Strategy Proof Classification

A thesis submitted in fulfillment
of the requirements for the degree of
Master of Science

by

Reshef Meir
Student ID: 040097503

supervised by
Jeffrey S. Rosenschein

School of Engineering and Computer Science
The Hebrew University of Jerusalem
Jerusalem, Israel

February 18, 2009

Abstract

We consider the following setting: a decision maker should classify a finite set of data points with binary labels, minimizing the expected error. Subsets of data points are controlled by different selfish agents, which might misreport the labels in order to sway the decision in their favor. We design mechanisms (both deterministic and randomized) that reach an approximately optimal decision and are Strategy-Proof, i.e. agents are best off when they tell the truth. We examine the best approximation ratio that can be achieved using a Strategy-Proof mechanism in various conditions, thereby matching our upper bounds with lower ones. We show that when the approximation ratio is constant, our results can be casted into a classical machine learning classification framework, where the decision maker must learn an approximately optimal classifier based only on a sampled subset of the agents' points.

Acknowledgments

First of all I would like to thank Ariel D. Proccacia for introducing me to this new, exciting research area, for his patient assistance in all aspects my work, and for long hours of fruitful discussions. Many thanks go to my advisor, Prof. Jeffrey S. Rosenschein, who supplied me with both the opportunity and the tools to pursue it. I thank Omri Abend for his enlightening comments on the draft.

My family deserve special thanks for bringing me up with the love for knowledge and education, and for their ongoing encouragement throughout the years.

Last, but not least, I am deeply indebted to my wife Adi, for her insightful notes, technical assistance and moral support. Without her I would not have made it.

Contents

Chapter 1

Introduction

The truth is more important than the facts

Frank Lloyd Wright

In the design and analysis of multiagent systems, one often cannot assume that the agents are cooperative. Rather, the agents are usually self-interested, seeking to maximize their own utility, possibly at the expense of the social good. With the growing awareness of this situation, game-theoretic notions and tools are increasingly brought into play.

In this work we shall consider two interrelated settings. In the first one, a decision has to be made based on data reported by multiple (possibly) selfish agents. The decision affects various points of interest of each agent, and the decision maker tries to decide in a way that will maximize the social welfare, according to the label that agents reported on each of their data points. The decision is expressed in the form of a classifier, which assigns a label for each data point. We emphasize that in this setting there is no notion of learning or generalization. The only requirement is that the decision (classifier) will produce a "correct" label (i.e. consistent with the relevant agent's interest) for as many data points as possible.

The second setting is a variation of the standard supervised classification problem. Samples are drawn from an unknown distribution, and are then labeled by experts. A classification mechanism receives the sampled data as input, and should output a classifier. Unlike the standard setting in machine learning (but similarly to our first setting), the experts are assumed to be selfish agents. The private interests of each agent are characterized by a general distribution function over samples and labels, and the result classifier should be consistent, as much as possible, with the average distribution, taken over all experts. This setting might seem far more complicated than the first, as it involves generalization from partial data. However, we show

how the learning problem effectively reduces to finding a classifier which best fits the available data.

In both settings the decision maker (or mechanism) would like to find the classifier which correctly classifies as much available data as possible. However, the agents might misreport their data (i.e. report a label inconsistent with their private interests) in an attempt to influence the final decision in their favor. The result of decision making based on such biased data may be totally unexpected and difficult to analyze. Mechanisms that encourage truthfull reporting will eliminate any such bias and allow the decision maker to select a classifier which best fits the reported data, without being concerned about agents' hidden interests.

Restricting ourselves to truthful mechanisms sometimes results in a suboptimal classifier, i.e. a classifier that does not minimize the number of errors. Nevertheless, in cases where we can show this classifier to be *approximately* optimal, it is preferred over a biased classifier whose consistency with agents' interests is unknown, and might be far worse. Indeed, most of this work is dedicated to showing upper and lower approximation bounds on such truthful classification mechanisms.

Mechanisms and Algorithms

A full definition of an *algorithm* typically includes much of the implementation details, e.g. how data is stored and accessed, how mathematical functions are computed and so on. Thus, although every algorithm uses some abstractions (e.g. of the syntax), it is still detailed enough to allow analysis of almost every aspect of it. A *mechanism* on the other way, may use far more general abstractions. Most implementation details may be omitted. As a result, some properties of a mechanism (e.g. runtime) cannot be analyzed. We would still like to be able to analyze some of the properties of course. The abstraction level of the mechanism description is thus chosen to allow analysis of exactly those properties in which we are interested, without imposing further restrictions on the implementation. In the context of this work, we are mainly interested in the properties of truthfulness and optimality (or approximate optimality), putting aside computational aspects such as runtime. Our mechanisms sometimes contain redundant steps, which are used only to facilitate the related proofs.

Proving that a mechanism with a (partial) list of requirements does not exist, clearly eliminates any possibility that there is an algorithm that obeys these requirements. Also, it is fairly easy to implement all mechanisms discussed in this work in a straightforward way, although such an implementation may not be very efficient. We sometimes give hints as to how to make these implementations more efficient, but this is not our main focus.

1.1 Motivating examples

Decision making

Consider a common central bank, such as the European Central Bank (ECB). The bank takes decisions which are based on reports from the various national central banks (so one can think of the national central bankers as the agents). The national central bankers, in turn, collect private information, by means of their own institutions, regarding various economic indicators (these are the data points). Naturally, decisions taken at the European level (on, for instance, whether or not to support certain monetary policies) affect all national central banks. This strongly incentivizes the national central bankers to misreport their national statistics in a way that supports a decision they find desirable.

This example demonstrates a case in which interests of agents may vary, but the decision itself is very simple (yes/no) and lacks structure. Consider another situation, in which two or more parties (the agents in this case) are in a conflict regarding the future use of a certain land property (e.g. industrial use vs. agricultural use). The property is abundant with resources in various locations (the data points), and the parties may attribute different (possibly negative) importance to each resource. A neutral arbitrator agrees to hear them out and divide the field in a way that will maximize the average utility of all the involved parties. It is not far fetched to assume that this division has some constraints: for example that the border has to be a straight line or that it has to pass through a specific location. This leaves us with a (large, possibly infinite) set of borders, or classifiers, from which the arbitrator may choose. Knowing how their reported preferences affect the decision, each party may misreport its true evaluation of each resource, in an attempt to achieve a favorable outcome.

Learning

A big organization is trying to fight the congestion on the internal mail system by designing a smart spam filter. In order to train the system, managers are asked to go over their last 1000 incoming e-mails and classify them as either **work related** or **spam**. Whereas some messages (e.g. "buy Viagra now!!!" or "Salary increase for all employees") will probably reach consensus, it is likely that others (e.g. "Joe from the Sales dept. goes on a lunch break") will be considered as work related by only part of the managers, while others will consider it spam. Moreover, as each manager is interested in filtering most of what she sees as spam, a manager might try to compensate for the "mistakes" of her colleagues by misreporting her real opinion on some cases. For example, the manager of R&D dept., 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 just the opposite to prevent her e-mails from being filtered, and so on.

A much more fertile ground for data manipulation is the Internet, which is abundant with agents that are not bound to professional etiquette and may be protected by anonymity. Second-hand car ad pages[1] sometimes build their own price list, based on the price reported by the site users.[2] Unfortunately, there is no way to verify this price, nor the fact that the car was actually sold, or even existed. This is a golden opportunity for traders, importers and anyone else with malicious intentions, to bias the price list by reporting false prices. This specific example is more related to regression learning, but there are also examples (e.g. online recommendation systems) that have a stronger flavor of classification.

1.2 Thesis Structure

This work is organized as follows: The next chapter supplies the required background in game theory and machine learning. Chapter 3 presents the formal definitions and notations used throughout the thesis. In chapter 4 we present a simplified scenario, as a first demonstration of both positive and negative results, as well as generalization techniques. Chapter 5 extends our results to a wide variety of concept classes, under the assumption that the agents are interested in similar data points. We also show how results in the decision making setting can be leveraged to learning algorithms with a bounded error. In chapter 6 we drop the last assumption and observe how it affects the previous results. In the final chapter, we conclude our results, compare them with related work, and suggest future directions.

Appendices A and B discuss in detail questions that are related to this work. Appendix C contains some of the longer proofs, which were omitted to allow easier and continuous reading.

[1] E.g. www.yad2.co.il

[2] The seller is asked for the real price he got for the car, when removing the ad.

Chapter 2

Background

In this chapter we provide some background that is required for this work. The background on Machine Learning is in a lower, technical level, as similar formal models and notations are later used as a basis for our Strategy-Proof Classification models. In contrast, some background concepts in Mechanism Design are described in a rather informal way, and are intended to supply the reader with better understanding of the intuition behind the formal models, examples and proof techniques. Where needed, terms are formally defined in the relevant chapters. We conclude with an overview of related work in both fields.

2.1 Game Theory and Mechanism Design

The presentation in this section is mainly based on introductory chapters from the book "Algorithmic Game Theory" by Nisan, Roughgarden, Tardos and Vazirani [36; 26]. Informally, a *game* consists of players, actions and payoffs (or utilities). The game defines the outcome of every possible joint action of players, and each player is awarded a certain amount of utility based on this outcome. A *rational player* is assumed to act in a way that will maximize his utility. In cases where the outcome is stochastic, rational players are assumed to try and maximize their *expected utility*. The rationality assumption (i.e. assuming that all players are rational) somewhat simplifies the analysis of game outcome, but even if an agent is rational it is not always clear what he will do. As the outcome of the game depends also on actions of other players, the utility maximizing action may not be well-defined. However, in some cases the behavior of all rational players can be expected, due to the properties of the specific game.

An action, or strategy, of a player is *[strictly] dominant*,[1] if the player is always [strictly]

[1] In some places, there is an additional requirement on non-strict dominant strategies, that the inequality will be strict in at least one combination of agents' actions. We do not impose such a restriction. E.g. if all actions are equivalent, then all actions are (non-strictly) dominant.

better off (i.e. gets a higher utility) by choosing that action. An immediate corollary of the rationality assumption is that every player will always play a Dominant Strategy (DS), if such is available for him (there may be more than one weakly dominant strategy). Note that if a joint action is a [strictly] DS for *all* players, then it is also a [unique] Nash Equilibrium. In Mechanism Design, we are typically interested in designing games that will motivate agents to play in a specific, desired way. A common solution concept is to construct the payoffs of a game such that the desired behavior of the players will reside in the set of DS. Even if there is more than one DS, we can assume the players will play as told, as they cannot gain by playing any other action even if it is also DS.

Strategy-Proofness

In a wide collection of games the players hold some private information (typically a description of their preferences over possible outcomes), and they are required to reveal this information, or parts of it. A designated *mechanism* then aggregates reported information from all players, and decides on the outcome of the game. As players' information is private, players can lie, i.e. mis-report their private preferences. In auctions, for example, the players are required to report how much they are willing to pay for the item. In Voting Games (see next section), the voters are asked for their order of preference over candidates. Of course in both cases nothing prevents the players from bidding a different price, or from throwing into the ballot box the name of a candidate other than their favorite. As mechanisms are typically designed to yield optimal results (e.g. to elect the most popular candidate or to maximize the profit of the auctioneer) based on their input, an untruthful input may results in suboptimal results and thus lying is considered an undesirable behavior. Therefore, there is much interest in mechanisms that enforce or encourage truth telling. As stated in the previous paragraph, under the assumption of rationality, the designer of the mechanism can promote truth telling if she makes sure that telling the truth is always a DS for all agents. Mechanisms in which truth-telling is always dominant are known as *Strategy-Proof*, *Truthful*, *Honest*, or *Incentive Compatible*.[2] SP mechanisms also have some other virtues: They relieve players of the *burden* of strategic thinking. The players benefit since they can decide on their optimal action without the computational and/or communicational overhead in considering actions of other players. Moreover, the mechanism designer can be certain that the outcome of the game is not affected by factors such as information privacy and whether players communicate. We again emphasize that in an SP mechanism no agent can gain by lying,

[2]In some places Incentive Compatibility is interpreted in a weaker sense, i.e. that truth telling is a Nash Equilibrium. Throughout this work only the first interpretation is used. To avoid confusion, we avoid the term Incentive Compatible. Classification mechanism will be referred to as Strategy-Proof or Truthful, and other mechanisms (e.g in voting) will be described by the term *honest*.

and thus we may assume that if an SP mechanism is used, all agents truthfully report their private information. An even more demanding requirement is *Group-Strategy-Proofness* (GSP), i.e. that no subset of agents is able to gain by lying.

VCG mechanisms

A generic way of turning any mechanism into a Strategy-Proof mechanism, is augmenting it with Vickrey-Clarke-Groves (VCG) payments (see, e.g., [26] for an overview of the VCG mechanism). In some games, e.g. single-item auctions, the payment mechanism is simple and easy to implement. This supplies us with an efficient and widely used SP mechanism, namely the Vickrey Auction [39]. However, in the general case, VCG requires arbitrary payments to be transfered to and from agents, and in many multiagent systems such payments are often not feasible (e.g. in many internet settings) or illegal (in most elections). Therefore, this work concentrates on achieving good mechanisms *without* payments. See Dekel et al. [11] for a discussion of this point. More criticism on VCG mechanisms is in [32].

SP mechanisms without payments

Finding SP mechanisms that do not require side payments has drawn attention in many fields with varying levels of success. In a simple voting game with two candidates, the trivial Majority Rule is SP (even GSP), and it is hard to think of reasons why to use any other rule. The results in most other fields typically expose a tradeoff between SP and other properties of the designed mechanisms. For example, SP (or GSP) mechanisms are available in Exchange Economics [3], Allotment Rules [4] and Cake Cutting [6], in all of which truthfulness comes at the expense of other desired properties of the mechanism. In the negative extremity, we meet Voting Rules again: If there are at least three candidates and no restriction on voters' preferences, then truthful mechanisms do not exist unless we are willing to relax fundamental requirements of the voting procedure.

Voting

The players in Voting problems are the voters, whose private information is their order of preferences over candidates, i.e. the ranking of candidates. The set of different rankings by all voters is known as the *profile*. A Voting Mechanism gets as input the reported profile, which is not necessarily true, and outputs the winning candidate. A Seminal result in voting theory [17; 33] shows that all possible SP voting mechanisms are extremely limited: either there are only two candidates that can be elected, or there is a single voter (a dictator) which solely decides on

the winner. Extensions and similar impossibility results have been provided for multi-winner voting [13], Social Welfare mechanisms [1], Voting mechanisms that use randomization [18], and for other related problems. A review of SP mechanisms (with and without payments) in many related domains can be found in [35].

2.2 Machine Learning

As Machine Learning is a general name for many different areas, this work only deals with *Supervised Learning*, and more specifically with *Classification*. The presentation of classification and related issues in this section is mainly based on the textbook by Devroye, Györfi and Lugosi [12], with some simplifications and adjusted notations. The PAC model was introduced by Valiant in [37].

The PAC model

A key model in the field of classification is the *Probably Approximately Correct (PAC)* model. In this model, a learning algorithm is presented with a finite amount of labeled examples. Examples are pairs $\langle x, y\rangle \in \mathcal{X} \times \mathcal{Y}$, where $x \in \mathcal{X}$ is a data point and $y \in \mathcal{Y}$ is its corresponding label. In the general case $\mathcal{Y}$ is any finite set, but in *binary* classification problems, on which we focus, $\mathcal{Y} = \{-, +\}$. The algorithm then generates a function $h : \mathcal{X} \to \mathcal{Y}$, also known as *classifier*, *hypothesis* or *concept*, which is thereafter used to predict the labels of new examples. Let $\mathcal{H}$ the set of all the possible classifiers.

The underlying assumptions of the PAC model are that:

1. The set of allowed classifiers is restricted, i.e. $h \in \mathcal{C}$ for some fixed, predefined $\mathcal{C} \subseteq \mathcal{H}$. $\mathcal{C}$ is known as the *concept class*.

2. There is some fixed but unknown distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$ from which training samples are drawn independently. The evaluation of the result classifier is also according $\mathcal{D}$, which is known as the *target distribution*.

We use the distribution $\mathcal{D}$ to obtain a reasonable evaluation of different classifiers. Bold square parentheses ($[\![a]\!]$) are used to denote the indicator random variable of an event a, i.e. $[\![a]\!] = 1$ iff a is true, and 0 otherwise.

Definition 2.2.1. The *risk* of classifier is the probability that it will give a false prediction according to the distribution $\mathcal{D}$:

$$risk(h) = \mathbb{E}_{\mathcal{D}}\left[[\![h(x) \neq y]\!]\right] = \Pr_{\langle x,y\rangle \sim \mathcal{D}}(h(x) \neq y)$$

Remark 2.2.2. More generally, the risk depends on the *loss function* being used, and defined as $\mathbb{E}_{\mathcal{D}}\left[\ell(c(x), y)\right]$, where ℓ can be any non-negative function. The common $0-1$ loss is defined as $\ell(a, b) = [\![a \neq b]\!]$. Throughout this work we only use the $0-1$ loss, which gets us definition 2.2.1, but other loss functions are also used in literature.

We define the *bayes classifier* to be the classifier with the minimum risk:

$$h_{bayes} = argmin_{h \in \mathcal{H}} risk(h),$$

and denote its risk by $r_{bayes} = risk(h_{bayes}) = min_{h \in \mathcal{H}} risk(h)$. [3]

Idealy, we would like a perfect classification algorithm, which always returns a bayes classifier, as it is evidently the best thing we can hope for. Unfortunately, this is not possible since the distribution $\mathcal{D}$ is unknown, and the algorithm is only exposed to a finite number of samples from this distribution. Moreover, we cannot even tell the actual risk of a given classifier, but only estimate it:

Definition 2.2.3. Let $S = \{\langle x_i, y_i \rangle\}_{i=1}^{m}$ a dataset of m samples sampled i.i.d. from $\mathcal{D}$. The *empirical risk* of a classifier h w.r.t. a dataset S is the relative number of errors h makes on S.

$$\widehat{risk}(h, S) = \frac{1}{m} \sum_{i \leq m} [\![h(x_i) \neq y_i]\!]$$

The next natural step will be to try and minimize the empirical error, i.e. to sample some dataset S, and then take $h = argmin_{h \in \mathcal{H}} \widehat{risk}(h, S)$ as our chosen classifier. Unfortunately, a well known phenomenon in machine learning is that finding a classifier with low empirical risk on a given sampled dataset, does not guarantee that the actual risk of this classifier is low. In the general case, we have no means to bound $|risk(h) - \widehat{risk}(h, S)|$, and it may be arbitrarily large.

However, some major breakthroughs in machine learning during the recent decades demonstrated that this problem can be effectively solved by restricting the concept class in which we search for classifiers.

Let $\mathcal{C} \subseteq \mathcal{H}$ a concept class and S a dataset. The *risk minimizer* of $\mathcal{C}$ is just the best classifier in $\mathcal{C}$, [4]

[3] For simplicity, we restrict ourselves throughout this work to cases where h_{bayes} is well-defined. Note that this is not such a hard restriction: If $\mathcal{X}$ is either finite or a closed subset of $\mathbb{R}^d$ (as is usually the case in classification) then the risk obtains a minimum in $\mathcal{H}$

[4] Unlike h_{bayes}, c_{min} may not be well-defined in some cases (e.g. if $\mathcal{C}$ is not closed). A more accurate way of presenting the equations in this section and in appendix A would define r_{min} and r_{bayes} using limits, as done in [12]. Note however, that c^*, r^* are well-defined and should not be replaced. For the sake of exposition we leave the definition as is.

$$c_{min} = argmin_{c \in \mathcal{C}} risk(c).$$

Yet unfortunately, c_{min} is also out of reach, for the same reasons we could not find h_{bayes} - our algorithms may only access the sampled dataset.

Definition 2.2.4. The *Empirical Risk Minimizer* (ERM) of $\mathcal{C}$ w.r.t. S is the classifier with the lowest empirical risk in $\mathcal{C}$:

$$c^* = argmin_{c \in \mathcal{C}} \widehat{risk}(c, S).$$

We also denote $r^* = \widehat{risk}(c^*, S)$.

Note that there is only a finite number of ways to classify the finite dataset S, thus both c^* and r^* are well defined, although c^* is not necessarily unique.

By showing that our empirical estimation is not too far from the truth, we can justify picking the ERM. Indeed, it is known that if $\mathcal{C}$ is of bounded complexity, then given enough samples[5] one may achieve:

$$pr\left(\forall c \in \mathcal{C}\left(|risk(c) - \widehat{risk}(c, S)| < \epsilon\right)\right) > 1 - \delta, \tag{2.1}$$

for arbitrarily small ϵ and δ and for every distribution $\mathcal{D}$. That is, almost every sample guarantees that the empirical and the real risk are close. Equation (2.1) is a the key to the generalization process, and will be used in the formal generalizations of decision mechanisms to machine learning settings. In the remainder of this section, we overview this process in a more structured (albait less accurate) way, with the purpose of convincing the reader that the ERM is the right concept to focus on.

