mathrm{R}_I(c^*, F) = (3 - 2w_{\min}) r^*
\end{aligned}$$

Further, if we have an individually realizable profile F', then for any agent j, $d(f_j, c_j) = \mathrm{R}_j(c_j, F') = 0$ (from (25)), in which case

$$\mathrm{R}_I(\mathbf{WRD}(F'), F') = \sum_{i\in I} w_i \mathrm{R}_I(c_i, F') \leq \sum_i \sum_j w_i w_j d(f_i, f_j) \leq (2 - 2w_{\min}) r^* \,.$$

Thus the proof of Theorem 4.4' (and Theorem 4.4 as a special case) is complete. ■

Proposition 4.7. *There is a dataset S with three agents, such that*

$$R_I(\textbf{\textit{SRD}}(S), S) > 2.4 \cdot r^* > \left(3 - \frac{2}{n}\right) r^*.$$

Example B.5. *We set our concept class to $\mathcal{C} = \{c_-, c_+\}$. Assume w.l.o.g. that an agent that is indifferent between the concepts dictates the c_- concept. Let S_1, S_2 be all positive. S_3 contains exactly half negative samples. We set agents' weights as follows: $w_1 = w_2 = 0.29$, and $w_3 = 0.42$.*

*Observe first that $R_I(c_-, S) = 0.79$, whereas $r^* = R_I(c_+, S) = 0.21$. However, the **SRD** mechanism selects agent 3 (and thus the concept c_-) with probability of $\frac{0.42^2}{0.29^2+0.29^2+0.42^2} \cong 0.511$. Therefore,*

$$R_I(\textbf{\textit{SRD}}(S), S) > 0.51 \cdot 0.79 + 0.49 \cdot 0.21 = 0.5058 > 2.4 \cdot 0.21 = 2.4 \cdot r^*,$$

which proves the lower bound. ◊

Proposition B.6. *There is an individually realizable dataset S with three agents, such that*

$$R_I(\textbf{\textit{SRD}}(S), S) > 1.39 \cdot r^* > \left(2 - \frac{2}{n}\right) r^*.$$

Example B.7. *We keep $\mathcal{C} = \{c_-, c_+\}$. Let S_1, S_2 be all positive, and S_3 be all negative. We set agents' weights as follows: $w_1 = w_2 = 0.363$, and $w_3 = 0.274$.*

*We have that $R_I(c_-, S) = 0.763$, and $r^* = R_I(c_+, S) = 0.274$. The **SRD** mechanism selects agent 3 with probability of $\frac{0.274^2}{0.363^2+0.363^2+0.274^2} \cong 0.222$. Therefore,*

$$R_I(\textbf{SRD}(S), S) > 0.222 \cdot 0.763 + 0.778 \cdot 0.274 > 0.382 > 1.39 \cdot 0.274 = 1.39 \cdot r^*,$$

which proves the lower bound for the realizable case. ◊

Theorem 4.8'. *The following hold for Mechanism 6, w.r.t. any* profile F*:*

- $\alpha_{\mathbf{w}} \leq 2 - \frac{2}{n}$.
- ***CRD** has an approximation ratio of $\alpha_{\mathbf{w}} + 1$, i.e., at most $\mathbf{3 - \frac{2}{n}}$.*
- *if S is individually realizable, then the approximation ratio is $\frac{\alpha_{\mathbf{w}}}{2} + 1$, i.e., at most $2 - \frac{1}{n}$.*

Proof. We first prove that $\alpha_{\mathbf{w}} \leq 2 - \frac{2}{n}$.

Let $g(x) = \frac{1}{2-2x}$. Note that g is convex. Also, since $\sum_{i \in I} w_i = 1$, we have that

$$\frac{1}{n} \leq \sum_{i \in I} w_i^2 \leq 1. \tag{29}$$

$$\begin{aligned}
(\alpha_{\mathbf{w}})^{-1} &= \sum_{i \in I} p'_i = \sum_{i \in I} w_i \frac{1}{2 - 2w_i} = \sum_{i \in I} w_i g(w_i) \\
&\geq g\left(\sum_{i \in I} w_i \cdot w_i\right) = \frac{1}{2 - 2\sum_{i \in I} w_i^2} && \text{(from Jensen's inequality)} \\
&\geq \frac{1}{2 - 2(1/n)}, && \text{(from (29))}
\end{aligned}$$

thus $\alpha_{\mathbf{w}} \leq 2 - \frac{2}{n}$.

We denote by $d(f, f')$ the number of disagreements between f and f'. f_i, c_i denote the labels of agent i, and the classifier in $\mathcal{C}$ that is the closest to them (i.e., $c \in \mathcal{C}$ that minimizes $d(c, f_i)$). For any c, it holds that

$$\mathrm{R}_I(c, F) = \sum_{i \in I} w_i \mathrm{R}_i(c, F) = \sum_{i \in I} w_i d(c, f_i).$$

Note that for all i, $d(c_i, c^*) \leq 2d(f_i, c^*)$, since otherwise c^* is closer to f_i than c_i.