To begin with, observe that with high probability our ERM is not much worse than the real risk minimizer:

$$risk(c^*) < \widehat{risk}(c^*, S) + \epsilon \leq \widehat{risk}(c_{min}, S) + \epsilon < risk(c_{min}) + 2\epsilon \tag{2.2}$$

(With high probability). This already supplies us with a generic learning mechanism: First choose a "good" concept class, then sample a dataset large enough, and output the ERM. If our basic assumptions hold, and the dataset contains enough samples, then we *Probably* (with probability $> 1 - \delta$) get a classifier which is *Approximately Correct* (with $risk < \epsilon$) - Hence the name *PAC*. Indeed, many learning algorithms such as SVM and Fisher's LDA [34; 16] are based on this approach.

[5]The actual size of the dataset, as well as the complexity of concept classes, are discussed in appendix A.

Approximating the ERM

Sometimes the ERM cannot be found due to computational or other limitations but can only be approximated to some degree. Nevertheless, we can still use equation (2.1) to prove a PAC-style bound on the actual risk.

Let $c' \in \mathcal{C}$ s.t. $\widehat{risk}(c', S) \leq \alpha r^*$, then from equation (2.1)

$$risk(c') < \widehat{risk}(c', S) + \epsilon \leq \alpha \cdot \widehat{risk}(c^*, S) + \epsilon < \alpha \cdot risk(c_{min}) + (\alpha + 1)\epsilon \tag{2.3}$$

also holds with probability of at least $1 - \delta$.

Randomized Algorithms

An immediate (and weaker) corollary of equation (2.2) gives us a bound on the expected risk, taken over all sampled datasets of size m:

$$\mathbb{E}_{S \sim \mathcal{D}^m} \left[risk(c^*(S)) \right] < risk(c_{min}) + \epsilon',$$

where $c^*(S)$ is the ERM computed on dataset S.

Theorem 2.2.5. *Let $\mathcal{C}$ of a bounded complexity. Suppose that we have some (possibly randomized) algorithm which returns a concept $c_{\mathcal{A}}$, guaranteeing*

$$\forall S \in (\mathcal{X} \times \mathcal{Y})^m \left(\mathbb{E}_{\mathcal{A}} \left[\widehat{risk}(c_{\mathcal{A}}(S), S) \right] \leq \alpha \cdot r^* \right).$$

If S is sampled i.i.d. from $\mathcal{D}$ and $m = |S|$ is high enough then

$$\mathbb{E} \left[risk(c) \right] = \mathbb{E}_S \left[\mathbb{E}_{\mathcal{A}} \left[risk(c_{\mathcal{A}}(S)) \right] \right] < \alpha \cdot risk(c_{min}) + \epsilon.$$

Notably, $m > \Omega \left(\frac{\alpha^2}{\epsilon^2} log \left(\frac{\alpha}{\epsilon} \right) \right)$ examples are sufficient.

The proof is given in appendix A.

It is important to emphasize that unlike equation (2.3), there is a non-negligible probability that the approximation will be far worse than α, thus this is not really a PAC-style bound. However, bounding the expected risk still explains the motivation behind randomized approximation algorithms for the ERM, and this is indeed the approach taken in this work. Unless explicitly stated otherwise, we take the ERM as our target classifier, to which we compare the output of our mechanisms.

Chapter 3

Model and Notations

3.1 Basic Notations

Our model inherits most of its notations from the classification model we presented in section 2.2. Thus the definitions of $\mathcal{X}, \mathcal{Y}, \mathcal{H}, \mathcal{C}$ remain the same. As we reduced our problem to finding classifiers that are empirically good, we are relieved of the need to consider distributions and sampling procedures, and our input will simply be a dataset. However, this dataset is actually a collection of smaller datasets, each of them "belongs" to a different agent. We define the set of agents $I = \{1, \ldots, n\}$, where $n \geq 2$. Each agent i *controls* the data points $\{x_{i,1}, \ldots, x_{i,m_i}\}$, where $x_{i,j} \in \mathcal{X}$. The real label of $x_{i,j}$, denoted $y_{i,j} \in \{-, +\}$, is a private data known only to agent i, which reports in turn the *reported label* $\overline{y}_{i,j}$ for each data point. The pairs $s_{i,j} = \langle x_{i,j}, y_{i,j} \rangle$; $\overline{s}_{i,j} = \langle x_{i,j}, \overline{y}_{i,j} \rangle$ are a *real example* and its related *reported example*. The partial dataset S_i contains all of agent i's real examples, and the full dataset is the multiset containing all partial datasets $S = \bigcup_{i \in I} S_i$. As before, $m = |S| = \sum_{i=1}^{n} m_i$. We also assume that S preserves the distinction between the datasets, i.e. the owner of each example is known. In some places S is denoted as the *set* or *tuple* of partial datasets (i.e. $\{S_1, \ldots, S_n\}$ or $\langle S_1, \ldots, S_n \rangle$). Similarly, $\overline{S}_i, \overline{S}$ contain the reported partial and full datasets.

We denote by $\mathcal{S}_m$ all possible datasets of size m, and $\mathcal{S} = \bigcup_{m=1}^{\infty} \mathcal{S}_m$. It will sometimes be convenient to refer only to the unlabeled examples of agent i, or only to his labels. We denote these sets by X_i and Y_i, respectively, and denote each partial dataset also as $S_i = \langle X_i, Y_i \rangle$. Likewise, $\overline{S}_i = \langle X_i, \overline{Y}_i \rangle$. We also refer to X_i as agent i's *interest*, and to Y_i [$\overline{Y}_i$] as his *[reported] preferences*.

Evaluation of Classifiers

Let S a dataset. The *risk* of a classifier $c \in \mathcal{C}$ to each agent i, is exactly the relative number of errors c makes on i's examples:

$$risk_i(c, S) = \frac{1}{m_i} \sum_{j=1}^{m_i} [\![c(x_{i,j}) \neq y_{i,j}]\!]. \tag{3.1}$$

As the private risk depends only on the examples of the relevant agent, therefore we use $risk_i(c, S)$ and $risk_i(c, S_i)$ interchangeably. The risk can be thought as the (negative) utility that each agent receives when c is the selected classifier.

The differentiation we made in section 2.2 between risk and empirical risk is not needed throughout most sections of this work, as we only try to minimize the error on the dataset. This is true for both settings described in the introduction, i.e. whether S reflects the full interests of the agents or it is sampled i.i.d from some unknown distribution.

From the decision maker's point-of-view, the total number of errors is the right metric to evaluate a classifier. We define the *global risk* of a classifier as

$$risk(c, S) = \sum_{i=1}^{n} \frac{m_i}{m} \cdot risk_i(c, S) = \frac{1}{m} \sum_{\langle x,y \rangle \in S} [\![c(x) \neq y]\!]. \tag{3.2}$$

We sometimes write the global risk as $risk_I$ for emphasis.

As in the basic machine learning model, we denote the (empirically) best classifier by $c^*(S) = argmin_{c \in \mathcal{C}} risk(c, S)$, and the minimal global risk by $r^*(S) = risk(c^*, S)$. When the dataset is clear from the context, we may simply use c^* and r^*.

Decision Problems

A *Decision Problem* is characterized by an input space $\mathcal{X}$, a set of possible labels $\mathcal{Y}$ and a concept class $\mathcal{C}$. $\mathcal{Y} = \{-, +\}$ unless explicitly stated otherwise. Each decision problem that we discuss is assigned a name in block letters (e.g. DET-SYNTHETIC) to allow easy reference. In each such problem we assume that a truthful learning mechanism gets as input a dataset $S \in \mathcal{S}$, and must output a classifier $c_{\mathcal{M}} \in \mathcal{C}$ with a low as possible global risk. We explore the lower and upper bounds on the approximation ratio of such learning mechanisms in each problem, by comparing the result of the mechanism to r^*. The description of a decision problem may also contain further restrictions or assumptions (e.g. on the structure of the dataset). We emphasize that in the decision problems, unlike their generalization versions (or *learning problems*), the dataset is not sampled. Rather, it is simply given as input.

We make a distinction between two types of classification mechanisms, according to whether their output is deterministic or not.

3.2 Deterministic Mechanisms

A *Deterministic Classification Mechanism* is a function $\mathcal{M} : \mathcal{S} \to \mathcal{C}$. I.e. it maps every possible dataset S to a single classifier $c_{\mathcal{M}} = \mathcal{M}(S)$.

An example of such a mechanism is the ***ERM mechanism***, which simply outputs $c^*(\overline{S})$, i.e. the classifier with the lowest risk in $\mathcal{C}$ (if there is more than one, pick one arbitrarily).

Another example of such mechanism is taking agent 1 as a dictator, i.e. always output $c^*(\overline{S}_1)$, the best classifier for agent 1 (if there is more than one, pick one arbitrarily).

Recall however our fundamental assumption: that $\mathcal{M}$ cannot be directly presented with the real dataset. Rather, it can only be given as input the reported dataset $\overline{S}$.

Definition 3.2.1. A mechanism $\mathcal{M}$ is **Strategy-Proof** (SP) if no agent can lower his risk by misreporting his true labels. More formally, let $S = \bigcup_{i \le n} S_i$ a dataset, $j \in I$, and let $\overline{S} = \bigcup_{i \neq j} S_i \cup \overline{S}_j$. That is, the reported labels of all examples are truthful, except for some of the labels of agent j's examples. $\mathcal{M}$ is SP if for every such S, j and $\overline{S}$,

$$risk_j(\mathcal{M}(S), S) \le risk_j(\mathcal{M}(\overline{S}), S).$$

A stronger demand is that agents would not gain even if some of them coordinate their lies.

Definition 3.2.2. A mechanism $\mathcal{M}$ is **Group Strategy-Proof** (GSP) if no subset of agents can misreport their true labels, such that all of them strictly gain. More formally, let $S = \bigcup_{i \le n} S_i$ a dataset, and let $J \subseteq I$, $\overline{S} = \bigcup_{i \notin J} S_i \cup \bigcup_{i \in J} \overline{S}_j$, where $\overline{S}_j = \langle X_j, \overline{Y}_j \rangle$. That is $S, \overline{S}$ are equal, except for some of the labels of agents in J. $\mathcal{M}$ is GSP if for every such $S, \overline{S}$,

$$\exists j \in J \left(risk_j(\mathcal{M}(S), S) \le risk_j(\mathcal{M}(\overline{S}), S) \right).$$

There are also alternative, non-equivalent definitions of Group Strategy-Proofness in literature (see remark 4.1.3).

When dealing with SP mechanisms, the assumption is that all agents are indeed truthful and therefore $\overline{S} = S$. We thus simply write $risk(\mathcal{M}, S)$ instead of $risk(\mathcal{M}(S), S)$ or $risk(\mathcal{M}(\overline{S}), S)$.

Definition 3.2.3. A mechanism $\mathcal{M}$ is **β-approximating** if for any dataset S it holds that

$$risk(\mathcal{M}(S), S) \le \beta \cdot r^*.$$

Claim 3.2.4. *The ERM mechanism is 1-approximating but not SP. On the other hand, the dictator mechanism is SP but has an unbounded approximation ratio.*

Proof. The ERM mechanism is clearly 1-approximating, since by definition $r^* = risk(c^*, S)$. An example for a very simple scenario in which ERM is not SP appears in section 4.

As for the dictator mechanism, agent 1 has no interest to lie as it may only lead to an outcome that does not minimize his risk, and is thus worse for him. Other agents cannot gain by lying simply because their reported labels are ignored anyway, therefore the mechanism is SP. To see why the approximation ratio is unbounded, just add more agents (or more examples to agents) that do not agree with agent 1, but agree with each other. □

We are interested in mechanisms that are both SP, and have a bounded (preferably low) approximation ratio. As we will see, this is sometimes a difficult requirement. However, we are willing to settle for mechanisms that perform well in expectation.

3.3 Randomized Mechanisms

A *Randomized Classification Mechanism* allows randomization of the output classifier, where the actual probabilities are a function of the input dataset, i.e. $\mathcal{M} : \mathcal{S} \to \Delta(\mathcal{C})$. We denote by $p_{\mathcal{M}}$ the conditional distribution induced by $\mathcal{M}$, that is $\mathcal{M}$ returns classifier c according to the conditional probability distribution $p_{\mathcal{M}}(c|S)$. Note that $c_{\mathcal{M}}$ (and any function of it, such as the risk) is a random variable.

Definition 3.3.1. The global risk of the result is defined to be the *expected risk* of $c_{\mathcal{M}}$ over possible outcomes of the mechanism:

$$risk(\mathcal{M}, S) = \mathbb{E}\left[risk(c_{\mathcal{M}}(S), S)\right] = \mathbb{E}_{c\sim p_{\mathcal{M}}}\left[risk(c, S)|S\right].$$

If the mechanism selects its result from a discrete set $C \subseteq \mathcal{C}$, as is usually the case, then can write the expected risk explicitly:

$$risk(\mathcal{M}, S) = \sum_{c\in C} p_{\mathcal{M}}(c|S) \cdot risk(c, S).$$

Similarly, the private risk of the mechanism for agent i is $\mathbb{E}\left[risk_i(c_{\mathcal{M}}(S), S)\right]$. Since both private and global risk are real non-negative numbers, we can extend definitions 3.2.1 and 3.2.3 to randomized mechanisms in a straightforward way. Intuitively, a randomized mechanism is SP if no agent can lower his *expected risk* by lying. Similarly, a randomized mechanism is β-approximating if the *expected risk* of the outcome does not exceed βr^*.

We emphasize that the mechanism must guarantee expected β-approximation on *every input dataset*, and expectation is taken only over possible outcomes of the mechanism for a given dataset.

Conditional Risk

In some places we define *events* on some probability space p. If Q is an event, then $risk(\mathcal{M}|Q)$ is a shorthand for $\mathbb{E}_p\left[risk(\mathcal{M})|Q\right]$. This holds for both private and global risk. Note that for any set of complementary mutually exclusive events $Q_1, \ldots, Q_q$, $risk(\mathcal{M}) = \sum_{i \leq q} \Pr_p(Q_i) risk(\mathcal{M}|Q_i)$.

3.4 Generalization Schemes

Most of this work is focused on the decision-making setting, in which the dataset S is a given input. However, in sections 4.3 and 5.2 we generalize our results to the learning-theoretic setting where datasets are *sampled* from unknown distributions. As to avoid overloading the reader with notations, the relevant definitions and assumptions are detailed in the opening of each section. One issue that is relevant to all these sections, is the distinction between the *actual risk* and the *empirical risk*. The empirical risk, which is computed w.r.t. a given dataset, is denoted by $\widehat{risk}(c, S)$, whereas the actual risk, which is computed w.r.t. a fixed distribution, is denoted by $risk(c)$. Further, our goal to which we compare our classifiers is the minimal risk r_{min} rather that r^* (which depends on S).

In all other sections this distinction does not exist. The risk is denoted by $risk(c, S)$, as in equations (3.1), (3.2), and we compare it with $r^* = r^*(S)$.

Chapter 4

A Simple Scenario

We start by analyzing a very simple scenario, in which the concept class contains only 2 classifiers: Either all examples are labeled by the mechanism as positive, or all of them are labeled as negative. As the results for this scenario are interesting in their own right, it also serves us as an "easy start" for the analysis of SP mechanisms, before we dive into more complex models. The contents of this chapter were published in [23], with the omission of some proofs.

Definition of the PLUS-MINUS Problem

We define the problem PLUS-MINUS by setting the concept class to be $\mathcal{C} = \{c_-, c_+\}$. c_+ can be interpreted as a *positive decision*, assigning a positive label to any $x \in \mathcal{X}$, whereas c_- is the opposite, negative decision. Note that an equivalent way to define PLUS-MINUS is by restricting the input space instead of the concept class. That is, assume that all examples are located in a single spot $\mathcal{X} = \{x\}$, which is necessarily classified as either positive or negative by every classifier.

As in the general case, agents report to the mechanism the labels of the points they control. If all agents report truthfully, the above problem is trivially solved by choosing c according to the majority of labels. This is a special case of the Empirical Risk Minimization (ERM) mechanism, which was described in section 3.2.

ERM is not SP

Unfortunately, if we choose ERM as our mechanism then agents may lie in order to decrease their subjective risk - even in the extremely restricted case of two possible concepts. Indeed, consider the following dataset (illustrated in Figure 4.1): agent 1 controls 3 examples, 2 positive and 1 negative. Agent 2 controls 2 examples, both negative. Since there is a majority of negative

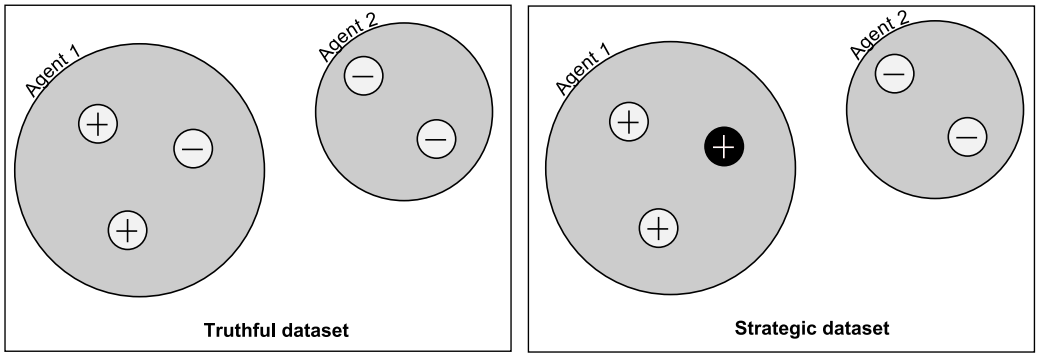


Figure 4.1: ERM is not strategy-proof. Agent 1 changes one of its points from negative to positive, thus changing the risk minimizer from c_- to c_+, to agent 1's advantage. In this illustration, $\mathcal{X} = \mathbb{R}^2$.

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 an agent's utility is measured with respect to its real labels, rather than with respect to the reported labels. Therefore the ERM mechanism is not SP. However, see appendix B for a more detailed analysis of the ERM approach.

Remark 4.0.1. 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 (which reduces to the simple majority rule in this case) returns c_-, falsely reporting a negative label will only reinforce the mechanism's decision. This is a striking contrast to the regression learning setting considered in Dekel et al. [11], where some deep technical results concern the single-point-per-agent scenario.

Despite the fact that ERM is not SP, the concept that minimizes the global risk is clearly optimal. Thus we would like to use it to evaluate other concepts and mechanisms. Applying definition 2.2.4 in our restricted setting, the optimal risk is

$$r^* = risk(c^*, S) = \min\{risk(c_+, S), risk(c_-, S)\}.$$

As is common in computer science, we will be satisfied with only approximate optimality (if this guarantees strategy-proofness).

4.1 Deterministic Mechanisms and Bounds

We start with some observations. Note that the identity of each sampled point is not important, only the *number* of positive and negative points each agent controls. Thus we denote by $P_i = |\{\langle x, y\rangle \in S_i : y = +\}|, N_i = m_i - P_i = |\{\langle x, y\rangle \in S_i : y = -\}|$. For convenience we also let $P = \sum_{i\in I} P_i$, $N = \sum_{i\in I} N_i$. We emphasize that $\{P_i, N_i\}_{i\in I}$ contain all the information relevant for our problem and can thus replace S.

Now, 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 following mechanism:

Mechanism 1 THE DOUBLE MAJORITY MECHANISM ($\mathcal{M}_{DM}$)

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 4.1.1. The DM mechanism can be thought of as the classification counterpart of the Project-and-Fit regression mechanism of Dekel et al. [11]. However, the results regarding the mechanism's guarantees do not follow from their results.

Theorem 4.1.2. *The DM mechanism is a 3-approximation SP mechanism for PLUS-MINUS .*

Proof. Strategy-Proofness is obvious, since if a positive agent lies, his lie is either ineffective or results in a negative (i.e. worse) result, and likewise for a negative agent.

Remark 4.1.3. We correct here a mistake in our AAAI paper [23]. In that paper we stated that the DM mechanism, as well as some other mechanisms in this chapter, are *Group Strategy-Proof*.

Under definition 3.2.2 for GSP this is correct. However, In the paper we used an alternative definition of group strategy-proofness, i.e. that a joint lie is considered beneficial even if only some manipulators gain from it (as long as the others are not harmed). Under this definition, the mechanism is no longer GSP.

A counter example is when there is an agent with an equal number of positive and negative examples, which might collude with a negative agent.

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.

Lemma 4.1.4. $P \geq \frac{1}{4}m$.

Proof. 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 so:

$$P \geq \frac{P'}{2} \geq \frac{m}{4}$$

□

Now, we know that $P + N = m$, so:

$$N = m - P \leq m - (\frac{m}{4}) = \frac{3m}{4} \leq 3P$$

Clearly if the mechanism decided "correctly", i.e., $P \geq m/2$, then

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

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

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

In any case we have that $risk(c, S) \leq 3r^*$, proving that the DM mechanism 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 the SP property. We show the this is *not* the case by proving a matching lower bound on the best possible approximation ratio achievable by an SP mechanism.