$$\begin{aligned}
\mathrm{R}_I(\mathbf{CRD}(F), F) &= \sum_{i\in I} p_i \mathrm{R}_I(c_i, F) = \sum_{i\in I} p_i \sum_{j\in I} w_j d(c_i, f_j) \\
&= \sum_{i\in I} \left(\sum_{j\neq i} p_i w_j d(c_i, f_j) + p_i w_i d(c_i, f_i) \right) \\
&\leq \sum_{i\in I} \left(\sum_{j\neq i} p_i w_j (d(c_i, c^*) + d(c^*, f_j)) + p_i w_i d(c^*, f_i) \right) \\
&= \sum_{i\in I} p_i d(c_i, c^*) \sum_{j\neq i} w_j + \sum_{i\in I}\sum_{j\in I} p_i w_j d(c^*, f_j) \\
&= \alpha_{\mathbf{w}} \sum_{i\in I} \frac{w_i}{2(1-w_i)} d(c_i, c^*)(1-w_i) + \sum_{j\in I} w_j d(c^*, f_j) \sum_{i\in I} p_i \\
&\leq \alpha_{\mathbf{w}} \sum_{i\in I} \frac{w_i}{2} 2d(f_i, c^*) + \sum_{j\in I} w_j d(c^*, f_j) \\
&= (\alpha_{\mathbf{w}} + 1) \sum_{j\in I} w_j d(c^*, f_j) = (\alpha_{\mathbf{w}} + 1)\mathrm{R}_I(c^*, F) \\
&\leq \left(3 - \frac{2}{n}\right) r^*
\end{aligned}$$

Now, in the realizable case, $f_i = c_i$ for all i.

$$\begin{aligned}
\mathrm{R}_I(\mathbf{CRD}(F), F) &= \sum_{i\in I} p_i \mathrm{R}_I(c_i, F) = \sum_{i\in I} p_i \sum_{j\in I} w_j d(f_i, f_j) = \sum_{i\in I} p_i \sum_{j\neq i} w_j d(f_i, f_j) \\
&\leq \sum_{i\in I}\sum_{j\neq i} p_i w_j (d(f_i, c^*) + d(f_j, c^*)) \qquad \text{(T.I.)} \\
&= \sum_{i\in I} p_i d(f_i, c^*) \sum_{j\neq i} w_j + \sum_{i\in I} p_i \sum_{j\neq i} w_j d(f_j, c^*) \\
&= \sum_{i\in I} p_i d(f_i, c^*)(1-w_i) + \sum_{i\in I} p_i (r^*(F) - w_i d(f_i, c^*)) \\
&= \alpha_{\mathbf{w}} \sum_{i\in I} \frac{w_i}{2(1-w_i)} d(f_i, c^*)(1-w_i) + r^*(F) - \sum_{i\in I} p_i w_i d(f_i, c^*) \\
&= \frac{\alpha_{\mathbf{w}}}{2} \sum_{i\in I} w_i d(f_i, c^*) + r^*(F) - \sum_{i\in I} p_i w_i d(f_i, c^*) \\
&= \frac{\alpha_{\mathbf{w}}}{2} r^*(F) + r^*(F) - \sum_{i\in I} p_i w_i d(f_i, c^*) \\
&\leq \left(\frac{\alpha_{\mathbf{w}}}{2} + 1\right) r^*(F) \leq \left(2 - \frac{1}{n}\right) r^*(F),
\end{aligned}$$

which completes the proof. ■

Theorem 4.9. *The following hold for Mechanism 7:*

- $\beta_{\mathbf{w}} \leq 1 - \frac{2}{n}$.
- ***RRD** has an approximation ratio of at most 4, and at least 3 (in the worst case).*
- *if S is individually realizable, then the approximation ratio is $1 + \beta_{\mathbf{w}}$, i.e., at most $\mathbf{2 - \frac{2}{n}}$.*

Proof. Let $q(x) = \frac{1}{1-2x}$. Note that q is convex.

$$\begin{aligned}
(\beta_{\mathbf{w}})^{-1} &= \sum_{i \in I} p'_i = \sum_{i \in I} w_i \frac{1}{1 - 2w_i} = \sum_{i \in I} w_i q(w_i) \\
&\geq q\left(\sum_{i \in I} w_i \cdot w_i\right) = \frac{1}{1 - 2\sum_{i \in I} w_i^2} && \text{(from Jensen's inequality)} \\
&\geq \frac{1}{1 - 2(1/n)}, && \text{(from (29))}
\end{aligned}$$

thus $\beta_{\mathbf{w}} \leq 1 - \frac{2}{n}$.

For the upper bound, we will need the following.

Lemma B.8. *For all $i \in I$, $p_i \leq 2w_i$.*

Proof. Let $h(x) = \frac{x}{1-2x}$. Note that h is convex. Thus by Jensen's inequality

$$\frac{1}{n-1} \sum_{j \neq i} h(w_j) \geq h(\frac{1}{n-1} \sum_{j \neq i} w_j) = h(\frac{1 - w_i}{n-1}). \tag{30}$$

Next,

$$\begin{aligned}
\sum_{j \in I} \frac{w_j}{1 - 2w_j} &= \frac{w_i}{1 - 2w_i} + \sum_{j \neq i} \frac{w_j}{1 - 2w_j} = \frac{w_i}{1 - 2w_i} + \sum_{j \neq i} h(w_j) \\
&\geq \frac{w_i}{1 - 2w_i} + (n-1) h\left(\frac{1 - w_i}{n-1}\right) && \text{(by Eq. (30))} \\
&= \frac{w_i}{1 - 2w_i} + (n-1) \frac{\frac{1-w_i}{n-1}}{1 - 2\frac{1-w_i}{n-1}} = \frac{w_i}{1 - 2w_i} + \frac{1 - w_i}{1 - 2\frac{1-w_i}{n-1}} \\
&\geq \frac{w_i}{1 - 2w_i} + \frac{1/2}{1 - 2\frac{1/2}{n-1}} = \frac{w_i}{1 - 2w_i} + \frac{1}{2\frac{n-2}{n-1}} > \frac{w_i}{1 - 2w_i} + \frac{1}{2}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
p_i = \beta_{\mathbf{w}} p'_i &= \left(\sum_{j \in I} \frac{w_j}{1 - 2w_j}\right)^{-1} \frac{w_i}{1 - 2w_i} < \frac{1}{\frac{w_i}{1-2w_i} + \frac{1}{2}} \cdot \frac{w_i}{1 - 2w_i} \\
&= \frac{w_i}{w_i + \frac{1-2w_i}{2}} = \frac{w_i}{w_i - w_i + \frac{1}{2}} = 2w_i. && \square
\end{aligned}$$

We now bound the risk of **RRD**. We skip some steps that are detailed in the upper bound proof of the **CRD** mechanism.

$$
\begin{aligned}
\mathrm{R}_I(\mathbf{RRD}(S), S) &= \sum_{i\in I} p_i \mathrm{R}_I(c_i, S) = \sum_{i\in I} p_i d(c_i, c^*) \sum_{j\neq i} w_j + \sum_{i\in I}\sum_{j\in I} p_i w_j d(c^*, f_j) \\
&= \beta_{\mathbf{w}} \sum_{i\in I} \frac{w_i}{1-2w_i} d(c_i, c^*)(1-w_i) + \sum_{j\in I} w_j d(c^*, f_j) \sum_{i\in I} p_i \\
&\leq \beta_{\mathbf{w}} \sum_{i\in I} \frac{2w_i(1-w_i)}{1-2w_i} d(f_i, c^*) + r^*(S) \\
&= \beta_{\mathbf{w}} \sum_{i\in I} \left(\frac{w_i(1-2w_i)}{1-2w_i} d(f_i, c^*) + \frac{w_i}{1-2w_i} d(f_i, c^*) \right) + r^*(S) \\
&= \beta_{\mathbf{w}} \sum_{i\in I} w_i d(f_i, c^*) + \beta_{\mathbf{w}} \sum_{i\in I} \frac{w_i}{1-2w_i} d(f_i, c^*) + r^*(S) \\
&= \beta_{\mathbf{w}} r^*(S) + \beta_{\mathbf{w}} \sum_{i\in I} \frac{w_i}{1-2w_i} d(f_i, c^*) + r^*(S) \leq \sum_{i\in I} p_i d(f_i, c^*) + 2r^*(S) \\
&\leq 2 \sum_{i\in I} w_i d(f_i, c^*) + 2r^*(S) = 2r^*(S) + 2r^*(S) = 4r^*(S).
\end{aligned}
$$

In the realizable case, recall that $f_i = c_i$ for all i.

$$
\begin{aligned}
\mathrm{R}_I(\mathbf{RRD}(S), S) &= \sum_{i\in I} p_i \mathrm{R}_I(c_i, S) = \sum_{i\in I} p_i \sum_{j\in I} w_j d(f_i, f_j) = \sum_{i\in I} p_i \sum_{j\neq i} w_j d(f_i, f_j) \\
&\leq \sum_{i\in I}\sum_{j\neq i} p_i w_j (d(f_i, c^*) + d(f_j, c^*)) \qquad \text{(T.I.)} \\
&= \sum_{i\in I} p_i d(f_i, c^*) \sum_{j\neq i} w_j + \sum_{i\in I} p_i \sum_{j\neq i} w_j d(f_j, c^*) \\
&= \sum_{i\in I} p_i d(f_i, c^*)(1-w_i) + \sum_{i\in I} p_i (r^*(S) - w_i d(f_i, c^*)) \\
&= \beta_{\mathbf{w}} \sum_{i\in I} \frac{w_i}{1-2w_i} d(f_i, c^*)(1-w_i) - \beta_{\mathbf{w}} \sum_{i\in I} \frac{w_i}{1-2w_i} w_i d(f_i, c^*) + r^*(S) \\
&= \beta_{\mathbf{w}} \sum_{i\in I} \frac{w_i(1-2w_i)}{1-2w_i} d(f_i, c^*) + r^*(S) \\
&= \beta_{\mathbf{w}} \sum_{i\in I} w_i d(f_i, c^*) + r^*(S) = \beta_{\mathbf{w}} r^*(S) + r^*(S) \\
&= (1+\beta_{\mathbf{w}})\, r^*(S) \leq \left(2 - \frac{2}{n}\right) r^*(S),
\end{aligned}
$$

which proves the upper bound. ■

B.2 Proofs of Generalization Results (Section 4.3)

As in Section 3.3, we distinguish the notation $\mathrm{R}(c)$ (the risk w.r.t. the fixed input distribution) from $\widehat{\mathrm{R}}(c, S)$ (the empirical risk, w.r.t. the sampled dataset S). For the proofs in this section, we will also need the following fundamental result from machine learning theory.

Theorem B.9 (Vapnik and Chervonenkis [41])**.** *Let m be s.t.*

$$m > \frac{V_C}{\epsilon^2} \log\left(\frac{V_C}{\epsilon^2 \delta}\right).$$

Let S be a dataset that contains m data points sampled i.i.d. from a distribution $\mathcal{D} \in \Delta(\mathcal{X} \times \mathcal{Y})$. Then with probability of at least $1 - \delta$,

$$\forall c \in \mathcal{C} \left(|R(c) - \widehat{R}(c, S)| < \epsilon \right) \tag{31}$$

where V_C is a constant which depends only *on the concept class $\mathcal{C}$, and not on the distribution $\mathcal{D}$ or on any other property of the problem.*

V_C is known as the *VC-dimension* of $\mathcal{C}$, introduced in [41]. We do not give a formal definition of V_C here. However, detailed and accessible overviews of both VC theory and PAC learning are abundant (for example, [12]). While V_C may be very large, or even infinite in some cases, it is known to be finite for many commonly used concept classes (e.g., linear classifiers).