Theorem 4.1.5. *Let $\epsilon > 0$. There is no deterministic $(3 - \epsilon)$-approximation SP mechanism for PLUS-MINUS .*

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.

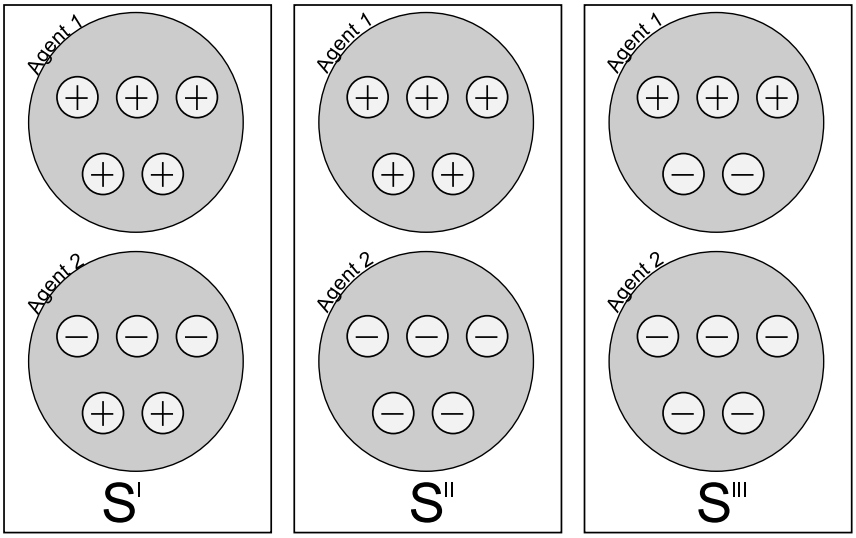


Figure 4.2: The examples of each agent in the three datasets are shown (for $k = 2$). Agent 1 can make dataset II look like dataset III and vice versa by reporting false labels. The same goes for agent 2 regarding datasets I and II.

Let $\epsilon > 0$. We will use $I = \{1, 2\}$, and an integer $k = k(\epsilon)$ to be defined later. Note that in all 3 datasets $m_1 = m_2 = 2k + 1$. We define the three datasets as follows (see Figure 4.2 for an illustration):

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

Let $\mathcal{M}$ be some SP mechanism. Then it must hold that $\mathcal{M}(S^I) = \mathcal{M}(S^{II})$. Indeed, otherwise assume first that $\mathcal{M}(S^I) = c_+$ and $\mathcal{M}(S^{II}) = c_-$. 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 $\mathcal{M}(S^I) = c_-$ and $\mathcal{M}(S^{II}) = c_+$, 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 $\mathcal{M}(S^{II}) = \mathcal{M}(S^{III})$.

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

$$risk(c, S^{III}) = risk(c_+, S^{III}) = \frac{N_1 + N_2}{m} = \frac{3k + 1}{4k + 2} \tag{4.1}$$

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

$$r^* = risk(c_-, S^{III}) = \frac{k+1}{4k+2}$$

By combining the last two equations:

$$\frac{risk(c, S^{III})}{r^*} = \frac{\frac{3k+1}{4k+2}}{\frac{k+1}{4k+2}} = \frac{3k+1}{k+1}$$

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

4.2 Randomized Mechanisms and Bounds

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?

Recall that we formally defined the risk of randomized mechanism as the *expected* risk of the result classifier. Applying definition 3.3.1 to our simple concept class $C = \{c_+, c_-\}$, yields the following simple definition for the risk:

$$risk(\mathcal{M}(S), S) = p_{\mathcal{M}}(c_+|S)risk(c_+, S) + p_{\mathcal{M}}(c_-|S)risk(c_-, S)$$

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

Theorem 4.2.1. *Let $\epsilon > 0$. There is no $(2-\epsilon)$-approximation SP randomized mechanism for PLUS-MINUS .*

Proof. We will use the same datasets used in the proof of Theorem 4.1.5, and illustrated in Figure 4.2. Let $\mathcal{M}$ be a SP randomized mechanism, and denote by $p_{\mathcal{M}}(c|S)$ 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 4.2.2. $p_{\mathcal{M}}(c_+|S^I) = p_{\mathcal{M}}(c_+|S^{II}) = p_{\mathcal{M}}(c_+|S^{III})$.

Proof. If $p_{\mathcal{M}}(c_+|S^I) \neq p_{\mathcal{M}}(c_+|S^{II})$ then agent 2 will report its labels in a way that guarantees a higher probability of c_-. Similarly, $p_{\mathcal{M}}(c_+|S^{II}) \neq p_{\mathcal{M}}(c_+|S^{III})$ implies that agent 1 can increase the probability of c_+ by lying. □

Denote

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

and

$$p_- = p_{\mathcal{M}}(c_-|S^I) = p_{\mathcal{M}}(c_-|S^{II}) = p_{\mathcal{M}}(c_-|S^{III}).$$

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

$$\begin{aligned} risk(\mathcal{M}(S^{III}), S^{III}) &= p_+ risk(c_+, S^{III}) + p_- risk(c_-, S^{III}) \\ &= p_+ \cdot \frac{3k+1}{4k+2} + p_- \cdot \frac{k+1}{4k+2} \\ &\geq \frac{1}{2} \cdot \frac{3k+1}{4k+2} + \frac{1}{2} \cdot \frac{k+1}{4k+2} = \frac{1}{2}, \end{aligned}$$

whereas

$$r^* = risk(c_-, S^{III}) = \frac{k+1}{4k+2}$$

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

$$\frac{risk(\mathcal{M}(S^{III}), S^{III})}{r^*} = \frac{4k+2}{2(k+1)} = 2 - \frac{1}{k+1} > 2 - \epsilon$$

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

Exploring the upper approximation bound in the randomized case appears to be a less trivial task than in the deterministic case. Let us start by putting forward the *Random Dictator mechanism*. Note that in the PLUS-MINUS problem it is equivalent to return c_+ with proba-

Mechanism 2 THE RANDOM DICTATOR MECHANISM ($\mathcal{M}_R$)

Select agent i w.p. m_i/m.
return c_i.

bility P'/m and c_- with probability N'/m. Unfortunately, this simple randomization (which is clearly SP for *any* concept class) cannot even beat the deterministic bound of $3-\epsilon$. As an example, consider the dataset S of n agents with the following examples: one agent with $P_1 = k+1$, $N_1 = k$, and $n-1$ additional agents each holding $2k+1$ negative examples. Thus $P = k+1; N = (n-1)(2k+1)$ but $P' = 2k+1; N' = (n-1)(2k+1)$. The optimal classifier makes $|P| = k+1$ mistakes, whereas the expected number of mistakes done by the mechanism

is

$$\begin{aligned} m \cdot risk(\mathcal{M}_R, S) =& p_- \cdot |P| + p_+ \cdot |N| = \frac{N'}{m} \cdot (k+1) + \frac{P'}{m} \cdot ((n-1)(2k+1)+k) \\ =& \frac{(n-1)(2k+1)}{n(2k+1)}(k+1) + \frac{2k+1}{n(2k+1)}(2nk+n-k-1) \\ =& \frac{(n-1)(k+1)}{n} + \frac{2nk+n-k-1}{n} = \frac{nk+n-k-1+2nk+n-k-1}{n} \\ =& \frac{3nk+2n-2k-2}{n} \end{aligned}$$

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

$$\frac{risk(\mathcal{M}_R, S)}{r^*} = \frac{3nk+2n-2k-2}{n(k+1)} \overset{k\to\infty}{\to} 3 - \frac{2}{n}. \tag{4.2}$$

Thus, for every $\epsilon > 0$, there is a large enough k s.t. the approximation ratio is worse than $3 - \frac{2}{n} - \epsilon$. Note that in this example all agents control datasets of the same size $(2k+1)$. A similar example can be crafted where there are only 2 agents and the second holds $k(2k+1)$ samples, providing us with a lower bound of $3 - \epsilon$.

We presently put forward a randomized SP 2-approximation mechanism, thereby matching the lower bound with an upper bound. We will calculate P' and N' as in our deterministic DM mechanism.

Crucially, a more sophisticated (and less intuitive) randomization can do the trick.

Mechanism 3 ($\mathcal{M}_{R2}$)

Compute P' and N' as in mechanism 1.
if $P' \geq N'$ **then**
  Set $t = \frac{N'}{m}$, **return** c_+ with probability $\frac{2-3t}{2-2t}$, and c_- with probability $\frac{t}{2-2t}$.
else
  Set $t = \frac{P'}{m}$, **return** c_- with probability $\frac{2-3t}{2-2t}$, and c_+ with probability $\frac{t}{2-2t}$.
end if

Theorem 4.2.3. *Mechanism 3 is an SP 2-approximation randomized mechanism for PLUS-MINUS .*

Proof. Similarly to the DM mechanism, mechanism 3 is clearly 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 $P' \geq N'$, so $t = \frac{N'}{m}$, and c_+ is returned with probability $\frac{2-3t}{2-2t}$, c_- with probability $\frac{t}{2-2t}$. Recall that N, P denote the number of negative and positive examples, respectively. First notice that

$$N' \leq 2N. \tag{4.3}$$

This is trivially true since every negative agent has a majority of negative examples.

Case 1: $P \geq N$. From (4.3), the real risk of the best classifier satisfies:

$$r^* = risk(c_+, S) = \frac{N}{m} \geq \frac{N'}{2m} = \frac{t}{2}, \tag{4.4}$$

whereas mechanism 3 satisfies:

$$\begin{aligned} risk(\mathcal{M}_{R2}(S), S) &= \frac{2-3t}{2-2t} r^* + \frac{t}{2-2t}(1-r^*) \\ &= \frac{2-3t}{2-2t} r^* + \frac{tr^*}{2-2t}(\frac{1}{r^*} - 1) \\ &= \frac{r^*}{2-2t}(2 - 3t + t(\frac{1}{r^*} - 1)) \\ &= \frac{r^*}{2-2t}(2 - 3t + \frac{t}{r^*} - t) = \frac{r^*}{2-2t}(2 + \frac{t}{r^*} - 4t) \\ &\leq \frac{r^*}{2-2t}(2 + 2 - 4t) = r^* \frac{4-4t}{2-2t} = 2r^*, \end{aligned}$$

where the inequality follows from (4.4).

Case 2: $N > P$. In this case, the optimal risk is $r^* = risk(c_-, S) = \frac{P}{m}$.

Lemma 4.2.4. $\frac{1}{r^*} - 1 \leq \frac{1+t}{1-t}$

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, we have that $N \leq N' + \frac{P'}{2}$, and thus:

$$\frac{N}{m} \leq \frac{N'}{m} + \frac{P'}{2m} = \frac{N'}{m} + \frac{m - N'}{2m} = \frac{N'}{2m} + \frac{1}{2}.$$

It follows that $1 - r^* \leq \frac{t}{2} + \frac{1}{2}$; therefore $r^* \geq \frac{1-t}{2}$. We now conclude that

$$\frac{1}{r^*} - 1 \leq \frac{2}{1-t} - 1 = \frac{1+t}{1-t}$$

□

Now, we have:

$$
\begin{aligned}
risk(\mathcal{M}(S), S) &= \frac{t}{2-2t}r^* + (\frac{2-3t}{2-2t})(1-r^*) \\
&= \left(\frac{t}{2-2t}r^* + (\frac{2-2t-t}{2-2t})(\frac{1}{r^*}-1)\right) r^* \\
&= \left(\frac{t}{2-2t} + (1-\frac{t}{2-2t})(\frac{1}{r^*}-1)\right) r^* \\
&\leq \left(\frac{t}{2-2t} + (1-\frac{t}{2-2t})\frac{1+t}{1-t}\right) r^* \\
&= \left(\frac{t}{2-2t} + \frac{1+t}{1-t} - \frac{t+t^2}{2(1-t)(1-t)}\right) r^* \\
&= \left(\frac{t(1-t)}{2(1-t)^2} + \frac{2(1-t)(1+t)}{2(1-t)^2} - \frac{t+t^2}{2(1-t)^2}\right) r^* \\
&= \frac{t(1-t)+2(1-t)(1+t)-(t+t^2)}{2(1-t)^2} r^* \\
&= \frac{t-t^2+2-2t^2-t-t^2}{2(1-t)^2} r^* \\
&= \frac{1-2t^2}{(1-t)^2} r^* = f(t)r^*,
\end{aligned}
$$

where the inequality is due to Lemma 4.2.4.

It is now sufficient to show that $f(t) \leq 2$. By taking the derivative of $f(t)$ we find that

$$f'(t) = \frac{2-4t}{(1-t)^3}.$$

Note that both numerator and denominator are nonnegative in the range $t \in [0, 1/2]$, thus $f'(t)$ is nonnegative and $f(t)$ is monotonically nondecreasing:

$$\forall t \in [0, 1/2] \left(f(t) \leq f\left(\frac{1}{2}\right) = 2 \right).$$

As in the deterministic proof, we have that in any case $risk(\mathcal{M}_{R2}(S), S) \leq 2r^*$, thus 2-approximation is assured. □

4.3 A Learning Theoretic Setting

In this section we extend our simple decision-making model to the general machine learning framework described in section 2.2. Our previous results will be leveraged to obtain encour-

aging learning theoretic results. More specifically, we show that a good approximation of the *minimal empirical risk* (r^*), provides us with an approximation of the *minimal actual risk* r_{min}.

Instead of looking at a fixed set of examples and selecting the concept that fits them best, 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. In addition, every agent holds a (known) distribution $\mathcal{D}_i$ over the input space, which reflects the relative importance it attributes to each point. The new definition of the subjective risk naturally extends the previous setting by expressing the errors a concept makes when compared to Y_i, given the distribution $\mathcal{D}_i$:

$$risk_i(c) = \mathbb{E}_{x \sim \mathcal{D}_i} \left[[\![c(x), Y_i(x)]\!] \right]$$

The global risk is calculated similarly to the way it was before. For ease of exposition, we will assume in this section that all agents have equal weight.[1] $(n = |I|)$

$$risk(c) = \sum_{i \in I} \frac{1}{n} \cdot risk_i(c)$$

Three Game-Theoretic Assumptions

We exploit this very simple setting, to clarify the distinction between three alternative game-theoretic assumptions on agents' behavior.

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 5.2 we refer to this assumption as the "Truthful Approach". This approach is also known as "ϵ-truthfullness", and is taken by Dekel et al. [11].

A second assumption is that agents will *always* report in a way that minimizes their expected risk. Note that for some mechanisms it might not be well-defined, since the best action of an agent depends on the (unknown) actions of other agents. This assumption is also stronger than the standard rationality assumption, (which does not assume anything on agents' behavior when truth-telling is suboptimal), but it is not comparable to the first assumption. In section 5.2 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 in the truthfull approach.

An agent that always obeys the first assumption is called *ϵ-truthful*. An agent that always obeys the second assumption is called *purely rational*.

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

The third assumption, which is also the weakest, lets any agent act in one of the two ways described. I.e. each agent can decide to report either true labels, or lie in a way that will decrease his expected risk (but will never lie in a way that will not help him).

In this section we employ the third, weakest assumption, as it supplies us with the strongest results. Thus our results are "stronger" in a way than the results in Dekel et al. [11] (regression) and in chapter 5 (classification).

Remark 4.3.1. Suppose we employ the second assumption, and consider the following simple mechanism: sample one point from $\mathcal{D}$, 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 the DM mechanism or mechanism 3. This clearly supplies us with upper approximation bounds of 3 and 2 respectively, using only one sampled data point. This suggests that the difference between the assumptions is non-trivial. Compare also with the analysis of the two first approaches in section 5.2.

The Learning Mechanism

Since no mechanism can 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. The empirical risk is the risk on the sampled dataset, as defined in the previous section.

Mechanism 4 ($\tilde{\mathcal{M}}$)

---

for each agent $i \in I$ **do**
  Sample m points i.i.d. from $\mathcal{D}_i$.
  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 $\overline{S}_i = \{\langle x_{i,j}, y_{i,j} \rangle\}_{j=1}^{m}$.
end for
Use mechanism 3 on $\overline{S} = \{\overline{S}_1, \ldots, \overline{S}_n\}$, **return** $\mathcal{M}_{R2}(\overline{S})$.

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 3's randomization) mechanism 4 yields close to a 2-approximation with relatively few examples, even in the face of strategic behavior. The subtle point here is that mechanism 4 is not strategy-proof. Indeed, even if an agent gives greater weight to negative points (according to Y_i and $\mathcal{D}_i$), it might be the case that (by miserable chance) the agent's sampled dataset only contains positive points.

However, since mechanism 3 is SP in the previous section's setting, if an agent's sampled dataset faithfully represents its true distribution, and the agent is strongly inclined toward c_+ or c_-, the agent still cannot benefit by lying. 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.

Theorem 4.3.2. *Given sampled datasets, assume that agents are either ϵ-truthful or purely rational. Let $risk(\tilde{\mathcal{M}})$ denote the expected risk of mechanism 4, where the expectation is taken over the randomness of the sampling and mechanism 3. 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*

$$risk(\tilde{\mathcal{M}}) \leq 2r^* + \epsilon.$$

Specifically, sampling $m > 50\frac{1}{\epsilon^2}\ln(\frac{10n}{\epsilon})$ will suffice. The proof of the theorem is given in appendix C.1.

Chapter 5

Agents with Similar Interests

In this chapter we explore the following problem: Given that all agents control the same data points, or data points that are distributed in a similar way, what is the best approximation ratio that can be achieved with SP mechanisms? The simple scenario PLUS-MINUS is a special case of this problem, since PLUS-MINUS is equivalent to a scenario where all data points of all agents are located in a single spot, i.e. $|\mathcal{X}| = 1$. In such case, every classifier either classifies all of $\mathcal{X}$ as positive, or as negative. Thus every classifier is equivalent to either c_+ or c_-, which outs us in the PLUS-MINUS scenario.

A Formal Definition

In the problem SIMILAR$_\mathcal{C}$ the concept class is $\mathcal{C}$, but all agents control the same data points. I.e. the input is restricted to datasets of the sort $S = \{\langle X, Y_i\rangle\}_{i\leq n}$, where X is fixed over all datasets, and labels vary between agents. We refer to such a dataset as *dataset with similar interests*, an example of such dataset is in figure 5.1.

In addition, since the number of examples is the same for all agents, every agent is assigned a *weight* w_i.

For every $x \in X$, $i \in I$, let $y_i(x)$ the label that agent i holds for x. The global risk is thus defined as

$$risk_I(c, S) = \sum_{i\in I} w_i \cdot risk_i(c, S) = \sum_{i\in I} w_i \frac{1}{m'} \sum_{\langle x,y\rangle\in S_i} [\![c(x) \neq y]\!] = \frac{1}{m'} \sum_{x\in X} \sum_{i\in I} w_i [\![c(x) \neq y_i(x)]\!],$$

where $m' = |X|$ is the size of each partial dataset, and $m = |S| = n \cdot m'$.

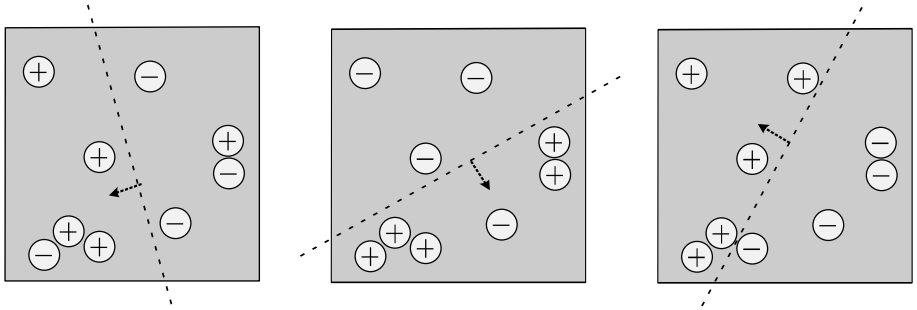


Figure 5.1: An instance of SIMILAR$_{\mathcal{C}}$ ($\mathcal{C}$ here is the class of linear separators in $\mathbb{R}^2$). We can see that the data points of all 3 agents are the same, but the labels, i.e. their preferences, are different. The best classifier of each agent is also shown (the arrow marks the positive half space of the separator).

5.1 Upper Bounds

In this section we show that there is a SP mechanism that guarantees a 3-approximation under the assumption of similar interests.

We first prove a similar result for a more general model, and later derive the upper bound for SIMILAR$_{\mathcal{C}}$ as a special case. 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 $\mathcal{X} \times \mathcal{Y}$. 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.

We adjust the definition of the private and global risk to handle distributions. We denote it by $risk(c, F)$ to distinguish it from the other definitions of risk we use.