Theorem 4.10. *Assume all agents are ϵ-truthful, and let $\mathcal{C}$ be any concept class with a bounded dimension. For any $\epsilon > 0$, there is k (polynomial in $\frac{1}{\epsilon}$ and $\ln(n)$) s.t. if at least k datapoints are sampled, then the expected risk of Mechanism 8 is at most $\left(3 - \frac{2}{n}\right) \cdot r_{\min} + \epsilon$.*

Proof. Let $S_i = \langle X, Y_i(X) \rangle$ be the partial dataset of agent i, with its true private labels. Denote by $Q_i = Q_i(\epsilon)$ the event that

$$\forall c \in \mathcal{C} \left(|\mathrm{R}_i(c) - \widehat{\mathrm{R}}_i(c, S)| < \epsilon \right). \tag{32}$$

We emphasize that Q_i is a property of S, i.e., for some random samples S the event Q_i holds, whereas for others it does not hold. Our proof sketch can now be reformulated as follows:

(a) Q_i happens for all i simultaneously with high probability.

(b) Whenever Q_i occurs, agent i will report truthfully (under the ϵ-truthfulness assumption).

(c) When all Q_i occur, the risk of Mechanism 8 is bounded by $\left(3 - \frac{2}{n}\right) \cdot r_{\min} + \epsilon$.

(d) Otherwise the risk can be high, but this has a small effect on the total expected risk.

Let $\delta > 0$. As S_i is an i.i.d. random sample from $\mathcal{D}_i$, then from Theorem B.9 every Q_i occurs with probability of at least $1 - \delta$ (provided that there are enough samples). Also, from the union bound the probability of the event $\forall j Q_j$ is at least $1 - \delta'$, where $\delta = \frac{\delta'}{n}$.

Lemma B.10. *If Q_i occurs, then agent i can gain at most 2ϵ by lying.*

Proof. Assume agent i is selected by the mechanism, otherwise it is trivially true.

We denote by $\hat{c}_i \in \mathcal{C}$ the concept returned by the mechanism when i reports truthfully, i.e., $\hat{c}_i = \text{argmin}_{c \in \mathcal{C}} \widehat{\mathrm{R}}_i(c, S_i)$.

Let any $c' \in \mathcal{C}$,

$$\begin{aligned}
\mathrm{R}_i(\hat{c}_i) - \mathrm{R}_i(c') =& \mathrm{R}_i(\hat{c}_i) - \widehat{\mathrm{R}}_i(\hat{c}_i, S_i) + \widehat{\mathrm{R}}_i(\hat{c}_i, S_i) - \mathrm{R}_i(c') \\
\leq& |\mathrm{R}_i(\hat{c}_i) - \widehat{\mathrm{R}}_i(\hat{c}_i, S_i)| \\
&+ |\widehat{\mathrm{R}}_i(c', S_i) - \mathrm{R}_i(c')| \qquad \text{(since } \hat{c}_i \text{ is empirically optimal)} \\
<& \epsilon + \epsilon = 2\epsilon, \qquad \text{(from (32))}
\end{aligned}$$

□

By Lemma B.10, i cannot gain more than 2ϵ by reporting c'. By taking $\epsilon < \frac{\epsilon'}{2}$, we complete the proof of parts (a) and (b) from the proof sketch.

Now, for part (c), we assume $\forall i Q_i$. Thus, from Lemma B.10 and the ϵ-truthfulness assumption, all agents are truthful (i.e., $\overline{S} = S$).

Lemma B.11. *If S holds that Q_i occurs for* all $i \in I$*, then*

$$\widehat{R}_I(c^*(S), S) \leq r_{\min} + \epsilon,$$

where $c^*(S) =$ argmin$_{c \in \mathcal{C}} \widehat{R}_I(c, S)$.

Proof. For *any* $c \in \mathcal{C}$, $|\mathrm{R}_i(c) - \widehat{\mathrm{R}}_i(c, S_i)| < \epsilon$, from Equation (32). Therefore

$$\begin{aligned}
\widehat{\mathrm{R}}_I(c^*(S), S) \leq& \widehat{\mathrm{R}}_I(c, S) = \sum_{i \in I} p_i \widehat{\mathrm{R}}_i(c, S) = \sum_{i \in I} p_i \widehat{\mathrm{R}}_i(c, S_i) \\
<& \sum_{i \in I} p_i (\mathrm{R}_i(c) + \epsilon) = \mathrm{R}_I(c) + \epsilon,
\end{aligned}$$

and in particular $\widehat{\mathrm{R}}_I(c^*(S), S) < r_{\min} + \epsilon$. □

We now bound the expected risk of the mechanism. We denote by $c_{\mathbf{M}}(S)$ the (random) classifier that is returned by Mechanism 6 on the input S. For any random variable A, $\mathbb{E}_{\mathbf{M}}[A \mid S]$ is the expectation of A over the random dictator selection for a fixed dataset S. Similarly, $\mathbb{E}_S[A \mid i]$ is the expectation of A over the random sampling, given that i is the selected dictator.

$$\mathbb{E}\left[\mathrm{R}_I(c_{\mathbf{M}}(S)) \mid \forall j Q_j\right] = \mathbb{E}_S\left[\mathbb{E}_{\mathbf{M}}\left[\mathrm{R}_I(c_{\mathbf{M}}(S)) \mid S\right] \mid \forall j Q_j\right] = \mathbb{E}_{\mathbf{M}}\left[\mathbb{E}_S\left[\mathrm{R}_I(c_{\mathbf{M}}(S)) \mid i, \forall j Q_j\right]\right]$$

(changing the order of randomizations)

$$= \sum_{i \in I} p_i \mathbb{E}_S\left[\mathrm{R}_I(\hat{c}_i(S)) \mid i, \forall j Q_j\right]$$

$$\leq \sum_{i \in I} p_i \mathbb{E}_S\left[\widehat{\mathrm{R}}_I(\hat{c}_i(S), S_i) + \epsilon \mid i, \forall j Q_j\right] \qquad \text{(from (32))}$$

$$= \sum_{i \in I} p_i \mathbb{E}_S\left[\widehat{\mathrm{R}}_I(\hat{c}_i(S), S_i) \mid i, \forall j Q_j\right] + \epsilon$$

$$= \mathbb{E}_{\mathbf{M}}\left[\mathbb{E}_S\left[\widehat{\mathrm{R}}_I(c_{\mathbf{M}}(S), S) \mid i, \forall j Q_j\right]\right] + \epsilon$$

$$\leq \mathbb{E}_{\mathbf{M}}\left[\mathbb{E}_S\left[\left(3 - \frac{2}{n}\right)\widehat{\mathrm{R}}_I(c^*(S), S) \mid i, \forall j Q_j\right]\right] + \epsilon \qquad \text{(from Theorem 4.8)}$$

$$\leq \mathbb{E}_{\mathbf{M}}\left[\mathbb{E}_S\left[\left(3 - \frac{2}{n}\right)(r_{\min} + \epsilon) \mid i, \forall j Q_j\right]\right] + \epsilon \qquad \text{(from Lemma B.11)}$$

$$= \left(3 - \frac{2}{n}\right)(r_{\min} + \epsilon) + \epsilon \leq \left(3 - \frac{2}{n}\right) \cdot r_{\min} + 4\epsilon = \left(3 - \frac{2}{n}\right) r_{\min} + \epsilon',$$

which proves part (c) of the proof sketch.

Finally, we bound the total risk of the mechanism, taking part (d) into account.

$$\begin{aligned}
\mathrm{R}_I(\widetilde{\mathbf{CRD}}) =& \mathbb{E}\left[\mathrm{R}_I(c_{\mathbf{M}}(\overline{S}))\right] = \mathbb{E}_S\left[\mathbb{E}_{\mathbf{M}}\left[\mathrm{R}_I(c_{\mathbf{M}}(\overline{S})) \mid S\right]\right] \\
=& \Pr(\forall j Q_j)\mathbb{E}_S\left[\mathbb{E}_{\mathbf{M}}\left[\mathrm{R}_I(c_{\mathbf{M}}(\overline{S})) \mid S\right] \mid \forall j Q_j\right] \\
&+ \Pr(\neg\forall j Q_j)\mathbb{E}_S\left[\mathbb{E}_{\mathbf{M}}\left[\mathrm{R}_I(c_{\mathbf{M}}(\overline{S})) \mid S\right] \mid \neg\forall j Q_j\right] \\
\leq& \mathbb{E}_S\left[\mathbb{E}_{\mathbf{M}}\left[\mathrm{R}_I(c_{\mathbf{M}}(\overline{S})) \mid S\right] \mid \forall j Q_j\right] + \delta' \cdot 1 \\
=& \mathbb{E}_S\left[\mathbb{E}_{\mathbf{M}}\left[\mathrm{R}_I(c_{\mathbf{M}}(S)) \mid S\right] \mid \forall j Q_j\right] + \delta' \\
& \qquad \text{(since all agents are truthful in this case)} \\
& \leq \left(3 - \frac{2}{n}\right) r_{\min} + \delta' + \epsilon' = \left(3 - \frac{2}{n}\right) r_{\min} + \epsilon'',
\end{aligned}$$

as required. ■

We conclude by computing the exact number of samples needed by Mechanism 8 under the ϵ-truthfulness assumption.

Lemma B.12. *If* $k > 64\frac{V_C}{\epsilon^2}\log\left(256\frac{V_C \cdot n}{\epsilon^3}\right)$, *then*

$$R_I(\widetilde{\boldsymbol{CRD}}) \leq \left(3 - \frac{2}{n}\right) r_{\min} + \epsilon.$$

Proof. From Theorem B.9, if $|S_j| > \frac{V_C}{(\epsilon^*)^2}\log\left(\frac{V_C}{(\epsilon^*)^2\delta^*}\right)$, then $\Pr(\neg Q_j(\epsilon^*)) < \delta^*$ and from the union bound it holds that

$$\Pr(\exists j \in I, \neg Q_j(\epsilon^*)) \leq \sum_{j \in I} \neg Q_j(\epsilon^*) < n\delta^*.$$

Taking $\epsilon^* < \frac{\epsilon}{8}$ and $\delta^* < \frac{\epsilon}{4n}$, and unfolding all the residues we used in the proof, we get that

$$\begin{aligned}\mathrm{R}_I(\widetilde{\mathbf{CRD}}) &\leq \left(3-\frac{2}{n}\right) r_{\min} \leq \left(3-\frac{2}{n}\right) r_{\min} + 4\epsilon^* + 2n\delta^* \\ &< \left(3-\frac{2}{n}\right) r_{\min} + 4\frac{\epsilon}{8} + 2n\frac{\epsilon}{4n} = \left(3-\frac{2}{n}\right) r_{\min} + \epsilon,\end{aligned}$$

while

$$\frac{V_C}{(\epsilon^*)^2}\log\left(\frac{V_C}{(\epsilon^*)^2\delta^*}\right) = \frac{V_C}{(\epsilon/8)^2}\log\left(\frac{V_C}{(\epsilon/8)^2(\epsilon/4n)}\right) = 64\frac{V_C}{\epsilon^2}\log\left(256\frac{V_C\cdot n}{\epsilon^3}\right).$$