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

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

As usual, the global risk is defined as

$$risk_I(h, F) = \sum_{i \in I} w_i risk_i(h, F).$$

This definition has much resemblance to definition 2.2.1 from section 2.2. This resemblance in not accidental, and will later be used to generalize our results.

As in chapter 4, we denote by c_i the ERM classifier (limited to the concept class $\mathcal{C}$), w.r.t. the preferences of player i, i.e. $c_i = argmin_{c \in \mathcal{C}}(risk_i(c, F))$. [1]

[1] We will assume for now that the concept class $\mathcal{C} \subseteq \mathcal{H}$ is a closed set, and thus the risk gets a minimum in it.

Also, as usual,

$$c^* = argmin_{c \in \mathcal{C}} risk_I(c, F).$$

We also denote by $r^* = risk(c^*, F)$ the minimal global risk.

The next three theorems bound the global risk of a selected classifier.

Theorem 5.1.1. *Let j be $argmax_{i \in I} w_i$, i.e. the heaviest agent, then*

$$risk_I(c_j, F) \leq (2n - 1) \cdot r^*.$$

As in chapter 4, allowing randomization (i.e. select the dictator randomly according to agents' weights) can take us further:

Theorem 5.1.2.

$$\sum_{i \in I} w_i risk_I(c_i, F) \leq 3r^*.$$

This bound can be further improved, if we attribute equal weight to all agents.

Theorem 5.1.3. *Let all n agents have equal weights, then*

$$\frac{1}{n} \sum_{i \in I} risk_I(c_i, F) \leq \left(3 - \frac{2}{n}\right) r^*.$$

The key idea in all proofs is showing that both dictatorial risk and optimal risk depend on the average "distance" between agents' preferences. That is, the price of using a single dictator increases when agents tend to disagree on large parts of the space, but this also implies that the optimal risk must increase as well. Therefore, the approximation ratio remains constant (in the randomized cases).

Back to the Original Problem

We now derive upper bounds for the finite dataset case. Recall the Random Dictator mechanism from section 4.2. We slightly modify it to use weights instead of dataset sizes:[2]

Mechanism 5 THE WEIGHTED RANDOM DICTATOR MECHANISM ($\mathcal{M}_w$)

Select agent i w.p. w_i
Ask agent i for his labeled dataset S_i, **return** $c_i = argmin_{c \in \mathcal{C}} risk_i(c, S_i)$.

This assumption can be dropped when we return to the SIMILAR$_\mathcal{C}$ problem with finite datasets.

[2]Note that c_i is well defined for any $\mathcal{C}$, since S is finite.

Theorem 5.1.4. *The following statements hold for the SIMILAR$_\mathcal{C}$ problem, for any concept class $\mathcal{C}$:*

1. *The (deterministic) Heaviest Dictator mechanism is an SP $O(n)$-approximation mechanism.*

2. *The Weighted Random Dictator mechanism is an SP 3-approximation mechanism.*

3. *If all agents have equal weights, then the Weighted Random Dictator mechanism guarantees a $(3 - \frac{2}{n})$-approximation ratio.*

Proof. Clearly both mechanisms are SP, since the selected agent gets exactly what he wants, and all other agents are ignored.

Let S a dataset with similar interests. Set $\mathcal{X} = X$. The partial datasets $S_1, \ldots, S_n$ induce the following distributions $F_1, \ldots, F_n$ over $\mathcal{X} \times \mathcal{Y}$:

$$\Pr_{\langle x,y\rangle \sim F_i}(x = x', y = y') = \frac{1}{m'}$$

if $\langle x', y'\rangle \in S_i$ and 0 otherwise.

Note that according to this mapping, both definitions of risk are equivalent:

$$\forall i \in I \forall c \in \mathcal{C}\left(risk_i(c, S) = risk_i(c, F)\right),$$

and in particular, $argmin_{c\in\mathcal{C}} risk_i(c, S) = argmin_{c\in\mathcal{C}} risk_i(c, F) = c_i$, and $risk(c^*, S) = risk(c^*, F)$. [3] Therefore, any upper bound in the general distribution case also applies in the finite dataset case, and the three parts of the theorem follow directly from theorems 5.1.1, 5.1.2 and 5.1.3, respectively:

1. $risk_I(c_j, S) = risk_I(c_j, F) \leq (2n - 1)r^* = O(n)r^*$ where j is the heaviest agent,

2. $risk_I(\mathcal{M}_w, S) = \sum_{i\in I} w_i risk_I(c_i, S) = \sum_{i\in I} w_i risk_I(c_i, F) \leq 3r^*$, and

3. $risk_I(\mathcal{M}_w, S) \leq \left(3 - \frac{2}{n}\right) r^*$ when all weights are equal.

□

[3] The minimum is defined since F_i are distributions over a finite set.

5.2 Two Models of Generalization

Recall the generalization process described in section 2.2. We now repeat this process in detail, showing how one can build a learning algorithm for a given concept class, which guarantees a bounded approximation ratio of the risk.

As in section 4.3, we are required to distinguish the empirical risk (denoted by $\widehat{risk}(c, S)$) from the actual risk. The actual risk of each agent is computed w.r.t. a private distribution $\mathcal{D}_i$ over $\mathcal{X} \times \mathcal{Y}$. Using the same notations we used in section 2.2, $risk_i(c) = \mathbb{E}_{\langle x,y\rangle \sim \mathcal{D}_i}[[\![c(x) \neq y]\!]]$. We define the global distribution $\mathcal{D}$ to be $\sum_{i \in I} w_i \mathcal{D}_i$, i.e. the weighted average of agents' private distributions. We assume that the marginal of all the distributions on $\mathcal{X}$ is the same, and we denote it by $\mathcal{D}_{\mathcal{X}}$. The definition of the global risk does not change. Note that

$$risk_I(c) = \mathbb{E}_{\langle x,y\rangle \sim \mathcal{D}}[[\![c(x) \neq y]\!]] = \sum_i w_i risk_i(c).$$

Our goal is to find a classifier with low risk, thus we compare our result to the risk of the best concept in $\mathcal{C}$, denoted by $r_{min} = inf_{c \in \mathcal{C}} risk(c)$. Following the standard assumption in machine learning, we have no direct access to $\mathcal{D}$ or $\mathcal{D}_i$, and we can only sample from $\mathcal{D}_{\mathcal{X}}$ and ask the agents for their labels.

5.2.1 The Truthful Approach

In this section we show that if agents remain truthful even in face of a small potential gain, then we can learn any concept class and guarantee an approximation ratio close to 3.

Mechanism 6 THE GENERIC LEARNING MECHANISM ($\tilde{\mathcal{M}}_1$)

Sample m' data points i.i.d. from $\mathcal{D}_{\mathcal{X}}$. Denote the unlabeled dataset X.
Ask all agents to label X, and return their labeled datasets $\overline{S}_i = \langle X, \overline{Y}_i \rangle$.
Run mechanism 5 on $\overline{S} = \{\overline{S}_i\}_{i \leq n}$, using the original weights $w_1, \ldots, w_n$.
return $\mathcal{M}_w(\overline{S})$.

A few observations on mechanism 6 are in place. First, nothing changes if we swap the order of randomizations, i.e. select i as a dictator w.p. w_i and then sample a dataset. Thus, although $m = |\overline{S}| = n \cdot m'$ only m' data points are actually considered. Secondly, the real risk of agent i is computed w.r.t $\mathcal{D}_i$ and not w.r.t the sampled data points. This means that the mechanism is not necessarily SP, as it is possible that the sample X does not accurately reflect the interests of agents, and the reported labels Y_i may not be truthful. We denote by Y_i, S_i the real labels and labeled datasets, respectively. Thirdly, S_i is an i.i.d random sample from $\mathcal{D}_i$.

Thus if i is truthful $\overline{S}_i$ is also an i.i.d. sample, otherwise we do not assume anything on $\overline{S}_i$. However, $S = \bigcup_{i \in I} S_i$ is *not* i.i.d samples from $\mathcal{D}$.

Denote by $risk(\tilde{\mathcal{M}}_1)$ the expected risk of mechanism 6, where the expectation is taken over the randomness of the sampling and the randomization done by mechanism 5.

Theorem 5.2.1. *Let $\mathcal{C}$ any concept class with a bounded dimension. For any $\epsilon, \delta > 0$, there is m' (polynomial in $\frac{1}{\epsilon}, log\left(\frac{n}{\delta}\right)$), s.t. if we sample m' points, then with probability of at least $1-\delta$,*

1. *No agent can gain more than ϵ by lying.*
2. *If all agents are truthful then $risk_I(\tilde{\mathcal{M}}_1) \leq 3 \cdot r_{min} + \epsilon$.*

Note that probabilities and expectations are taken over both sampling and selection done by the mechanism.

Proof. Denote by $Q_i = Q_i(\epsilon)$ the event that

$$\forall c \in \mathcal{C} \left(|risk_i(c) - \widehat{risk}_i(c, S)| < \epsilon \right). \tag{5.1}$$

Recall equation (2.1) from section 2.2. It states that if S is a large enough i.i.d. sample from some distribution $\mathcal{D}$, then

$$\Pr\left(\forall c \in \mathcal{C} \left(|risk(c, \mathcal{D}) - \widehat{risk}(c, S)| < \epsilon \right)\right) > 1 - \delta.$$

As S_i is an i.i.d. random sample from $\mathcal{D}_i$, then for a large enough sample every Q_i occurs with probability of at least $1 - \delta$. Also, from the union bound the probability of the event $\forall j Q_j$ is at least $1 - \delta'$, where $\delta = \frac{\delta'}{n}$.

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.

Lemma 5.2.2. *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. $\hat{c}_i = argmin_{c \in \mathcal{C}} \widehat{risk}_i(c, S_i)$.

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

$$\begin{aligned}
risk_i(\hat{c}_i) - risk_i(c') =& risk_i(\hat{c}_i) - \widehat{risk}(\hat{c}_i, S_i) + \widehat{risk}(\hat{c}_i, S_i) - risk_i(c') \\
\leq& |risk_i(\hat{c}_i) - \widehat{risk}(\hat{c}_i, S_i)| \\
&+ |\widehat{risk}(c', S_i) - risk_i(c')| \qquad \text{(since } \hat{c}_i \text{ is empirically optimal)} \\
<& \epsilon + \epsilon = 2\epsilon, \qquad \text{(from (5.1))}
\end{aligned}$$

□

and thus i cannot gain more than 2ϵ by reporting c'. By taking $\epsilon < \frac{\epsilon'}{2}$, this proves the first part of the theorem.

Now, assume all agents are truthful (i.e. $\overline{S} = S$).

Lemma 5.2.3. *If S holds that Q_i occurs for* all $i \in I$*, then*

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

where $c^*(S) = argmin_{c \in \mathcal{C}} \widehat{risk}_I(c, S)$.

Proof. As we assumed all agents are truthful, $S_i = \overline{S}_i$ is an i.i.d sample from $\mathcal{D}_i$. From (5.1), if Q_i occurs then for any $c \in \mathcal{C}$, $|risk_i(c) - \widehat{risk}(c, S_i)| < \epsilon$. Therefore

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

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

We now bound the expected risk of the mechanism. We denote by $c_{\mathcal{M}}(S)$ the (random) classifier that is returned by mechanism 5 on the input S. For any random variable A, $\mathbb{E}_{\mathcal{M}}[A|S]$ is the expectation of A over the randomization for a fixed dataset S. $\mathbb{E}_S[A|i]$ is the expectation

of A over the random sampling, given that i is the selected dictator.

$$\begin{aligned}
risk_I(\tilde{\mathcal{M}}_1) =& \mathbb{E}\left[risk_I(c_{\mathcal{M}}(S))\right] = \mathbb{E}_S\left[\mathbb{E}_{\mathcal{M}}\left[risk_I(c_{\mathcal{M}}(S))|S\right]\right] \\
=& \Pr(\forall j Q_j)\mathbb{E}_S\left[\mathbb{E}_{\mathcal{M}}\left[risk_I(c_{\mathcal{M}}(S))|S\right]|\forall j Q_j\right] \\
&+ \Pr(\neg\forall j Q_j)\mathbb{E}_S\left[\mathbb{E}_{\mathcal{M}}\left[risk_I(c_{\mathcal{M}}(S))|S\right]|\neg\forall j Q_j\right] \\
\leq& \mathbb{E}_S\left[\mathbb{E}_{\mathcal{M}}\left[risk_I(c_{\mathcal{M}}(S))|S\right]|\forall j Q_j\right] + \delta' \cdot 1 \\
=& \mathbb{E}_{\mathcal{M}}\left[\mathbb{E}_S\left[risk_I(c_{\mathcal{M}}(S))|i, \forall j Q_j\right]\right] + \delta' \quad \text{(changing the order of randomizations)} \\
=& \sum_{i\in I} w_i \mathbb{E}_S\left[risk_I(\hat{c}_i(S))|i, \forall j Q_j\right] + \delta' \\
\leq& \sum_{i\in I} w_i \mathbb{E}_S\left[\widehat{risk}_I(\hat{c}_i(S), S_i) + \epsilon|i, \forall j Q_j\right] + \delta' \qquad \text{(from eq. (5.1))} \\
=& \sum_{i\in I} w_i \mathbb{E}_S\left[\widehat{risk}_I(\hat{c}_i(S), S_i)|i, \forall j Q_j\right] + \delta' + \epsilon \\
=& \mathbb{E}_{\mathcal{M}}\left[\mathbb{E}_S\left[\widehat{risk}_I(c_{\mathcal{M}}(S), S)|i, \forall j Q_j\right]\right] + \delta' + \epsilon \\
\leq& \mathbb{E}_{\mathcal{M}}\left[\mathbb{E}_S\left[3\widehat{risk}_I(c^*(S), S)|i, \forall j Q_j\right]\right] + \delta' + \epsilon \qquad \text{(from theorem 5.1.4)} \\
\leq& \mathbb{E}_{\mathcal{M}}\left[\mathbb{E}_S\left[3(r_{min} + \epsilon)|i, \forall j Q_j\right]\right] + \delta' + \epsilon \qquad \text{(from lemma 5.2.3)} \\
=& 3(r_{min} + \epsilon) + \delta' + \epsilon = 3 \cdot r_{min} + \delta' + 4\epsilon = 3 \cdot r_{min} + \epsilon'
\end{aligned}$$

This proves the second part of the theorem. □

Corollary 5.2.4. *If agents can be assumed to be truthful when they cannot benefit more than ϵ from lying, then $\tilde{\mathcal{M}}_1$ is $(3 + \epsilon'')$-approximating.*

Proof. This assumption means simply that whenever the first part of theorem 5.2.1 is true (i.e. all agents have a small incentive to lie), then the header of the second part is also true and the approximation ratio holds. In other words, the mechanism is $3 + \epsilon'$ approximating w.p. of at least $1 - \delta'$.

Denote by T the event that the first part of the theorem holds.

$$\begin{aligned}
risk_I(\tilde{\mathcal{M}}_1) =& \Pr(T) risk_I(\tilde{\mathcal{M}}_1|T) + \Pr(\neg T) risk_I(\tilde{\mathcal{M}}_1|\neg T) \\
\leq& risk_I(\tilde{\mathcal{M}}_1|T) + \delta' risk_I(\tilde{\mathcal{M}}_1|\neg T) \\
\leq& 3 \cdot r_{min} + \epsilon' + \delta' = 3 \cdot r_{min} + \epsilon'' \qquad \text{(from the second part of the theorem)}
\end{aligned}$$

□

We conclude this section by computing the exact number of samples needed by the mechanism $\tilde{\mathcal{M}}_1$.

Lemma 5.2.5. *If* $m' > 64\frac{V_C}{\epsilon^2} log\left(256\frac{V_C \cdot n}{\epsilon^3}\right)$*, then*

$$risk_I(\tilde{\mathcal{M}}_1) \leq 3 \cdot r_{min} + \epsilon.$$

Proof. From theorem A.0.1, 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

$$risk_I(\tilde{\mathcal{M}}_1) \leq 3 \cdot r_{min} = 3 \cdot r_{min} + 4\epsilon^* + 2n\delta^* < 3 \cdot r_{min} + 4\frac{\epsilon}{8} + 2n\frac{\epsilon}{4n} = 3 \cdot r_{min} + \epsilon,$$

while

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

□

5.2.2 The Rational Approach

In this section we make a slightly different assumption on agents' behavior, i.e. that they always report labels in a way that will maximize their utility (minimize their risk). We also slightly alter the learning mechanism to show an improved bound on the global risk under the new assumption.

Definition 5.2.6. An agent $i \in I$ is *purely rational* if he always follows a dominant strategy when one exists. In our context, a purely rational agent will always report his labels such that $\mathbb{E}_{\mathcal{M}}\left[risk_i(c_{\mathcal{M}}(\overline{S}))|\overline{S}\right]$ is minimal, if such labeling exists. Otherwise (i.e. if the optimal labeling depends on other agents' strategies), the strategy of i is undefined.

Consider mechanism 7. We define the real dataset as $S = \bigcup_{i \in I} S_i$. Note that $|\overline{S}_j| = m$ while S_j is smaller, and that unlike mechanism 6, here the real dataset S is an i.i.d sample from $\mathcal{D}$. Also, mechanism 7 does *not* explicitly apply mechanism 5 as part of the computation. Like in the previous mechanism, S_i is still an i.i.d sample from $\mathcal{D}_i$. Lastly, note that the selected agent does not need to consider the labels of other agents when reporting his own, thus $\overline{Y}_j$ is well-defined.

Mechanism 7 THE GENERIC LEARNING MECHANISM ($\tilde{\mathcal{M}}_2$)

for $t = 1, \ldots, m$ **do**
  Select i at random using weights $w = w_1, \ldots, w_n$.
  Sample $\langle x, y \rangle$ from $\mathcal{D}_i$.
  Add $\langle x, y \rangle$ to S_i.
end for
Select j at random using weights w.
Ask agent j for the labels of *all* data points $X = \langle X_1, \ldots, X_n \rangle$.
Create the dataset $\overline{S}_j = \langle X, \overline{Y}_j \rangle$ using the reported labels of j.
return $\hat{c}_j = argmin_{c \in \mathcal{C}} \widehat{risk}(c, \overline{S}_j)$.

Theorem 5.2.7. *Let $\mathcal{C}$ any concept class with a bounded dimension. Assume all agents are purely rational. For any ϵ, there is m (polynomial in $\frac{1}{\epsilon}$) such that*

$$risk_I(\tilde{\mathcal{M}}_2) \leq 3 \cdot r_{min} + \epsilon,$$

where expectation is taken over both sampling and selection.

Proof. We denote by $c^*(S)$ the best concept (in $\mathcal{C}$) for the real dataset S. Note that S is an i.i.d sample from $\mathcal{D}$, but our mechanism has no access to S (we do not assume agents are truthful).

We denote by T the event

$$risk_I(c^*(S)) < r_{min} + 2\epsilon.$$

As in the previous section, T is a property of S, i.e. its occurrence depends only on the sampling.

Lemma 5.2.8. *If $m = m(\delta, \epsilon)$ is large enough then*

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

Proof. This is an immediate corollary of equation (2.1). As $c^* = argmin_{c \in \mathcal{C}} \widehat{risk}_I(c, S)$, $\mathcal{C}$ is of a bounded dimension and S is sampled i.i.d. from $\mathcal{D}$, then for any $c \in \mathcal{C}$

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

holds w.p. of at least $1 - \delta$, for a large enough m. In particular,

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

□

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 $\mathcal{F}$ the class of all dichotomies of X, i.e. all $f s.t. f : X \to \{-,+\}^m$, then $\mathcal{C}_X = \mathcal{C} \cap \mathcal{F}$. In other words, $\mathcal{C}_X$ contains all dichotomies of X that are allowed by $\mathcal{C}$.

Observe that $c^*(S) \in \mathcal{C}_X$ and $\hat{c}_j \in \mathcal{C}_X$ for all agents.This is 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 $\tilde{c} = argmin_{c \in \mathcal{C}_X} risk_I(c)$. Clearly $risk_I(\tilde{c}) \leq risk_I(c^*(S))$, since $c^*(S)$ is also a member of $\mathcal{C}_X$. Thus when T occurs the inequality

$$risk_I(\tilde{c}) < r_{min} + 2\epsilon \tag{5.2}$$

also holds.

We next show how to approximate $\tilde{c}$ using the general distribution theorem we proved in the beginning of the chapter. Let the profile $F = \langle \mathcal{D}_1, \ldots, \mathcal{D}_n \rangle$. This is a valid profile with similar interests, thus for any concept $c \in \mathcal{C}$, $risk(c) = risk(c, F)$ for private and global risk alike.

Lemma 5.2.9. *Let j be the selected dictator, then*

$$\hat{c}_j = argmin_{c \in \mathcal{C}_X} risk_j(c) = argmin_{c \in \mathcal{C}_X} risk_j(c, \overline{F}).$$

Proof. Recall that $\hat{c}_j \equiv argmin_{c \in \mathcal{C}} \widehat{risk}(c, \overline{S}_j)$. 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 7 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 $argmin_{c \in \mathcal{C}_X} risk_j(c)$ is the best that agent j can hope for, and he can also achieve it by reporting the appropriate labels $\overline{Y}_j$. □

We now apply theorem 5.1.2 on F, using the class $\mathcal{C}_X$, getting

$$\sum_{j \in I} w_j risk_I(\hat{c}_j, F) \leq 3 \cdot risk_I(\tilde{c}, F) = 3 \cdot risk_I(\tilde{c}). \tag{5.3}$$

We emphasize that equation (5.3) *always* holds, independently of the sampling or selection.

Finally, we bound the risk of the result concept:

$$
\begin{aligned}
risk_I(\tilde{\mathcal{M}}_2) =& \mathbb{E}_S\left[\mathbb{E}_{\mathcal{M}}\left[risk_I(c_{\mathcal{M}})|S\right]\right] \\
=& \Pr(T)\mathbb{E}_S\left[\mathbb{E}_{\mathcal{M}}\left[risk_I(c_{\mathcal{M}})|S\right]|T\right] + \Pr(\neg T)\mathbb{E}_S\left[\mathbb{E}_{\mathcal{M}}\left[risk_I(c_{\mathcal{M}})|S\right]|\neg T\right] \\
\leq& \mathbb{E}_S\left[\mathbb{E}_{\mathcal{M}}\left[risk_I(c_{\mathcal{M}})|S\right]|T\right] + \delta \cdot 1 && \text{(from lemma 5.2.8)} \\
=& \mathbb{E}_S\left[\sum_{j\in I} w_j risk_I(\hat{c}_j(S))|T\right] + \delta \\
\leq& \mathbb{E}_S\left[3 \cdot risk_I(\tilde{c}(S))|T\right] + \delta && \text{(from eq. (5.3))} \\
<& 3\mathbb{E}_S\left[(r_{min} + 2\epsilon)|T\right] + \delta && \text{(from eq. (5.2))} \\
=& 3(r_{min} + 2\epsilon) + \delta = 3 \cdot r_{min} + 6\epsilon + \delta.
\end{aligned}
$$

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

Similarly to lemma 5.2.5, it follows from theorem A.0.1 that

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

is sufficient for the theorem to hold. □

Comparing the two approaches

We discussed the different assumptions underlying the two generalization models (i.e. theorems 5.2.1 and 5.2.7) in section 4.3. Note that not only the second model uses a different assumption on agents' behavior, it also provides us with an improved bound that does not depend on the number of agents n. Somewhat surprisingly, the rational approach supplies 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 agent's real preferences than a truthful labeling, as random samples might represent the agent's interest in an inaccurate way. We stress this point by reverting to the PLUS-MINUS scenario. By assuming all agents are purely rational, we can generalize any of the bounds we showed in chapter 4 by sampling only one data point (see remark 4.3.1).

5.3 Lower Bounds

We show that our analysis of the weighted random dictator mechanism is tight, even in the most simple setting.

Theorem 5.3.1. *For every concept class of size at least 2, there is an instance S with equal weights for which $risk_I(\mathcal{M}_w, S) = \left(3 - \frac{2}{n}\right) r^*$.*

Proof. We first prove for a concept class that contains only two classifiers: c_+, c_-, which classify *all* data points as positive or all of them as negative, respectively. This is equivalent to the PLUS-MINUS scenario presented in chapter 4. Moreover, we contruct a counter example that is very similar to the one we used to demostrate tightness of mechanism 2.

Our dataset is constructed as follows: there are n agents. One agent (w.l.o.g. agent 1) with n positive examples and n negative ones , and $n-1$ additional agents each holding $2n$ negative examples. Thus there are only n positive examples in total. Note that $m = n(2n)$. denote by j the agent selected by the mechanism as a dictator. The optimal classifier makes $n = m \cdot r^*$ mistakes, whereas the expected number of mistakes done by the mechanism is

$$\begin{aligned} m \cdot risk(\mathcal{M}_w, S) &= \Pr(j \neq 1) \cdot n + \Pr(j = 1) \cdot (m - n) \\ &= \frac{n-1}{n} n + \frac{n(2n) - n}{n} = n - 1 + 2n - 1 = 3n - 2 \end{aligned}$$

We get that the approximation ratio of $\mathcal{M}_w$ is $\frac{risk_I(\mathcal{M}_w, S)}{r^*} = \frac{m(3n-2)}{m \cdot n} = 3 - \frac{2}{n}$. For any other non-trivial concept class $\mathcal{C}$, there is at least one point $x' \in \mathcal{X}$ that is classified as positive by some concepts in $\mathcal{C}$, and as negative by all others. Therefore we build our example by placing all data points in x'. □

Thus the bound in theorem 5.1.3 is tight for $\mathcal{M}_w$. If weights are allowed, then a counter example with approximation of at least $3 - \epsilon$ can be constructed for every $\epsilon > 0$, using only two agents. this is simply by replacing agents $2, \ldots, n$ with a single agent of higher weight. A similar analysis of this counter-example shows that the deterministic mechanism $\mathcal{M}_h$ will attain an approximation ratio of $2n - 1$. Therefore, the bounds in theorems 5.1.3 and 5.1.1 are also tight.

However, it does not prove that there isn't some *other mechanism* capable of achieving a better approximation ratio than 3. Indeed, theorem 4.2.3 shows that at least for the special case of PLUS-MINUS , a better approximation of 2 is at hand. Since 2 is also a lower bound (theorem 4.2.1), we conclude from these results that for any profile $\overline{F}$ (or, alternatively, any dataset S with similar interests) and for any concept class $\mathcal{C}$, the worst case approximation ratio of the best randomized SP mechanism has to lie between 2 and 3. The exact value may depend on the characteristics of $\mathcal{C}$ and $\overline{F} \backslash S$.

What about deterministic mechanisms? For the PLUS-MINUS problem, we saw that the difference in the approximation ratio of deterministic and randomized SP mechanisms is not so

large. However, in the general case the situation is different.

Theorem 5.3.2. *There are concept classes for which any deterministic SP mechanism cannot guarantee a constant approximation ratio. More precisely, the approximation ratio of any such mechanism is bounded from below by $\Omega(n)$, the number of agents, even if all weights are equal.*

A sketch of the proof appears in the next chapter, as a by-product of a similar theorem. Recall that from theorem 5.1.4 there is a trivial SP $O(n)$-approximation mechanism, thus this bound is tight.

Chapter 6

Classification of Arbitrary Data

In this chapter we drop the rather strong assumption that agents have similar interests. Instead, our input is an arbitrary dataset $\langle S_1, \ldots, S_n \rangle$, without any assumptions on the structure of examples. We show that there are concept classes, including the widely-used class of Linear Separators, for which there are no reasonable Deterministic Strategy-Proof learning mechanisms. We also show that even allowing randomness cannot guarantee a constant approximation ratio with SP mechanisms. On the positive side, we show an SP mechanism whose approximation ratio is polynomially bounded by the size of the largest partial dataset. The underlying idea that guides our steps, as in previous chapters, is the tight coupling between Strategy-Proofness and dictatorial selection of the classifier.

6.1 Lower Bounds for Synthetic Scenarios

Deterministic Lower Bound

In this section, we describe a problem with the following property: the only deterministic learning mechanisms that are SP, are those in which there is a fixed dictator, and all other agents are ignored. For that matter we put forward some formal background from Social Choice theory, which is used thereafter in the proving the desired properties.

Voting and the Gibbard-Satterthwaite theorem

let $\mathbf{C}$ be a set of candidates. Each voter i holds a (private) strict order of preferences P_i over $\mathbf{C}$. For any $c_1, c_2 \in \mathbf{C}$, whenever voter i *prefers* c_1 over c_2 we denote it by $c_1 \succ_i c_2$ or, equivalently, by $(c_1, c_2) \in P_i$.

A set of preferences for all n voters is called a *profile*, and denoted by $\mathbf{P} = \{P_1, \ldots, P_n\}$.

We denote by P_{-i} the set of preferences for all agents except i, thus $\forall i\,(\mathbf{P} = (P_i, P_{-i}))$. We denote the candidate i prefers most by $P_i(1)$.

Definition 6.1.1. Let φ the set of all possible preference profiles. A *voting rule* $f : \varphi \to \mathbf{C}$ is a function from voting profiles to a winning candidate.

- A voting rule f is *manipulable* if there is a profile $\mathbf{P} \in \varphi$ and some preference P_i' of voter i, s.t. i strictly gains (according to P_i) by voting P_i' instead:

  $$f(P_{-i}, P_i') \succ_i f(\mathbf{P})$$

  if f is not manipulable, it is said to be *honest*.[1]

- A voting rule f is *dictatorial* if there is some voter i, whose favorite candidate is always the winner:

  $$\forall \mathbf{P} \in \varphi\,(f(\mathbf{P}) = P_i(1))$$

- A voting rule f is *onto* if any candidate can be elected:

  $$\forall c \in \mathbf{C} \exists \mathbf{P} \in \varphi\,(f(\mathbf{P}) = c)$$

- A *duple* is a voting rule whose range is of size at most 2. If $|\mathbf{C}| = 3$ then f is a duple iff it is not onto.

Theorem 6.1.2 (Gibbard '73 [17], Satterthwaite '75 [33])**.** *If there are at least 3 candidates, and f is onto and honest, then f is dictatorial.*

Classification as voting

We build a synthetic classification problem, in which the examples X_i of all agents are predetermined, and there are only 3 allowed classifiers. We then construct a reduction from the voting problem to our classification problem, in a way that preserves the properties and restrictions on voting rules. This will enable us to show that any reasonable classification algorithm induces a legal voting rule, and thus cannot be both SP and non-dictatorial. We then show that the approximation ratio of any dictatorial algorithm in the synthetic problem is at least $\Omega(m)$. Later in this chapter, we show lower bounds for more interesting concept classes using reductions to this synthetic scenario.

[1] Honest voting rules are also known as strategy-proof. For easy discrimination between properties of voting rules and those of classification mechanisms, we use the term "honest" for the first and "strategy-proof" for the latter.

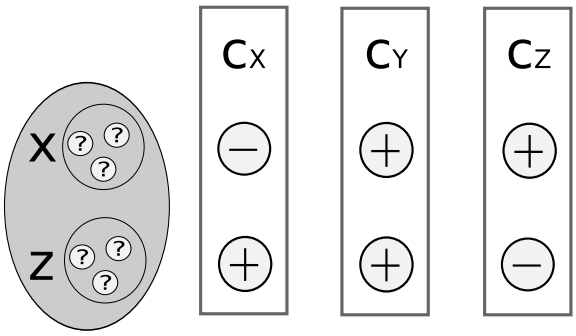


Figure 6.1: The classifiers of the DET-SYNTHETIC scenario.

In our scenario, $\mathcal{X} = \{x, z\}$, i.e. there are only two "locations" in which example may be found. There are 3 possible classifiers: c_X, which labels as negative all (and only) examples in x; c_Z, which does the opposite; and c_Y, which labels *all* examples as positive. Figure 6.1 shows how each classifier labels each example according to its location. Instead of considering the entire set of possible datasets $\mathcal{S}$, our proof will focus on a restricted set $\mathcal{S}^* \subset \mathcal{S}$, which will be defined later on. Clearly if approximation ratio is lower bounded in the restricted case, the bound also carries to the general case.

Definition 6.1.3. $\mathcal{M}$ is *dictatorial* if there exists agent i s.t. for any $S \in \mathcal{S}^*$ the result depends only on the labeling of S_i.

Definition 6.1.4. $\mathcal{M}$ is *onto* if for any concept $c \in \mathcal{C}$ there is some instance $S \in \mathcal{S}^*$ s.t.

$$\mathcal{M}(S) = c.$$

Recall that $m = |S|$, i.e. the total size of the dataset.

Theorem 6.1.5. *Let $\mathcal{M}$ a deterministic classification mechanism that learns the concept class $\mathcal{C}_s = \{c_X, c_Y, c_Z\}$.*

1. *If $\mathcal{M}$ is SP and onto, it is dictatorial.*
2. *If $\mathcal{M}$ is not onto, then it cannot guarantee any finite approximation ratio.*
3. *If $\mathcal{M}$ is dictatorial, then it cannot guarantee β-approximation, for any $\beta = o(m)$.*

The lower bound is an immediate result of the theorem:

Corollary 6.1.6. *Any deterministic SP classification mechanism for DET-SYNTHETIC cannot give any approximation ratio which is better than $\Omega(m)$.*

Note that an approximation of $O(m)$ is trivially achieved by *any* classifier unless all agents fully agree, so this is a very strong negative result.[2]

Proof of theorem 6.1.5, part 2. It is easy to see, that if $\mathcal{M}$ is not onto, then there is an instance for which $r^*(S) = 0$ but $risk(\mathcal{M}, S) > 0$.[3] Thus such a mechanism cannot guarantee any finite approximation ratio at all. □

Proof of theorem 6.1.5, part 1. The exact structure of S (without labels) is as follows: There are two kinds of agents, $\frac{n}{2}$ agents of each kind. Agents from the first kind hold 2 examples on x and $2k + 1$ examples on z. Agents from the second kind are symmetric w.r.t. x, z.

In our small scenario, the examples (X_i) are fixed, and only the labeling (Y_i) may change. As a classification mechanism in general is a function from labeled examples, we take into account only the labeling when analyzing the mechanism, that is $\mathcal{M}(S) = \mathcal{M}(\langle Y_1, \ldots, Y_n \rangle)$. Denote by $\mathbf{Y}$ the full labeling profile $\langle Y_1, \ldots, Y_n \rangle$, and let Υ the set of all possible labellings of X by the n agents. We use any of the notations $risk(c, S),\ risk(c, \mathbf{Y}),\ risk(c)$ to denote the risk (either private or global) of classifier c.

Any deterministic classification mechanism in our scenario is a function

$$\mathcal{M} : \Upsilon \to \{c_X, c_Y, c_Z\}$$

This simplified presentation of the mechanism is beginning to resemble the definition of voting rules, and this is not accidental. We now show how to reduce any voting instance to a classification instance, by translating any voting profile to labeling of the examples in X.

The properties of *onto* and *dictatorial* are similar for voting schemes and classification mechanisms. The property of *honesty* (in voting) is closely related to that of *Strategy-Proofness* (in classification), in a way that is elaborated in the remainder of this section.

We define our set of candidates to be $\mathbf{C} = \{c_X, c_Y, c_Z\} = \mathcal{C}_s$ (thus it is of size 3). When agent i labels his set of examples X_i with the labeling Y_i, it induces a preference ranking over all possible classifiers (possibly with ties).

Definition 6.1.7. Let Y_i, P_i. P_i *fits* Y_i if for all $c_1, c_2 \in \mathbf{C}$,

$$risk_i(c_1) < risk_i(c_2) \iff c_1 \succ_i c_2. \tag{6.1}$$

[2] It is still required to show that a deterministic mechanism can tell whether all agents fully agree or not. This can be achieved by any deterministic variant of the Iterative Random Dictator mechanism, which is presented in the next section.

[3] Just consider a labeling in which the out-of-range-classifier is the only perfect classifier.

That is, the order of preference over the three possible classifiers induced by the labeling Y_i, is exactly P_i.

Note that only one order may fit any given labeling.

From the definition it is clear that there is a natural mapping from $\mathbf{Y}$ to $\mathbf{P}$. This is not enough though, as our reduction requires a mapping in the other direction. For each agent i of the first kind, we define $g_i : P_i \to Y_i$ in the following way:

	$g_i(P_i)$					
P_i	x	z_1	z_2	$risk_i(c_X)$	$risk_i(c_Y)$	$risk_i(c_Z)$
$c_X \succ c_Y \succ c_Z$	-	+	+	0	2	$2k+1$
$c_X \succ c_Z \succ c_Y$	-	+	-	$k+1$	$k+3$	$k+2$
$c_Y \succ c_X \succ c_Z$	+	+	+	2	0	$2k+1$
$c_Y \succ c_Z \succ c_X$	+	-	+	$k+2$	k	$k+1$
$c_Z \succ c_X \succ c_Y$	-	-	-	$2k+1$	$2k+3$	2
$c_Z \succ c_Y \succ c_X$	+	-	-	$2k+3$	$2k+1$	0

The leftmost column enumerates all 6 possible orders over classifiers. The next 3 columns define the label of each example according to $Y_i = g_i(P_i)$: We use x to denote the 2 examples on x, z_1 to denote the first (arbitrary) k examples on z, and z_2 to denote the other $k+1$ examples on z.

We can now formally define the set $\mathcal{S}^*$: it simply contains all datasets in which the labellings $Y_1, Y_2, \ldots$ are restricted to one of the 6 labellings in the table.

The last 3 columns show the risk of each classifier according to Y_i, and can be used to verify that each order P_i indeed fits the labeling $Y_i = g_i(P_i)$. Since there are only 6 possible orders on $\mathbf{C}$, g_i is well defined. For agents of the second kind, g_i is defined in a symmetric way w.r.t. x, z. To conclude, the full mapping is naturally defined by taking $g(\mathbf{P}) = \langle g_1(P_1), \ldots, g_n(P_n) \rangle \in \Upsilon$.

By combining with g, every classification mechanism $\mathcal{M}$ induces a legal voting rule $(\mathcal{M} \circ g) : \varphi \to \mathbf{C}$.

Lemma 6.1.8. *Denote $f = \mathcal{M} \circ g$. If $\mathcal{M}$ is SP and onto, then f is honest and onto.*

Proof. The onto property of f derives immediately from the facts that g is well defined and $\mathcal{M}$ is onto.

Assume that f is not honest. Thus there are $\mathbf{P}, P_i'$ such that

$$f(P_{-i}, P_i') \succ_i f(\mathbf{P}).$$

From the definition of f,

$$\mathcal{M}(g(P_{-i}, P_i')) \succ_i \mathcal{M}(g(\mathbf{P})).$$

Let $\mathbf{Y} = g(\mathbf{P}),\ Y_i' = g(P_i')$. Thus

$$\mathcal{M}(Y_{-i}, Y_i') \succ_i \mathcal{M}(\mathbf{Y})$$

w.r.t. the labeling Y_i. From the definition of g, Y_i fits P_i, thus by using equation (6.1) we get that

$$risk_i(\mathcal{M}(Y_{-i}, Y_i')) < risk_i(\mathcal{M}(\mathbf{Y})),$$

therefore agent i strictly gains by misreporting his true labels, in contradiction to strategy-proofness of $\mathcal{M}$. □

From theorem 6.1.2 and lemma 6.1.8, the only deterministic classification algorithms that are immune to manipulation (i.e. in which agents cannot gain by mis-labeling) and onto, are those in which one of the agents is a dictator. Assuming $\mathcal{M}$ is both SP and onto, it must therefore be dictatorial. □

Proof of theorem 6.1.5, part 3. Let $\beta = o(m)$, and $k > 1$. Set m high enough, so $m > \frac{4}{\beta - 1}$. In our scenario $\forall i\ (|S_i| = 2k + 3)$, thus the total number of examples $|S| = \sum_{i \leq n} |S_i| = n(2k+3)$.

Given any (deterministic, onto) SP mechanism $\mathcal{M}$, we show how to add actual labels to our scenario DET-SYNTHETIC , s.t. $risk(\mathcal{M}(S), S) > \beta r^*(S)$. As $\mathcal{M}$ necessarily picks a dictator, and is not allowed to randomize, we may assume that the chosen dictator is some agent i', and i' is fixed across different possible labellings of S. W.l.o.g i' is of kind 1, i.e. holds 2 examples on x and $2k + 1$ examples on z. We label the examples of agent i' that are on x as negative, and label all other examples, of all other agents, as positive. We can verify that

$$r^*(S) = risk(c_Y, S) = \frac{1}{m}|\{\text{negative examples in } S\}| = \frac{2}{m}.$$

On the other hand, as $\mathcal{M}$ is onto and agent i' clearly prefers classifier c_X, he labels his examples s.t. c_X is always chosen by $\mathcal{M}$. Thus

$$risk(\mathcal{M}(S), S) = risk(c_X, S) = \frac{1}{m}|\{\text{positive examples on } x\}| = \frac{1}{2} - \frac{2}{m},$$

and the approximation ratio is bounded from below by

$$\frac{risk(\mathcal{M}(S), S)}{r^*(S)} = \frac{\frac{1}{2} - \frac{2}{m}}{\frac{2}{m}} = \frac{m}{4} - 1 > \beta.$$

□

Similar Interests

Suppose that all agents had similar interests as in chapter 5, e.g. that each agent holds half of its data points on x the other half on z. A very similar reduction would show that in this case too, the only reasonable SP mechanism is one that chooses an arbitrary dictator, w.l.o.g. agent 1. By letting agent 1 disagree with all other agents, we get that any deterministic SP learning mechanism cannot guarantee an approximation ratio better than $n = |I|$, even if all agents have similar interests and equal weights. This (almost) proves theorem 5.3.2. The remaining of the proof, i.e. the complete reduction to the voting problem is almost identical to the one in theorem 6.1.5, and is therefore omitted.