□

Theorem 4.11. *Assume all agents are purely rational, and let $\mathcal{C}$ be any concept class with a bounded dimension. For any $\epsilon > 0$, there is a k (polynomial* only *in $\frac{1}{\epsilon}$) s.t. if at least k datapoints are sampled, then the expected risk of Mechanism 8 is at most $\left(3-\frac{2}{n}\right) r_{\min} + \epsilon$.*

Proof. Note that the private distributions $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_n$ induce a global joint distribution on the input space, defined as $\mathcal{D} = \sum_{i=1}^{n} w_i \mathcal{D}_i$. We can alternatively define $r_{\min}$ as the minimal risk of any concept w.r.t. the distribution $\mathcal{D}$, i.e., $r_{\min} = \inf_{c\in\mathcal{C}} \mathbb{E}_{(x,y)\sim\mathcal{D}} [\![c(x) \neq y]\!]$. We would like to analyze the outcome of Mechanism 8 and compare the empirical risk to the actual risk. However, we have a technical problem with doing so directly, since S_i (as defined in the proof of Theorem 4.10) is sampled i.i.d. from $\mathcal{D}_i$, but not from $\mathcal{D}$.

In order to prove the theorem, we introduce a *virtual mechanism* (see Mechanism 9). This mechanism generates a *truthful* dataset S, which can be used as an i.i.d. sample from the joint distribution $\mathcal{D}$.

Mechanism 9 The Virtual Learning Mechanism

Sample k data points i.i.d. from $\mathcal{D}_X$ (assume we get the same dataset X as in Mechanism 8).
for each point $x \in X$ **do**
  Select agent i with probability w_i.
  Add $\langle x_j, Y_i(x)\rangle$ to S.
end for
return $c^*(S) = \mathbf{erm}(S)$.

The output of Mechanism 9, $c^*(S)$, is the best concept (in $\mathcal{C}$) for the real dataset S. Note that S is an i.i.d. sample from $\mathcal{D}$, but an actual mechanism such as Mechanism 8 cannot have access to the real labels Y_i—hence the term *virtual* mechanism.

We denote by $T = T(\epsilon)$ the event

$$\mathrm{R}_I(c^*(S)) < r_{\min} + 2\epsilon. \tag{33}$$

Similarly to Q_j in the previous proof, T is a property of S, i.e., its occurrence depends only on the sampling.

Lemma B.13. *If $k = k(\delta, \epsilon)$ is large enough then*

$$\Pr(\neg T) < \delta.$$

Proof. This is an immediate corollary of Theorem B.9. As $c^* = \text{argmin}_{c \in \mathcal{C}} \widehat{\mathrm{R}}_I(c, S)$, $\mathcal{C}$ is of a bounded dimension and S is sampled i.i.d. from $\mathcal{D}$, then for any $c \in \mathcal{C}$

$$\mathrm{R}_I(c^*(S)) < \widehat{\mathrm{R}}_I(c^*(S), S) + \epsilon \leq \widehat{\mathrm{R}}_I(c, S) + \epsilon < \mathrm{R}_I(c) + \epsilon + \epsilon$$

holds with probability of at least $1 - \delta$, for a large enough k. In particular,

$$\Pr(T) = \Pr(\mathrm{R}_I(c^*(S)) < r_{\min} + 2\epsilon) > 1 - \delta.$$

□

It is still not clear how to approximate $c^*(S)$, as our mechanism only has access to $\overline{S}$. For that purpose, we define a new concept class $\mathcal{C}_X \subseteq \mathcal{C}$ as the *projection* of $\mathcal{C}$ on X. Formally, let $H_X \subseteq \mathcal{H}$ be the class of all dichotomies of X, i.e., all h s.t. $h : X \to \{-, +\}$,[18] then $\mathcal{C}_X = \mathcal{C} \cap H_X$. In other words, $\mathcal{C}_X$ contains all dichotomies of X that are also allowed by $\mathcal{C}$.

Denote by $\overline{S}_i$ the dataset with the *reported labels* of agent i, and by $\hat{c}_i$ the best concept w.r.t. to this dataset. That is, $\overline{S}_i = \{\langle x, \overline{Y}_i(x)\rangle\}_{x \in X}$ and $\hat{c}_i = \text{argmin}_{c \in \mathcal{C}} \widehat{\mathrm{R}}_i(c, \overline{S}_i)$. Observe that $c^*(S) \in \mathcal{C}_X$ and $\hat{c}_j \in \mathcal{C}_X$ for all agents. This is the case since both $S, \overline{S}_j$ are labeled versions of the set X. Thus any classifier that is computed w.r.t. S or $\overline{S}_j$ is a dichotomy of X (which minimizes some function that depends on the labels). We define $\tilde{c} = \text{argmin}_{c \in \mathcal{C}_X} \mathrm{R}_I(c)$. Clearly $\mathrm{R}_I(\tilde{c}) \leq \mathrm{R}_I(c^*(S))$, since $c^*(S)$ is also a member of $\mathcal{C}_X$. Thus when T occurs, the inequality

$$\mathrm{R}_I(\tilde{c}) < r_{\min} + 2\epsilon \tag{34}$$

also holds, directly as a special case of (33).

We next show how to approximate $\tilde{c}$ using the generalized variant of Theorem 4.8, as it appears in the appendix. Consider a profile $F = \langle \mathcal{D}_1, \ldots, \mathcal{D}_n \rangle$. This is a valid profile with shared inputs; thus for any concept $c \in \mathcal{C}$, $\mathrm{R}(c) = \mathrm{R}(c, F)$ for private and global risk alike.

Lemma B.14. *Let j be the selected dictator, then*

$$\hat{c}_j = \textit{argmin}_{c \in \mathcal{C}_X} R_j(c) = \textit{argmin}_{c \in \mathcal{C}_X} R_j(c, F).$$

Proof. Recall that $\hat{c}_j \equiv \text{argmin}_{c \in \mathcal{C}} \widehat{\mathrm{R}}_j(c, \overline{S}_j)$. Since we assumed j is purely rational, he will always label all examples in X in a way that will minimize his private risk. From the way Mechanism 8 works, only concepts in $\mathcal{C}_X$ may be returned, and for *any* $c \in \mathcal{C}_X$, there is a labeling of X s.t. c is returned. This labeling $\overline{Y}(c)$ is simply $\forall x \in X\ (\overline{y}(x) = c(x))$. Thus $\text{argmin}_{c \in \mathcal{C}_X} \mathrm{R}_j(c)$ is the best that agent j can hope for, and he can also achieve it by reporting the appropriate labels $\overline{Y}_j$. □

[18] Put differently, H_X is a partition of $\mathcal{H}$ to equilivalence classes, according to their outcome on $X \subseteq \mathcal{X}$.

We now apply Theorem 4.8' on F, using the class $\mathcal{C}_X$, getting

$$\sum_{j\in I} p_j \mathrm{R}_I(\hat{c}_j, F) \leq \left(3 - \frac{2}{n}\right) r^*(F) = \left(3 - \frac{2}{n}\right) \mathrm{R}_I(\tilde{c}). \tag{35}$$

To see why this holds, observe that the left term is the expected risk of Mechanism 6 when the input is the profile F and the concept class $\mathcal{C}_X$; and $\tilde{c}$ is the globally optimal classifier for this input. We emphasize that Equation (35) *always* holds, independently of the sampling or selection.

Finally, we bound the risk of the result concept:

$$\begin{aligned}
\mathrm{R}_I(\widetilde{\mathbf{CRD}}) =& \mathbb{E}_S\left[\mathbb{E}_{\mathbf{M}}\left[\mathrm{R}_I(c_{\mathbf{M}}) \mid S\right]\right] \\
=& \Pr(T)\mathbb{E}_S\left[\mathbb{E}_{\mathbf{M}}\left[\mathrm{R}_I(c_{\mathbf{M}}) \mid S\right] \mid T\right] + \Pr(\neg T)\mathbb{E}_S\left[\mathbb{E}_{\mathbf{M}}\left[\mathrm{R}_I(c_{\mathbf{M}}) \mid S\right] \mid \neg T\right] \\
\leq& \mathbb{E}_S\left[\mathbb{E}_{\mathbf{M}}\left[\mathrm{R}_I(c_{\mathbf{M}}) \mid S\right] \mid T\right] + \delta \cdot 1 && \text{(from Lemma B.13)} \\
=& \mathbb{E}_S\left[\sum_{j\in I} w_j \mathrm{R}_I(\hat{c}_j(S)) \mid T\right] + \delta \\
\leq& \mathbb{E}_S\left[\left(3 - \frac{2}{n}\right) \mathrm{R}_I(\tilde{c}(S)) \mid T\right] + \delta && \text{(from (35))} \\
<& \left(3 - \frac{2}{n}\right) \mathbb{E}_S\left[(r_{\min} + 2\epsilon) \mid T\right] + \delta && \text{(from (34))} \\
=& \left(3 - \frac{2}{n}\right)(r_{\min} + 2\epsilon) + \delta = \left(3 - \frac{2}{n}\right) r_{\min} + 6\epsilon + \delta.
\end{aligned}$$

By taking $\delta = \epsilon = \frac{\epsilon'}{7}$, the proof is complete.

Similarly to Lemma B.12, it follows from Theorem B.9 that taking

$$k > 49\frac{V_C}{\epsilon^2}\log\left(343\frac{V_C}{\epsilon^3}\right)$$

is sufficient for Mechanism 8 to work well under the pure rationality assumption. ■

Acknowledgments

This work was partially supported by Israel Science Foundation grant #898/05, the Israel Ministry of Science and Technology grant #3-6797, and the Google Inter-University Center for Electronic Markets and Auctions. The authors thank Omri Abend, Shaull Almagor, Assaf Michaely and Ilan Nehama for their enlightening comments on drafts of this paper.

Vitae

Reshef Meir is a Ph.D. student in the Computer Science Department in the Hebrew University of Jerusalem, Israel, under the supervision of Prof. Jeffrey S. Rosenschein.

Ariel D. Procaccia is an assistant professor in Carnegie Mellon's Computer Science Department. Before coming to CMU he spent two years as a postdoc in Harvard's Center for Research on Computation and Society, and a year as a postdoc at Microsoft Israel R&D Center. He received his Ph.D. in Computer Science from the Hebrew University of Jerusalem.

Jeffrey S. Rosenschein is the director of the Multiagent Systems Research Group at Hebrew University. He is co-Editor-in-Chief (along with Peter Stone) of the Journal of Autonomous Agents and Multiagent Systems (JAAMAS), and on the Advisory Board of the Journal of Artificial Intelligence Research (JAIR). He is a AAAI Fellow, and previously served as president of the International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS). He was Department Chair of Computer Engineering, within the School of Computer Science and Engineering at Hebrew University and had previously served as Department Chair of Computer Science.

References

[1] N. Alon, M. Feldman, A. D. Procaccia, and M. Tennenholtz. Strategyproof approximation of the minimax on networks. *Mathematics of Operations Research*, 35(3):513–526, 2010.

[2] N. Alon, M. Feldman, A. D. Procaccia, and M. Tennenholtz. Walking in circles. *Discrete Mathematics*, 310(23):3432 – 3435, 2010.