Randomized Lower Bounds

A nice result would be to take DET-SYNTHETIC , and prove that the expected risk of any randomized SP mechanism is also bounded from below by some function of the dataset. Instead, we build an alternative synthetic scenario RAND-SYNTHETIC , and prove a weaker claim on it: If agents also take into account *private weights* of examples, then any SP randomized mechanism for RAND-SYNTHETIC cannot guarantee an approximation ratio which is better than the size of the largest partial dataset. As in the deterministic case, the heart of the proof lies in using results from Social Choice theory to show an inevitable linkage between Strategy-Proofness and dictatorial decision.

Randomized voting rules

As in the deterministic case, we put forward the relevant Social Choice theoretical background.

A randomized voting rule (decision scheme) $f : \varphi \to \Delta(\mathbf{C})$ is a function from profiles to lotteries over candidates.

Definition 6.1.9. A utility scale $U : C \to \mathbb{R}_+$ *fits* P_i if

$$U(c_1) > U(c_2) \iff c_1 \prec_i c_2$$

Any labeling of S_i induces a natural utility scale on $\mathbf{C}$: $\forall c \in \mathbf{C}, U_{Y_i}(c) = risk_i(c)$. Clearly U_{Y_i} fits P_i iff Y_i fits P_i. In order to comply with our definition of risk, we treat lower utility as better.

The definition of manipulability for randomized rules is more flexible than in the deterministic case.

Definition 6.1.10. A manipulation of a randomized voting rule f, consists of a profile $\mathbf{P}$, a manipulator i which votes P'_i, and a utility scale U, such that

1. U fits P_i

2. $U(f(P_{-i}, P'_i)) < U(f(\mathbf{P}))$

That is, i strictly gains utility (according to U) by lying. f is *manipulable* if there exists a manipulation of f.

Equivalently, f is *honest* if for all $\mathbf{P}, i, P'_i$, for any utility U that fits P_i,

$$U(f(P_{-i}, P'_i)) \geq U(f(\mathbf{P})).$$

Theorem 6.1.11 (Gibbard '77 [18])**.** *If f is an honest randomized voting rule, then it is a probability mixture of duples and dictatorial (deterministic) voting rules.*

The synthetic scenario RAND-SYNTHETIC

We first explain where our deterministic reduction from DET-SYNTHETIC fails in the randomized case.

In the learning problem, an algorithm is SP if for all profiles $\mathbf{P}$ and labellings that match $\mathbf{P}$ there is no beneficial lie. Since not any real-valued utility can be expressed by (discrete) labeling of the examples,[4] there may potentially exist a randomized learning algorithm that is SP in the learning sense, but the induced voting rule is not honest, and thus the Gibbard theorem does not apply to it.

To handle this shortcoming, we relax our definition of the private risk to allow each agent to hold private non-negative *weights* over examples, $\sum_{j=1}^{m_i} w_j = 1$. The private risk (or utility) of agent i w.r.t. some classifier c and a partial dataset S_i is now defined as

$$risk_i(c, S_i) = \sum_{j=1}^{m_i} w_{i,j} [\![c(x_{i,j}), y_{i,j}]\!]. \tag{6.2}$$

Just like in the unweighted case, we say that $\mathcal{M}$ is SP if no agent can make its expected risk (including the weights) strictly lower by reporting some false labels. We emphasize that in contrast to chapter 5, the weights are attributed to *examples* and not to *agents*. Also, the learning algorithm *is unaware* of these weights, and can only take into account the data points and their reported labels.

We now construct a reduction that preserves the properties of the randomized voting rule. Let $k > 5$. We build a new synthetic scenario RAND-SYNTHETIC , this time with 3 possible locations x, y, z, and only 2 agents. Agent 1 holds $2k+1$ examples on x, 7 examples on y and 7

[4] In the deterministic case we observed that the labeling contains *more* information than the ordering and this was ok. Here we face a problem since the labeling contains *less* information than an arbitrary utility scale.

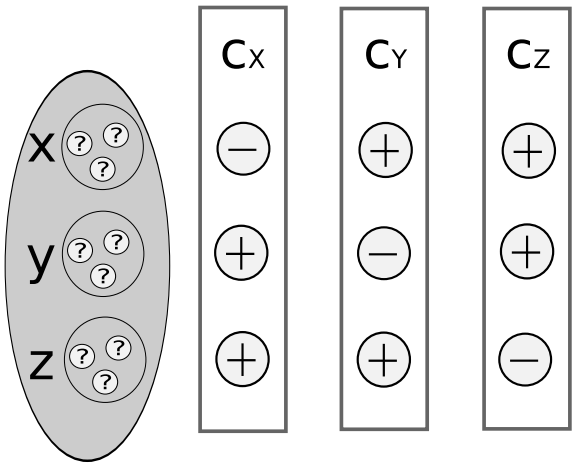


Figure 6.2: The classifiers of the RAND-SYNTHETIC scenario.

on z. Agent 2 is symmetric w.r.t. x, z. We have 3 possible classifiers. c_X classifies as negative all examples on x, and as positive all other examples. c_Y, c_Z are defined likewise. Figure 6.2 shows how each classifier labels each example according to its location. The weights and labels of all examples will be defined for each instance in the reduction.

If a classification mechanism is not onto (i.e. one of the 3 classifiers always has a probability of 0), we say it is a *duple*.

Theorem 6.1.12. *Let $\mathcal{M}$ a randomized classification mechanism for learning RAND-SYNTHETIC , and assume experts are allowed to attribute (private, real and non-negative) weights to their examples.*

1. *If $\mathcal{M}$ is SP then it is a probability mixture of duples and dictatorial mechanisms.*
2. *if $\mathcal{M}$ is a probability mixture of duples and dictatorial mechanisms, then it has an approximation ratio of $\Omega(k)$,* ***even without weights****.*

As in the deterministic case, joining the parts of the theorem leads to a lower bound:

Corollary 6.1.13. *Let $\mathcal{M}$ a randomized classification mechanism for learning RAND-SYNTHETIC , and assume experts are allowed to attribute weights to their examples, then $\mathcal{M}$ has an approximation ratio of $\Omega(k)$.*

proof of theorem 6.1.12, part 2. Suppose that $\mathcal{M}$ select a duple with probability $p > 0$, w.l.o.g. a mechanism which never outputs c_X. Let all agents label their examples according to c_X. As the other classifiers are distinct than c_X, $\mathcal{M}$ makes some errors w.p. of at least p. On the other hand, c_X makes no errors, thus $r^* = 0$ and the approximation ratio is infinite.

Now suppose that the mechanism must select one of the two agents as a dictator, with some probability assigned to each agent. We complete the labels in our scenario as follows: The labels of agent 1's examples on z are negative, and so are the labels of agent 2's on x. All other labels are positive. Clearly if either agent is a dictator, the the result concept must be c_X (if agent 2 is selected) or c_Z (if agent 1 is selected). Thus the risk of the mechanism is at least $2k+1$, while the risk of $c^* = c_Y$ is a constant (28). □

It is worthwhile to notice that a similar proof can be constructed for *almost any concept class* containing at least 3 distinct classifiers. Therefore, it should be clear that mixtures of dictatorial mechanisms (and duples) can almost never guarantee a constant approximation ratio.

We thereby turn to show that unfortunately, every SP mechanism must be such a mixture.

proof of theorem 6.1.12, part 1. In order to build the reduction we need to define again a function from profiles of orders to labellings of S. In addition, the reduction has to specify how to map every possible manipulation. That is, to determine the weights on manipulator's examples in the classification instance as a function of the utility scale used by the manipulator in the voting instance. This part of the mapping is denoted by g^*_w, which is a mapping from utilities to weights (will be defined later).

We first build the mapping $g^*_1 : \varphi_1 \to \Upsilon_1$ in the following way:

The table specifies only the number of negative examples in each location (note that it is always a majority, i.e. in each location there are more negative examples than positive ones. This will be important later in the proof).

	$\mathbf{Y}_1 = g^*_1(P_1)$					
P_1	x	y	z	$risk_1(c_X)$	$risk_1(c_Y)$	$risk_1(c_Z)$
$c_X \succ c_Y \succ c_Z$	2k+1	7	6	13	$2k+7$	$2k+9$
$c_X \succ c_Z \succ c_Y$	2k+1	6	7	13	$2k+9$	$2k+7$
$c_Y \succ c_X \succ c_Z$	k+3	7	4	$(k-2)+7+4 = k+9$	$k+7$	$(k+3)+7+3 = k+13$
$c_Y \succ c_Z \succ c_X$	k+1	7	6	$k+13$	$k+7$	$k+9$
$c_Z \succ c_X \succ c_Y$	k+3	4	7	$k+9$	$k+13$	$k+7$
$c_Z \succ c_Y \succ c_X$	k+1	6	7	$k+13$	$k+9$	$k+7$

g^*_2 is symmetric w.r.t. x, z. Using the table, one may verify that indeed $\mathbf{Y}_i = g^*_i(\mathbf{P}_i)$ fits $\mathbf{P}_i$.

The next stage is to build the mapping from utility scales to weights. Note that only the total weight of negative/positive examples on each location matters, and not the particular weight of each example.

Let $\mathcal{U}$ the set of all possible utility scales on $\mathbf{C}$. We first observe that any utility scale on $\mathbf{C}$ consists of 3 real numbers, thus any U can be written as u_X, u_Y, u_Z. We want to standardize the

utilities. Define $u'_X = \frac{u_X - min(u_X,u_Y,u_Z)}{max(u_X,u_Y,u_Z) - min(u_X,u_Y,u_Z)}$ and likewise u'_Y, u'_Z. Thus U' are all in the range $[0,1]$, and for any 2 distributions $p_1, p_2 \in \Delta(\mathbf{C})$,

$$U'(p_1) < U'(p_2) \iff U(p_1) < U(p_2)$$

Therefore, we may assume w.l.o.g that utilities are all in $[0,1]$ (if $u_X = u_Y = u_Z$ we assume they are all equal 0). Note that $U(p)$ is just another way to write $\mathbb{E}_p[U] = \sum_{c \in \mathbf{C}} p(c) u_c$.

The mapping has to specify $g^*_w(U) = (w_{x+}, w_{x-}, w_{y+}, w_{y-}, w_{z+}, w_{z-})$. The only restrictions are that all terms are non-negative, and sum up to 1. We define the mapping as follows:

- $w_{x-} = w_{y-} = w_{z-} = 0$, i.e. all negative examples get zero weight.
- $\alpha = 3 - u_X - u_Y - u_Z > 0$
- $w_{x+} = \frac{1}{\alpha}(1 - u_X) \geq 0$
- $w_{y+} = \frac{1}{\alpha}(1 - u_Y) \geq 0$
- $w_{z+} = \frac{1}{\alpha}(1 - u_Z) \geq 0$

We can see that if i weighs his examples according to these weights, then $risk_i(c_X, S_i) = w_{x+} + w_{y-} + w_{z-} = 1 - u_X$ and similarly for the other two classifiers.

Just like in the deterministic case, every randomized classification mechanism $\mathcal{M}$ induces a randomized voting rule by combining with g^*.

Finally, the restricted set $\mathcal{S}^* \subset \mathcal{S}$ in this case includes all datasets with labellings that are restricted to those that appear in the table

Lemma 6.1.14. *Denote $f = \mathcal{M} \circ g^*$. if $\mathcal{M}$ is SP, then f is honest.*

Proof. Assume that f is not honest. Then there are $\mathbf{P}, P'_i, U$, such that U fits P_i and

$$U(f(P_{-i}, P'_i)) > U(f(\mathbf{P}))$$

From the definition of f,

$$U(\mathcal{M}(g^*(P_{-i}, P'_i))) > U(\mathcal{M}(g^*(\mathbf{P})))$$

Let $\mathbf{Y} = g^*(\mathbf{P}),\ Y'_i = g^*(P'_i)$. Thus

$$U(\mathcal{M}(Y_{-i}, Y'_i)) > U(\mathcal{M}(\mathbf{Y}))$$

We denote by p_X, p_Y, p_Z the probabilities of classifiers without the manipulation (i.e. $\mathcal{M}(\mathbf{Y})$), and by p'_X, p'_Y, p'_Z the probabilities of classifiers after the manipulation (i.e. $\mathcal{M}(Y_{-i}, Y'_i)$). Re-writing the last equation, we get that

$$p'_X u_X + p'_Y u_Y + p'_Z u_Z > p_X u_X + p_Y u_Y + p_Z u_Z \tag{6.3}$$

This means that the risk also holds:

$$\begin{aligned} risk_i(\mathcal{M}(Y_{-i}, Y'_i)) &= p'_X risk_i(c_X, S_i) + p'_Y risk_i(c_Y, S_i) + p'_Z risk_i(c_Z, S_i) \\ &= p'_X(1-u_X) + p'_Y(1-u_Y) + p'_Z(1-u_Z) = 1 - (p'_X u_X + p'_Y u_Y + p'_Z u_Z) \\ &< 1 - (p_X u_X + p_Y u_Y + p_Z u_Z) = p_X(1-u_X) + p_Y(1-u_Y) + p_Z(1-u_Z) \\ &= p_X risk_i(c_X, S_i) + p_Y risk_i(c_Y, S_i) + p_Z risk_i(c_Z, S_i) = risk_i(\mathcal{M}(\mathbf{Y})) \end{aligned}$$

Where the inequality is due to equation (6.3). Therefore agent i strictly gains (in expectation) by misreporting his labeling, in contradiction to Strategy-proofness of $\mathcal{M}$. □

It is clear that f is a probability mixture of duples and dictatorial voting rules iff $\mathcal{M}$ is a probability mixture of duples and dictatorial classification mechanism (with exactly the same probabilities). It follows directly from lemma 6.1.14 and from Gibbard's theorem, that if $\mathcal{M}$ is SP, then f is honest and therefore f (and $\mathcal{M}$!) must be such a mixture. □

6.2 Some Observations on Mechanisms and Upper Bounds

Let $\mathcal{C}$ be any concept class, and $S = \{S_1, \ldots, S_n\}$ a set of examples such that $\forall m_i = |S_i| \leq k$, i.e. no more than k examples are controlled by a single agent. Recall the Random Dictator mechanism (mechanism 2), which selects agent i w.p. $\frac{m_i}{m}$, then picks the best concept for this agent.

First Shot: Try the RD Mechanism

Conjecture 6.2.1. The Random Dictator Mechanism (mechanism 2) is SP and $(k+1)$-approximating for all concept classes.

A fallacious proof. As n is the number of agents, clearly $nk \geq m$. Let c^* the best concept. denote by $T^* \subseteq I$ the agents whose examples are labeled perfectly by c^*, $T_B = T \setminus T^*$ are the agents that disagree with c^* on at least one example. Denote $t^* = |T^*|$, $T_B = |T_B|$. We observe that

$$r^* = risk(c^*, S) \geq \frac{T_B}{m}.$$

denote by r_i the risk of the agent i best classifier.

$$\begin{aligned}
risk(\mathcal{M}_R, S) =& \frac{1}{m}\sum_{i\leq n} m_i r_i \\
=& \frac{1}{m}\left(\sum_{i\in T^*} m_i r_i + \sum_{i\in T_B} m_i r_i\right) \overset{\#}{=} \frac{1}{m}\left(\sum_{i\in T^*} m_i r^* + \sum_{i\in T_B} m_i r_i\right) \\
=& r^*\frac{1}{m}\sum_{i\in T^*} m_i + \frac{1}{m}\sum_{i\in T_B} m_i r_i \leq r^*\frac{1}{m}\sum_{i\in T} m_i + \frac{1}{m}\sum_{i\in T_B} m_i r_i \\
=& r^* + \frac{1}{m}\sum_{i\in T_B} m_i r_i \leq r^* + \frac{1}{m}\sum_{i\in T_B} k\, r_i \\
=& r^* + \frac{k}{m}\sum_{i\in T_B} r_i \leq r^* + \frac{k}{m}\sum_{i\in T_B} 1 \\
=& r^* + k\frac{T_B}{m} \leq r^* + kr^* = (k+1)r^*
\end{aligned}$$

□

Why is this proof fallacious?

Because we did not formally define what is "the best concept of agent i". Each agent i may have more than one "best concept" (even infinitely many), and we need to define how $\mathcal{M}_R$ selects the one to use, otherwise we may not use the equality $r_i = r^*$ (see # in the proof). In chapter 5, where agents had similar interests, it did not matter, since any arbitrary choice of the classifier would do, as long as it minimized the risk of the selected dictator. What if we assume that $\mathcal{M}_R$ uses one of the "best concepts" of i that minimizes the global risk? that would be enough for the approximation proof. Unfortunately, $\mathcal{M}_R$ is no longer SP if ties are broken in this way.

To observe that, consider any setting where some agent has motivation to lie with the ERM algorithm. Now add a new agent with $100m$ examples that are very easy to classify, i.e. it poses no real constraint on the concept class. The risk of $\mathcal{M}_R$ is composed almost entirely of the "best classifier" of the new agent, but this is simply the ERM of the previous problem.

However, if we assume that agents are never indifferent between concepts (which is a rather strong assumption on the structure of the dataset), then the # equality is correct, and the conjecture becomes true.

Another possible direction is to drop the SP requirement and use the ERM mechanism, hoping the untruthful labels will not divert the result too far from the optimal one. Appendix B discusses the disadvantages of this approach.

We describe an improved mechanism which better handles the tie breaking. Recall that

$S = \langle X, Y \rangle$ where X are the data points and Y are the labels.

A Possible Solution: Breaking Ties Iteratively

For each input point $x \in \mathcal{X}$, assume x has a boolean field, $x.marked$. Intuitively, the field is set to T (for true) when our mechanism determines the final label of x.

Mechanism 9 ITERATIVE RANDOM DICTATOR ($\mathcal{M}_{IRD}$)

Initialize the given concept class as $\mathcal{C}_0 = \mathcal{C}$
Initialize $x.marked \leftarrow F$ for each $x \in \mathcal{X}$
Generate a random permutation that maps iterations to agents $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$
for iteration $t = 1, \ldots, n$ **do**
  Select agent $j = \pi(t)$
  $S_{t,j} \leftarrow \{\langle x, Y_j(x) \rangle \ : \ x \in X_j \wedge \neg x.marked\}$ // Consider all the examples of agent j that are *not* marked at time t
  Let $\tilde{c} \in \text{argmin}_{c \in \mathcal{C}_{t-1}} risk_j(c, S_{t,j})$ // $\tilde{c}$ is an ERM w.r.t. $\mathcal{C}_{t-1}$ and $S_{t,j}$
  Find a concept $\tilde{c} \in \mathcal{C}_{t-1}$ that is ERM w.r.t. $S_{t,j}$ // There is always at least one
  $C_t \leftarrow \{c \in C_{t-1} \ : \ \forall \langle x, y \rangle \in S_{t,j} \, (c(x) = \tilde{c}(x))\}$ // Remove concepts that disagree with $\tilde{c}$ on some example in $S_{t,j}$
  for each input point $x \in \mathcal{X}$ **do**
    if $\forall c, c' \in \mathcal{C}_t \, (c(x) = c'(x))$ **then**
      $x.marked \leftarrow T$
    end if
  end for
end for
Return an arbitrary concept from $\mathcal{C}_n$

Remark 6.2.2. Gibbard [18] noted that the random dictatorship result (i.e. theorem 6.1.11) only applies if voters/agents cannot express indifference between alternatives. As a counterexample he supplied the "serial dictatorship" mechanism, which breaks ties at turns. The IRD mechanism is indeed an example of such serial dictatorship, as an agent can clearly consider 2 or more classifiers as equally good.

We begin by showing that mechanism 9 can be implemented efficiently.

Lemma 6.2.3. *Let $m = |S|$. Let $f_{\mathcal{C}}(m)$ be a polynomial bound on the runtime required to find an ERM on S with respect to the concept class $\mathcal{C}$. Then it is possible to implement the IRD mechanism such that the runtime of the algorithm is polynomial in m.*

Proof. At any stage of the algorithm, we denote by $\tilde{c}(x)$ the label that is assigned to input point $x \in X$ by all the concepts in $\mathcal{C}_{t-1}$. Note that $\tilde{c}(x)$ is well defined for all input points that are

marked at time t, and once it is set, it remains defined and unchanged throughout the execution of the mechanism.

We shall represent $\mathcal{C}_t$ by simply remembering $\langle x, \tilde{c}(x)\rangle$ of all the input points that are marked as classified at time t. Denote this set of examples by $\hat{C}_t$, and note that

- c is in $\mathcal{C}_t$ if and only if for all $x \in \hat{C}_t$, $c(x) = \tilde{c}(x)$.
- $\mathcal{C}_t$ is empty if and only if $\hat{C}_t$ is not separable (by any concept from $\mathcal{C}$).