[3] I. Ashlagi, F. Fischer, I. Kash, and A. D. Procaccia. Mix and match. In *Proceedings of the 11th ACM Conference on Electronic Commerce (ACM-EC)*, pages 305–314, 2010.

[4] 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, 2005.

[5] J.-P. Barthèlemy, B. Leclerc, and B. Monjardet. On the use of ordered sets in problems of comparison and consensus of classifications. *Journal of Classification*, 3:187–224, 1986.

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

[7] V. Conitzer and T. Sandholm. Complexity of mechanism design. In *Proceedings of the 18th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 103–110, 2002.

[8] V. Conitzer and T. Sandholm. An algorithm for automatically designing deterministic mechanisms without payments. In *Proceedings of the 3rd International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pages 128–135, 2004.

[9] N. Dalvi, P. Domingos, Mausam, S. Sanghai, and D. Verma. Adversarial classification. In *Proceedings of the 10th International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 99–108, 2004.

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

[11] O. Dekel and O. Shamir. Good learners for evil teachers. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, pages 216–223, 2009.

[12] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag New York, Inc., 1997.

[13] E. Dokow, M. Feldman, R. Meir, and I. Nehama. Mechanism design on combinatorial domains. Working paper, 2011.

[14] E. Dokow and R. Holzman. Aggregation of binary evaluations. *Journal of Economic Theory*, 145:495–511, 2010.

[15] S. Dughmi and A. Ghosh. Truthful assignment without money. In *Proceedings of the 11th ACM Conference on Electronic Commerce (ACM-EC)*, pages 325–334, 2010.

[16] P. Fishburn and A. Rubinstein. Aggregation of equivalence relations. *Journal of Classification*, 3:61–65, 1986.

[17] A. Gibbard. Manipulation of voting schemes. *Econometrica*, 41:587–602, 1973.

[18] M. Guo and V. Conitzer. Strategy-proof allocation of multiple items between two agents without payments or priors. In *Proceedings of the 9th International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pages 881–888, 2010.

[19] M. Guo, V. Conitzer, and D. Reeves. Competitive repeated allocation without payments. In *Proceedings of the 5th International Workshop on Internet and Network Economics (WINE)*, pages 244–255, 2009.

[20] P. Harrenstein, M. M. de Weerdt, and V. Conitzer. A qualitative Vickrey auction. In *Proceedings of the 10th ACM Conference on Electronic Commerce (ACM-EC)*, pages 197–206, 2009.

[21] M. J. Kearns and U. V. Vazirani. *An Introduction to Computational Learning Theory*. MIT Press, 1994.

[22] J. Kemeny. Mathematics without numbers. *Daedalus*, 88:577–591, 1959.

[23] E. Koutsoupias. Scheduling without payments. In *Proceedings of the 4th Symposium on Algorithmic Game Theory (SAGT)*, 2011. To appear.

[24] B. Leclerc. Efficient and binary consensus functions on transitively valued relations. *Mathematical Social Sciences*, 8:45–61, 1984.

[25] D. Lowd, C. Meek, and P. Domingos. Foundations of adversarial machine learning. Manuscript, 2007.

[26] P. Lu, X. Sun, Y. Wang, and Z. A. Zhu. Asymptotically optimal strategy-proof mechanisms for two-facility games. In *Proceedings of the 11th ACM Conference on Electronic Commerce (ACM-EC)*, pages 315–324, 2010.

[27] R. Meir, S. Almagor, A. Michaely, and J. S. Rosenschein. Tight bounds for strategyproof classification. In *Proceedings of the 10th International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pages 319–326, Taipei, Taiwan, 2011.

[28] R. Meir, A. D. Procaccia, and J. S. Rosenschein. Strategyproof classification under constant hypotheses: A tale of two functions. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence (AAAI)*, pages 126–131, 2008.

[29] R. Meir, A. D. Procaccia, and J. S. Rosenschein. On the limits of dictatorial classification. In *Proceedings of the 9th International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pages 609–616, 2010.

[30] B. Mirkin. On the problem of reconciling partitions. In H. Blalock, editor, *Quantitative sociology: international perspectives on mathematical and statistical modeling*. Academic Press, New York, 1975.

[31] I. Nehama. Approximate judgement aggregation. In *Proceedings of the 7th International Workshop on Internet and Network Economics (WINE)*, pages 302–313, 2011.

[32] 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.

[33] A. Othman, E. Budish, and T. Sandholm. Finding approximate competitive equilibria: Efficient and fair course allocation. In *Proceedings of the 9th International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pages 873–880, 2010.

[34] J. Perote and J. Perote-Peña. Strategy-proof estimators for simple regression. *Mathematical Social Sciences*, 47:153–176, 2004.

[35] J. Perote-Peña and J. Perote. The impossibility of strategy-proof clustering. *Economics Bulletin*, 4(23):1–9, 2003.

[36] A. D. Procaccia and M. Tennenholtz. Approximate mechanism design without money. In *Proceedings of the 10th ACM Conference on Electronic Commerce (ACM-EC)*, pages 177–186, 2009.

[37] A. D. Procaccia, A. Zohar, Y. Peleg, and J. S. Rosenschein. The learnability of voting rules. *Artificial Intelligence*, 173(12–13):1133–1149, 2009.

[38] M. Satterthwaite. Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory*, 10:187–217, 1975.

[39] J. Schummer and R. V. Vohra. Mechanism design without money. In N. Nisan, T. Roughgarden, E. Tardos, and V. Vazirani, editors, *Algorithmic Game Theory*, chapter 10. Cambridge University Press, 2007.

[40] L. Valiant. A theory of the learnable. *Communications of the ACM*, 27, 1984.

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

[42] R. Wilson. On the theory of aggregation. *Journal of Economic Theory*, 10:89–99, 1975.