We need not represent all classified input points, but rather the support vectors that allow us to *determine the label* of each marked input point. Thus, we never explicitly represent X_t, but we (show that we) can check for any input point whether it is marked as classified at time t. On initialization $\hat{C}_0 = \emptyset$. Assume by induction that $\mathcal{C}_{t-1}$ is non-empty (and $\hat{C}_t$ is separable).

To find whether a specific input point $x \in \mathcal{X}$ is marked, we check whether $\hat{C}_{t-1} \cup \{\langle x, y'\rangle\}$ is separable for each label $y' \in \{+, -\}$. Since $\hat{C}_{t-1}$ is separable, separation with both labels is possible if and only if x is not classified at time t (that is, $x.marked = F$). In this way we can identify all of $S_{t,j}$, by applying this method on every input point x from $\langle x, y\rangle \in S_j$.

Now we want to find an ERM for $S_{t,j}$, which is in $\mathcal{C}_{t-1}$. This is equivalent to a concept in $\mathcal{C}$ that classifies all of $\hat{C}_{t-1}$ correctly, and minimizes the empirical error on $S_{t,j}$. To achieve that, we duplicate each example in $\hat{C}_{t-1}$ many times (say, more than $|S_{t,j}|$), such that any ERM is forced to label correctly all the input points of $\mathcal{C}_{t-1}$. Finally we define $\hat{C}_t$ by adding to $\hat{C}_{t-1}$ all the examples in $S_{t,j}$ that are labeled correctly by the ERM $\tilde{c}$ (it is possible that $\hat{C}_t = \hat{C}_{t-1}$). Since all the examples that we added are correctly separated by $\tilde{c}$, and $\tilde{c}$ also labels correctly all of $\hat{C}_{t-1}$, we get that $\hat{C}_t$ is also separable (by $\tilde{c}$), thus completing the induction step.

In every iteration we performed at most $2m$ calls to the function that computes ERM, each time on a dataset whose size is at most m^2. Thus our runtime is $O(2m \cdot f_{\mathcal{C}}(m^2))$, by the assumption on $f_{\mathcal{C}}$, is polynomial in m. □

Lemma 6.2.4. *Mechanism 9 is Strategy-Proof.*

Proof. First note that the order in which agents are selected as dictators is independent of the labeling and thus cannot be affected by it. Consider any agent j that is selected in time t. The examples of j that are not in $S_{t,j}$ are already classified, and whatever j reports their classification will not change. As for the examples in $S_{t,j}$, the mechanism minimizes the error and any lie may only result in sub-optimal results w.r.t these examples (which means it will be less favorable to j). □

We end this section with a conjecture:

Conjecture 6.2.5. There are many concept classes, including some that are in wide use, for which the IRD mechanism guarantees an $O(k)$-approximation ratio.

If the conjecture is true, then it stresses a clear connection between the approximation ratio of Strategy-Proof mechanisms and the number of examples controlled by each agent. It would complete our lower bound theorems with a matching upper bound, resulting in the following result:

a. Randomizing the dictator in a simple, iterative way guarantees Strategy-Proofness and $O(k)$-approximation ratio.

b. A better SP mechanism is out of reach.

In contrast to chapter 5, where a uniform upper bound was proven for all concept classes, we have to test conjecture 6.2.5 w.r.t. the specific characteristics of every concept class.

6.3 Linear Separators

In this section we supply lower approximation bounds for SP learning of Linear Separators in $\mathbb{R}^d$, using reductions to the synthetic scenarios which were described earlier in this chapter. An upper bound for the 1-dimensional case is also shown, due to the IRD mechanism.

A formal Definition

Definition 6.3.1. The concept class of *linear separators* in $\mathbb{R}^d$ is defined as

$$\mathcal{C}_d = \{(w, b) : w \in \mathbb{R}^d, \|w\| = 1, b \in \mathbb{R}\}.$$

Each classifier parts $\mathbb{R}^d$ to a positive and a negative part:

$$\forall c \in \mathcal{C}_d, \forall x \in \mathbb{R}^d \left(c(x) = sign(\langle x, w_c \rangle - b_c)\right),$$

where $sign(0) \equiv +$.

In the LINEAR$_d$ problem, the learning mechanism is presented with a dataset $S = \langle S_1, \ldots, S_n \rangle$ of arbitrary examples from $\mathbb{R}^d \times \{-, +\}$, and should output a classifier $c \in \mathcal{C}_d$ that minimize the global risk. The private and global risk are computed w.r.t. the datasets of all agents, as in equations (3.1),(3.2). We do not assume anything on the structure of the dataset. We denote by k the largest number of examples controlled by a single agent, i.e. $k = max_{i \leq n} m_i$.

Deterministic Lower Bound

Theorem 6.3.2. *For any $d \geq 1$, there is no deterministic mechanism for LINEAR$_d$ that is both SP and β-approximating, for any $\beta = o(\sqrt{m})$.*

Recall scenario DET-SYNTHETIC from section 6.1. We proved that for any mechanism $\mathcal{M}$ the following properties hold in DET-SYNTHETIC :

1. If $\mathcal{M}$ is SP and onto, it is dictatorial.
2. If $\mathcal{M}$ is not onto, then it cannot guaranty any finite approximation ratio.
3. If $\mathcal{M}$ is dictatorial, then it cannot guarantee β-approximation, for any $\beta = o(m)$.

Taking all 3 properties together, we bounded from below the approximation ratio that can be achieved in DET-SYNTHETIC . We now show how to use theorem 6.1.5 to attain a lower bound in the linear separation scenario, i.e. to prove theorem 6.3.2.

Proof of theorem 6.3.2. For any instance S of the original problem in DET-SYNTHETIC , we show how to create a new instance S' of LINEAR$_1$. Both scenarios contain the same number of agents. Whereas originally each agent controls $2k+3$ examples, he controls $k' = n(2k+3)^2$ examples in the new instance.

The total number of examples in the original instance is $m = n(2k+3) = \Theta(nk)$. The total number of new examples is thus $m' = nk' = (n(2k+3))^2 = \Theta((nk)^2) = \Theta(m^2)$. We translate any example originally on x to the location "-1", and each examples on z to the location "1" on the real line.

In addition, the remainder of examples ($k' - (2k+3)$ for each agent) are all placed on "0" with a *positive label*. Figure 6.3 shows the spatial arrangement of the examples in the new instance.

Let $\mathcal{M}$ any SP deterministic classification mechanism for LINEAR$_1$. We consider two cases.

Case 1 There is an instance in which "0" is classified as negative.

Case 2 For any instance, the classifier is picked from the following 3 equivalence classes:

- $w = 1, -1 < b < 0$. Those will classify as negative only examples on x, and thus they are all mapped to the c_X classifier in the synthetic scenario.
- $w = -1, 0 < b < 1$. Those will classify as negative only examples on z, and thus they are all mapped to the c_Z classifier in the synthetic scenario.

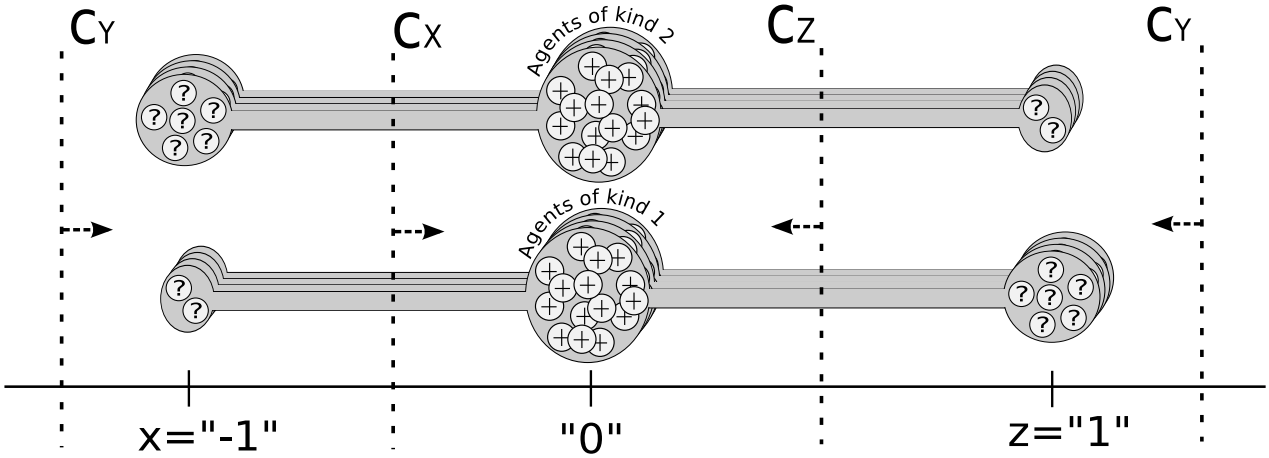


Figure 6.3: The construction of the lower bound example in the LINEAR scenario. The arrows indicate the positive half space of each classifier.

- all other classifiers that face 0. Those will classify *all* examples as positive, and thus they are all mapped to c_Y in the synthetic scenario.

The spatial counterpart of each classifier is shown in figure 6.3. In the first case, there must be an instance in which $\mathcal{M}$ misclassifies all examples on "0" as negative. This results in a risk of (at least) $m' - n(2k+3)$ whereas any other classifier achieves a risk of (at most) $n(2k+3)$. The approximation ratio in this case is thus

$$\Omega(\frac{m'}{nk}) = \Omega(\frac{m'}{m}) = \Omega(\frac{m'}{\sqrt{m'}}) = \Omega(\sqrt{m'}).$$

In the second case, let $\beta' = o(\sqrt{m'})$ and assume that $\mathcal{M}$ indeed guarantees an approximation ratio of β'. Let any instance S of the original problem in DET-SYNTHETIC , and let a denote the minimal number of mistakes in it, i.e. $a = m \cdot r^*$. The minimal number of mistakes in the new instance is also a, by using any of the linear classifiers in $\mathcal{C}_1$ corresponding to $c^* \in \mathcal{C}_s$. From our assumption, $\mathcal{M}$ does not make more than $\beta' a$ mistakes on the new instance. The corresponding classifier from $\mathcal{C}$ makes the same number of mistakes, providing us with an approximation ratio of $\beta' = o(\sqrt{m'}) = o(m)$ to the original problem.

Thus, if m is the number of examples, then the existence of any SP mechanism with approximation ratio of $o(\sqrt{m})$ for $\mathcal{C}_1$ will also provide us with a $o(m)$-approximating mechanism for the synthetic scenario, in contradiction to theorem 6.1.5.

For linear separators in higher dimensions $d > 1$, we apply the following natural reduction from LINEAR$_1$. We use the same setting, with $x =$"$(-1, 0, 0, \ldots)$", $z =$"$(1, 0, 0, \ldots)$". Thus all examples are on one axis, and for any classifier we may consider only its projection on this

axis. This provides us with a full reduction from $\mathcal{C}_1$, thus any solution for a higher dimension is necessarily also a solution for the 1-dimensional problem. Clearly the approximation ratio for the multi-dimensional problem cannot be $o(\sqrt{m})$ as well. □

What about an upper bound?

As the lower bound is of the same order as the total number of examples, there is clearly no sense in using any deterministic SP mechanism. A better approach must involve randomization.

Randomized Lower Bound

Denote by k the maximal number of examples controlled by a single agent in S, i.e. $k = max_{i \in I}|S_i|$. We would have liked to prove the following result for randomized mechanisms, similarly to theorem 6.3.2:

Conjecture 6.3.3. For $d \geq 1$, any SP randomized mechanism for learning LINEAR$_d$ has an approximation ratio of $\Omega(k)$.

Such a result is currently not at hand. Nevertheless, if we restrict the possible classification mechanisms allowed, the conjecture is true:

Theorem 6.3.4. *Let $\mathcal{M}$ a randomized classification mechanism for learning LINEAR$_d$. If $\mathcal{M}$ is a probability mixture of duples and dictatorial mechanisms, then it has an approximation ratio of $\Omega(k)$.*

The theorem (and also the proof) is very similar to the second part of theorem 6.1.12. We bring the proof for completeness:

Proof. If a duple is selected with positive probability, then there is at least one classifier $c' \in \mathcal{C}_d$ s.t. for any dataset, $p_{\mathcal{M}}(c') < 1$. If the dataset is labeled using c', the expected risk will be higher than 0, and thus the approximation ratio is infinite.

Conversely, consider the following setting with 2 agents: one agent holds 1 negative example on $x = (1, 1, 1, \ldots)$ and $k - 1$ positive examples on $z = (-1, -1, \ldots)$. The other agent holds 1 negative example on z, and $k - 1$ positive examples on x. clearly $r^* = 2/2k = \frac{1}{k}$, but the classifier selected by any of the dictators will misclassify *all* of the other agents examples. Thus it will have a risk of $\frac{1}{2} = \Omega(k)r^*$. □

Recall that in scenario RAND-SYNTHETIC we completed the lower bound proof, under the assumption that agents are allowed to hold private weights for their examples. Under the

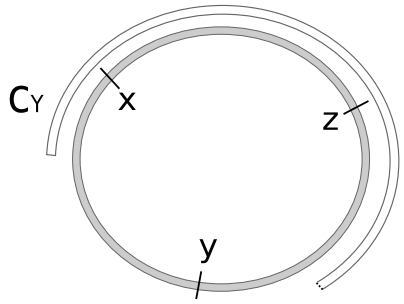


Figure 6.4: The constructed instance of INTERVAL. The interval shown classifies as negative only the examples on y, and is therefore mapped to c_Y.

same assumption, we can show a $\Omega(k)$ lower bound in a spatial setting which is very close to linear separators.

Let $\mathcal{X}$ be a circle of length 1. The class of possible classifiers contains all intervals of length $\frac{2}{3}$ on the circle. Intervals are open from one side and closed from the other, and classify as positive only examples inside the interval. We refer to this interval classification problem as INTERVAL.

Theorem 6.3.5. *Assume agents attribute weights for their examples (i.e. the risk is computed using equation* (6.2)*). Any randomized SP learning mechanism for INTERVAL has an approximation ratio of* $\Omega(k)$.

Proof. Given an instance in scenario RAND-SYNTHETIC , we build a reduction by placing the locations x, y, z in equal distances along the circle. Thus any classifier classifies exactly 1 location as negative, and the other two as positive. Any SP β-approximating classification mechanism for INTERVAL is also SP and β-approximating for RAND-SYNTHETIC (if weights are allowed). As a result, any SP $o(k)$-approximation mechanism for INTERVAL is in contradiction to theorem 6.1.13. □

Moreover, even without weights, if the randomized classification mechanism is restricted to mixtures of duples and dictatorial mechanisms, then it cannot achieve any approximation ratio better than $\Omega(k)$ (from the second part of theorem 6.1.12).

A natural question is "is this bound tight?". We next turn to show that it is possible to achieve an approximation bound close to k, using the IRD mechanism described in section 6.2.

Upper Bound for LINEAR$_1$

The following proof is a first step in the direction of conjecture 6.2.5. We show, via careful analysis of the IRD mechanism steps, that the number of data points controlled by each agent is

the factor which determines the number of errors. Apart from providing us with a first positive result for an important concept class, we believe that similar proof techniques may be used to verify our conjecture for other concept classes.

Theorem 6.3.6. *$\mathcal{M}_{IR}$ is an SP $O(k^2)$-approximation mechanism for LINEAR$_1$, where k is the size of the largest partial dataset.*

The outline of the proof is as follows: Let c^* any optimal classifier. We consider each agent as "good" if he completely agrees with c^*, and otherwise as "bad". In every iteration a single agent sets the final labels for his examples, but also enforces a label on other examples in the process. If the agent is good, then all the labels set in this iteration must agree with c^* as well, but if this agent is bad then some examples might be "ruined" in this iteration, i.e. labeled not according to c^*. The heart of the proof is bounding the expected number of examples that are ruined in each iteration, showing that this number decreases exponentially fast. This is since in every iteration the number of non-labeled examples is a fraction of the examples that had not been labeled until the previous iteration. The detailed proof can be found in appendix C.4. We believe that this analysis is not tight, and conjecture that the real bound is $O(k)$.

Chapter 7

Discussion

7.1 Summary of our Results

Decision Making Setting

On the positive side, we presented a simple randomized mechanism, namely the Weighted Random Dictator mechanism. We showed that it is Strategy-Proof (SP), and guarantees a 3-approximation bound on the risk for any concept class, under the assumption that all agents control the same set of data points (the SIMILAR$_\mathcal{C}$ problem). This bound is further improved to $3-\frac{2}{n}$ in the unweighted version of the problem, suggesting that it is better to use a *small* number of agents for the labeling procedure. We showed that in the general case, no SP deterministic mechanism can guarantee a constant approximation ratio, and that the trivial selection of the heaviest agent as a dictator is the best SP mechanism at hand.

In contrast, in the special case where there are only two possible decisions (the PLUS-MINUS problem), the heaviest dictator mechanism guarantees a 3-approximation, and this bound is tight. Also, the randomized bound can be improved to 2 using a non-trivial randomization, and this bound is also tight. All these results are summarized in table 7.1.

On the somewhat negative side, we demonstrated that if all assumptions on the structure of the dataset are dropped, then deterministic SP mechanisms are utterly useless. Even with randomized mechanisms, under slight alterations of the model, the approximation ratio is lower bounded by the size of the largest partial dataset controlled by a single agent. We further showed how these results carry from synthetic problems to the commonly used concept class of linear separators, and matched this randomized lower bound with an (untight) upper bound for the 1-dimensional case. These results suggest that when no structure can be assumed on the dataset, then it is better to use *many agents*, each controlling a small number of examples. Results are summarized in table 7.2.

	All Classes			PLUS-MINUS	
	Deterministic	Randomized	Rand., with Equal Weights	Det.	Rand.
Upper Bound	$O(n)$	3	$3-\frac{2}{n}$	3	2
Lower Bound	$\Omega(n)$	2_+	2_+	3	2

Table 7.1: Summary of results, when agents have similar interests. n is the number of agents. A $-/+$ subscript means that we conjecture the real bound is lower/higher, respectively.

	All Classes		$\textbf{LINEAR}_d$	
	Deterministic	Randomized	Det.	Rand. (1 dim)
Upper Bound	$O(m)$	$O(m)_-$	$O(m)$	$O(k^2)_-$
Lower Bound	$\Omega(m)$	$\Omega(k)$	$\Omega(\sqrt{m})_+$	$\Omega(k)_*$

Table 7.2: Summary of results, with arbitrary datasets. m is the total number of examples, k is the size of the largest dataset controlled by a single agent. $*$ means that this bound holds under some modifications of the model.

Straight-forwardness and Dictatorship

As mentioned, the PLUS-MINUS scenario is also a special case of classification with *arbitrary* datasets, where $|\mathcal{C}| = 2$. Interestingly, there is a striking, non-gradual leap in hardness when we move from this simple concept class to more complex ones. This difference can be partially explained by the notion of *straight-forwardness*. In the PLUS-MINUS problem, although truth-telling is not always dominant, each agent has a straight-forward (that is, *dominant*) strategy. By this we mean that the agent can play optimally based only on his private data, ignoring other agents' actions.[1] Our mechanisms exploit this fact by casting the action of each player to the optimal one, thus rendering any manipulation ineffective.

In contrast, it has been shown by Gibbard [19], that introducing a third candidate eliminates all straight-forward strategies in non-dictatorial voting rules. A similar thing occurs in the SP classification domain, and we are left only with dictatorial mechanisms. This point is further pressed in appendix B.

Machine Learning 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

[1] This is simply by labeling all of his examples according to the majority in his private dataset.

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 - r_{min}. In the general case of agents with similar interests (the SIMILAR$_{\mathcal{C}}$ problem) we bounded the risk by $3r_{min} + \epsilon$, and in the PLUS-MINUS problem we showed a better, $2r_{min} + \epsilon$ upper bound.

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

Non-constant Bounds

What about the LINEAR$_1$ problem and other scenarios where the the risk is bounded by a function of the input size, rather than a constant? Is it possible to use similar techniques to design a β-approximation learning mechanism for a concept class, where β is polynomial in k? As we sample more data points to drive ϵ lower and improve the bound, k will increase - and the bound will only get worse! Therefore, the generalization techniques we presented can work efficiently only if the original upper bound does not depend on sample size.

7.2 Related Work

Our work is closely related to the work of Dekel, Fischer and Procaccia [11]. They also investigated game-theoretic aspects of supervised machine learning, albeit in a *regression learning* setting. This work may be seen as an extension of theirs to the world of classification. Specifically, in their setting the label of each data point is a real number, and the risk of some hypothesis is the total *distance* to the correct labels. A significant part of their work concentrated on a setting where each agent only controls one point: they showed that no agent can benefit by lying under certain assumptions; this result holds trivially in our present classification setting. Notably, some of the bounds in our chapter 5 resemble the bounds in the regression setting. Moreover, similar intuition sometime accounts for both settings, although it is not clear if there is a way to derive the results in one setting from the other.

Perote and Perote-Peña [28] proposed a somewhat different model of linear regression in strategic setting, and introduced a family of mechanisms that are SP in this model. In contrast to [11] and to our work, they did not supply analytical bounds on the error rate. Rather, they showed via simulations that their SP mechanisms do perform better than the ERM[2] under some

[2]The ERM in their regression model is the Least Squares estimator.

complex assumptions on agents' behavior (these are necessary, as there is no nash equilibrium when ERM is used).

Another highly related work by Perote-Peña and Perote [29] has a negative flavor. They put forward a model of an unsupervised *clustering* problem, 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 is *always* a manipulation. I.e. there are no (deterministic) reasonable clustering mechanisms, which are SP. This strong negative result is close in spirit to those presented in section 6.1, which say that without further assumptions, there are no reasonable SP classification mechanisms as well.

There has been much interest in SP and GSP mechanisms in fields that are not normally considered as machine learning, but use models that much resemble ours. An example is the problem of selecting a candidate from some continuous set, subject to requirements such as efficiency. When the candidates are the real line, and agents' preferences are single peaked,[3] all deterministic and randomized SP mechanisms were characterized by Moulin [24; 25] and Ehlers et al. [15]. A stronger result by Dutta et al. [14] shows that in the multi-dimensional case, only dictatorial mechanisms are SP. These problems are somewhat related to that of finding spatial classifiers (such as linear separators in $\mathbb{R}^d$), and their definition of Strategy-Proofness is much closer to our definition than Gibbard's [18]. While the desired properties in the continuous decision model are qualitative (unlike the risk, which is quantitative), the connections between the models is still interesting.

Other existing works are related in a more indirect way, studying issues associated with both machine learning and game theory. Some employ machine learning as means to construct mechanisms with desired game-theoretical properties (e.g. Balcan et al. [2], Procaccia et al. [31; 30], and Kalai [20]). Several works attempt to learn in the face of malicious noise (e.g. Bshouty et al. [7], Kearns and Li [21]), but still lack the notion of incentives. Perhaps closer to our work is the paper of Dalvi et al. [10], who model classification as a game between a classifier and an adversary. This setting is relevant, for example, when a spammer is attempting to fool an anti-spam filter. Dalvi et al. examine the optimal strategies of the classifier and adversary, given their strategic considerations. Similarly, Barreno et al. [5] and Lowd and Meek [22] look at different models where machine learning algorithms are the target of attack by a malicious adversary; they discuss the different types of attacks on different concept classes, and possible defenses. In contrast to all these works (but similarly to [11]), our work concentrates on designing strategy-proof mechanisms, i.e. mechanisms that discourage agents from acting strategically in the first place, and not mechanism which work well in spite of such behavior.

[3] More accurately, the model assumes additional weak restrictions on the mechanisms, such as unanimity.

In this sense, our work has something in common with the fast growing body of work on algorithmic mechanism design (e.g. Conitzer and Sandholm [8; 9], Nisan and Ronen [27]). However, in algorithmic mechanism design approximation and randomization are employed as ways to circumvent computational complexity while achieving strategy-proofness. In contrast, in our work computational complexity is not an issue, but approximation nevertheless helps us to design strategy-proof mechanisms.

7.3 Conclusions

The tight coupling we showed between Strategy-Proofness and dictatorial mechanisms entails that without further assumptions or restrictions, SP classification is not possible in effect. We also supplied sufficient conditions that enable SP mechanisms with a bounded error.

Our mechanisms can serve human and automated decision makers that wish to maximize social welfare in the face of data that is biased by conflicting interests. Crucially, our results in the learning theoretic setting constitute first steps in designing classifiers that can function well in non-cooperative environments.

Future research may provide answers to some of the questions we laid 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 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, like misreporting the *location* of the data points rather than their labels.

All of these may reveal new parts of the whole picture and promote a better understanding of the conditions under which strategy-proof learning can occur effectively. This, in turn, might supply us with new insights on our results in the field and on its relations with other areas.

Appendix A

Sample Complexity

In section 2.2 we stated that for certain concept classes, the empirical risk is close to the real risk, provided that enough samples are take. We later used this statement to justify the approximation of the ERM as a proxy to the actual risk minimizer. However, we left open two important questions:

1. How many samples are sufficient?
2. What is a "good" concept class?

We devote this appendix to formally answer the two questions, showing how they are related to each other. A seminal result by Vapnik and Chervonenkis shows the following:

Theorem A.0.1 (Vapnik and Chervonenkis, '71 [38])**.** *Let m s.t.*

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

Then with probability of at least $1-\delta$,

$$\forall c \in \mathcal{C}\left(|risk(c) - \widehat{risk}(c,S)| < \epsilon\right). \tag{A.1}$$

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.*

While V_C may be very large, or even infinite in some cases[1], it is known to be finite for some commonly used concept classes (e.g. linear classifiers).

[1] V_C is known as the *VC-dimension* of $\mathcal{C}$, introduced in [38]. 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, [12] being just one of them.

This demand on m also supplies us with an answer to the second question: a concept class is good from machine learning point-of-view if on one hand it contains good classifiers (with low risk), and on the other hand V_C is low. The tradeoff between these two properties of $\mathcal{C}$ is widely known as the *Approximation-Estimation Tradeoff*.

Proof of theorem 2.2.5

Suppose that we have some (possibly randomized) algorithm which returns a concept $c_{\mathcal{A}}$, guaranteeing

$$\forall S \in \mathcal{S}\left(\mathbb{E}_{\mathcal{A}}\left[\widehat{risk}(c_{\mathcal{A}}(S), S)\right] \leq \alpha \cdot r^*\right).$$

Along with equation (A.1), we get that w.p. at least $1-\delta$

$$\begin{aligned}
\mathbb{E}_{\mathcal{A}}\left[risk(c_{\mathcal{A}})\right] <& \mathbb{E}_{\mathcal{A}}\left[\widehat{risk}(c_{\mathcal{A}}, S) + \epsilon\right] = \mathbb{E}_{\mathcal{A}}\left[\widehat{risk}(c_{\mathcal{A}}, S)\right] + \epsilon \\
\leq& \mathbb{E}_{\mathcal{A}}\left[\alpha \cdot \widehat{risk}(c^*, S)\right] + \epsilon \\
\leq& \mathbb{E}_{\mathcal{A}}\left[\alpha \cdot \widehat{risk}(c_{min}, S)\right] + \epsilon = \alpha \cdot \widehat{risk}(c_{min}, S) + (\alpha+1)\epsilon \\
<& \alpha \cdot risk(c_{min}) + \epsilon + (\alpha+1)\epsilon = \alpha \cdot risk(c_{min}) + (\alpha+2)\epsilon
\end{aligned}$$

If S is a "bad" dataset (occurs w.p. $< \delta$) then the risk is at most 1. thus

$$\mathbb{E}_S\left[\mathbb{E}_{\mathcal{A}}\left[risk(c_{\mathcal{A}})\right]\right] < \delta + (1-\delta)(\alpha \cdot risk(c_{min}) + (\alpha+2)\epsilon) < \alpha \cdot risk(c_{min}) + (\alpha+2)\epsilon + \delta.$$

Finally, Let ϵ' s.t. $\epsilon = \delta = \frac{\epsilon'}{\alpha+3}$. It follows from theorem A.0.1 that if $m > \frac{(\alpha+3)^2 V_C}{\epsilon'^2} log\left(\frac{(\alpha+3)^3 V_C}{\epsilon'^3}\right)$, then

$$\mathbb{E}_S\left[\mathbb{E}_{\mathcal{A}}\left[risk(c_{\mathcal{A}})\right]\right] < \alpha \cdot risk(c_{min}) + \epsilon'. \tag{A.2}$$

Note that

$$\frac{(\alpha+3)^2 V_C}{\epsilon'^2} log\left(\frac{(\alpha+3)^3 V_C}{\epsilon'^3}\right) = \frac{(\alpha+3)^2 V_C}{\epsilon'^2} 3log\left(\frac{(\alpha+3)V_C}{\epsilon'}\right) = \Theta\left(\frac{\alpha^2}{\epsilon^2} log\left(\frac{\alpha}{\epsilon}\right)\right)$$

□

Appendix B

How Bad is ERM?

ERM in the PLUS-MINUS scenario

Although the ERM mechanism is not SP, we can still use it the PLUS-MINUS scenario and guarantee a constant approximation, provided that agents only lie in a way which decreases their private risk.

Why is that true? Since any positive agent will only lie about his negative labels and vice versa - regardless of what any other agent does. Therefore if we denote by $\hat{S} = \langle \hat{P}, \hat{N} \rangle$ the dataset with the *reported* labels, then $P \leq \hat{P} \leq P'$ and $N \leq \hat{N} \leq N'$. From here, we can replace P', N' with $\hat{P}, \hat{N}$ in the proof of theorem 4.1.2 to obtain the following result:

$$risk_I(\hat{c}, S) \leq 3 \cdot r^*, \tag{B.1}$$

where $\hat{c} = c^*(\hat{S}) = ERM(\hat{S})$ is the concept returned by the ERM mechanism. Thus ERM is 3-approximating w.r.t. the real data, *despite* the fact that it is not SP.

Similar arguments show that if we use the randomization of mechanism 3 on the reported labels $\hat{P}, \hat{N}$ (instead of computing P', N'), then the new mechanism is 2-approximating even if agents lie.

As this may suggest that pursuing strategy-proofness may not be the best approach, we will see in the next section that this result is limited to the PLUS-MINUS problem.

ERM in General Settings

Conjecture B.0.2. Let $\mathcal{C}$ any concept class, and S a dataset with real labels. Assume that an agent only lies if he benefits from it, and denote by $\hat{S}$ the same dataset containing the reported labels.

$$risk_I(ERM(\hat{S}), S) \leq O(k) r^*.$$

A fallacious proof. Let $c^* = c^*(S)$ any ERM of the real dataset S, and denote by $\hat{c} = c^*(\hat{S})$ the result of the ERM mechanism. Denote by $I_B \subseteq I$ the set of all agents that do not fully agree with c^*, i.e. $I_B = \{i \in I \text{ s.t. } risk_i(\hat{c}, S) > 0\}$. We denote by $B \subseteq S$ (these are possibly "bad" examples) all examples of agents in I_B. Similarly, the "good" examples are $G = S \setminus B$. Let $b = |B|$ and $g = |G|$, clearly $b + g = m$. Only bad agents have incentive to lie, thus $\hat{G} = G$, but $\hat{B}, B$ may differ.

Note that the global risk bounds the number of bad agents:

$$mrisk_I(c^*, S) = \sum_{i \in I_B} m_i risk_i(c^*, S) + \sum_{i \in I_G} m_i risk_i(c^*, S) = \sum_{i \in I_B} m_i risk_i(c^*, S) \geq |I_B|,$$

and since each agent controls at most k examples,

$$\frac{b}{m} = \frac{|B|}{m} \leq k\frac{|I_B|}{m} \leq krisk(c^*, S) = k \cdot r^*. \tag{B.2}$$

Also, $risk(\hat{c}, \hat{S}) \leq risk(c^*, \hat{S})$, which means

$$\begin{aligned} \frac{g}{m}risk(\hat{c}, \hat{G}) + \frac{b}{m}risk(\hat{c}, \hat{B}) \leq \frac{g}{m}risk(c^*, \hat{G}) + \frac{b}{m}risk(c^*, \hat{B}) = \frac{b}{m}risk(c^*, \hat{B}) \Rightarrow \\ \frac{g}{m}risk(\hat{c}, \hat{G}) \leq \frac{b}{m}\left(risk(c^*, \hat{B}) - risk(\hat{c}, \hat{B})\right) \end{aligned}$$

Finally,

$$\begin{aligned} risk(\hat{c}, S) =& \frac{g}{m}risk(\hat{c}, G) + \frac{b}{m}risk(\hat{c}, B) \\ \leq& \frac{b}{m}\left(risk(c^*, \hat{B}) - risk(\hat{c}, \hat{B})\right) + \frac{b}{m}risk(\hat{c}, B) \\ \leq& 2\frac{b}{m} && \text{(since } risk \leq 1\text{)} \\ \leq& 2k \cdot r^* && \text{(from eq. (B.2))} \end{aligned}$$

□

It seems as if the ERM really works, but the "proof" collapses if our game-theoretic assumptions are carefully read. We stated that agents do not lie unless they strictly gain by lying, but whether or not a lie is beneficial depends on the labels reported by other agents. Thus, the statement that only agents in I_B lie holds only as long as *there is only one lier*. Once some agents changed their reported labels, the result of the mechanism changes, forming a *new* set of agents that disagree with it (I'_B). These agents in turn will now have a motivation to lie, and so on in an endless loop of strategic actions and re-evaluations. Therefore, the basic statement that

$\hat{G} = G$ is too naïve, and the whole proof collapses once it is dropped.

Why did the same assumption work well in the PLUS-MINUS scenario? Because when ERM is used in that simple case, each agent always has a *dominant strategy*, which does not depend on the behavior of other agents. Therefore there is never a reason for a second iteration of lies, and the result remains stable and can be analyzed. In contrast, dominant strategies are *not* available when more complex concept classes are used. Recall theorems 6.1.2,6.1.11 which state that under certain conditions truth-telling is not dominant. Another remarkable result by Gibbard [19] further shows that under the same conditions there are no dominant strategies at all! This result accounts for our assertion in the last paragraph that agents in I_G might lie as well.

Appendix C

Proofs

C.1 Proof of Theorem 4.3.2

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

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

Without loss of generality we assume that $r^* = risk(c_-) < risk(c_+)$. Notice that if $r^* = risk(c^*) = risk(c_-) > \frac{1}{2} - 3\epsilon$ 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

$$risk(c_-) + 3\epsilon \leq \frac{1}{2} \leq risk(c_+) - 3\epsilon. \tag{C.1}$$

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:

$$\forall c \in \{c_+, c_-\}, \forall i \in I,\ |\widehat{risk}_i(c, S_i) - risk_i(c)| < \epsilon. \tag{C.2}$$

Lemma C.1.1. *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 $i \in I$. Let $e(x)$ be the indicator random variable of the event $Y_i(x) = +$. We can now rewrite the empirical and real risk as the sum and the expectation of $e(x)$:

$$risk_i(c_-) = \mathbb{E}_{x\sim\mathcal{D}_i}[e(x)]$$

$$\widehat{risk}_i(c_-, S_i) = \frac{1}{m}\sum_{x \in S_i} e(x)$$

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 $risk_i(c_-)$. We derive from the Chernoff bound that for any data set of size $|S_i| = m$:

$$\Pr(|\widehat{risk}_i(c_-, S_i) - risk_i(c_-)| > \epsilon) < 2e^{-2\epsilon^2 m}$$

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

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

where the first inequality is due to the union bound.

Note that it is enough to show the above for c_- since

$$|\widehat{risk}_i(c_-, S_i) - risk_i(c_-)| = |\widehat{risk}_i(c_+, S_i) - risk_i(c_+)|.$$

□

If G occurs, then from (C.2) and the triangle inequality it holds that for all $c \in \{c_+, c_-\}$ and $i \in I$,

$$|risk(c) - \widehat{risk}(c, S)| \leq \sum_{i \in I} \frac{1}{n}|risk_i(c) - \widehat{risk}_i(c, S)| \leq \epsilon. \tag{C.3}$$

Using (C.3) we could have bounded the risk of $\mathcal{M}_{R2}(S)$, but unfortunately this would not do as the mechanism may only access $\overline{S}$ and not S. In order to bound $risk(\mathcal{M}_{R2}(\overline{S}))$, 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 $\mathcal{M}_{R2}$ on $\tilde{S}$), and $\tilde{S}, S$ are very similar.

We now divide I into two types of agents:

$$I' = \{i \in I :\ |risk_i(c_-) - \frac{1}{2}| < \epsilon\},$$

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 $P_i = m\widehat{risk}_i(c_-, S_i)$. Since $risk(c_-) < risk(c_+)$ 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 3 (which is used by mechanism 4 in the last step) 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 C.1.2. *If G occurs, then for all agents in I*

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

Proof. If $i \in I''$ then w.l.o.g. $risk_i(c_-) \leq risk_i(c_+) - 2\epsilon$, thus from (C.2)

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

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

$$risk_i(c_+) \leq risk_i(c_-) \leq risk_i(c_+) + 2\epsilon.$$

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

$$\begin{aligned}\tilde{P}_i \geq P_i + m\epsilon \; &= m\hat{risk}_i(c_-, S_i) + m\epsilon \geq mrisk_i(c_-) \\ &\geq mrisk_i(c_+) \geq m(\widehat{risk}_i(c_+, S_i) - \epsilon) \geq N_i - m\epsilon \geq \tilde{N}_i\end{aligned}$$

□

Lemma C.1.2 implies that, if G occurs, agents cannot do better than report $\tilde{S}$ under Mechanism 4, 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 $\mathcal{M}_{R2}$ only increases with the number of positive examples (the probability of the mechanism choosing the positive classifier increases),

$$risk(\mathcal{M}_{R2}(S)) \leq risk(\mathcal{M}_{R2}(\overline{S})) \leq risk(\mathcal{M}_{R2}(\tilde{S})). \tag{C.4}$$

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

Lemma C.1.3. *If G occurs,*

$$\forall c \in \{c_+, c_-\},\ |risk(c) - \widehat{risk}(c, \tilde{S})| \leq 3\epsilon. \tag{C.5}$$

As in Lemma C.1.1, it will suffice to show this only for c_-.

Proof. From (C.2), for $m > \frac{1}{\epsilon}$,

$$\begin{aligned}
\widehat{risk}(c_-, \tilde{S}) = \frac{\tilde{P}_i}{m} = \frac{P_i + \lceil m\epsilon \rceil}{m} &\leq \frac{P_i + m\epsilon + 1}{m} \\
&\leq \frac{P_i}{m} + 2\epsilon = \widehat{risk}(c_-, S) + 2\epsilon \\
&\leq risk(c_-) + \epsilon + 2\epsilon = risk(c_-) + 3\epsilon
\end{aligned}$$

□

From (C.1) and (C.5)

$$\widehat{risk}(c_-, \tilde{S}) \leq risk(c_-) + 3\epsilon \leq risk(c_+) - 3\epsilon \leq \widehat{risk}(c_+, \tilde{S}) \tag{C.6}$$

So c_- is also empirically the best concept for $\tilde{S}$; Mechanism 3 guarantees:

$$\widehat{risk}(\mathcal{M}_{R2}(\tilde{S}), \tilde{S}) \leq 2\widehat{risk}(c_-, \tilde{S}) \tag{C.7}$$

Furthermore, since the risk of Mechanism 4 is a convex combination of the risk of c_+, c_-, we get from (C.5),

$$risk(\mathcal{M}_{R2}(\tilde{S})) \leq \widehat{risk}(\mathcal{M}_{R2}(\tilde{S}), \tilde{S}) + 3\epsilon \tag{C.8}$$

Finally, by using (C.4), (C.8), (C.7) and (C.6) in this order, we get that if G occurs:

$$\begin{aligned} risk(\mathcal{M}_{R2}(\overline{S})) &\leq risk(\mathcal{M}_{R2}(\tilde{S})) \leq \widehat{risk}(\mathcal{M}_{R2}(\tilde{S}), \tilde{S}) + 3\epsilon \\ &\leq 2\widehat{risk}(c_-, \tilde{S}) + 3\epsilon \\ &\leq 2(risk(c_-) + 3\epsilon) + 3\epsilon = 2r^* + 9\epsilon \end{aligned} \tag{C.9}$$

From the definition of the mechanism, $risk(\tilde{\mathcal{M}}) = \mathbb{E}_S\left[risk(\mathcal{M}_{R2}(\overline{S}))\right]$. If G does not occur, the risk cannot exceed 1. Thus by applying Lemma C.1.1 with $\delta = \epsilon = \frac{\epsilon'}{10}$ we find that for $m > 50\frac{1}{\epsilon'^2}\ln(\frac{10n}{\epsilon'})$:

$$\begin{aligned} risk(\tilde{\mathcal{M}}) = \mathbb{E}_S\left[risk(\mathcal{M}_{R2}(\overline{S}))\right] &\leq \Pr(G)\mathbb{E}_S\left[risk(\mathcal{M}_{R2}(\overline{S}))|G\right] + \Pr(\neg G)1 \\ &\leq (2r^* + 9\epsilon) + \Pr(\neg G)1 && \text{(from eq.(C.9))} \\ &\leq 2r^* + 9\epsilon + \epsilon \\ &\leq 2r^* + \epsilon' \end{aligned}$$

□

C.2 proof of theorems 5.1.2, 5.1.1

The more complicated proof is of regards the non-deterministic case, and is brought first. Using the same notations and lemmas, the deterministic case is given a short proof in the last section.

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

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

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' preference are independent, thus for every two agents $i \neq j$,

$$\forall x \in \mathcal{X}\forall y, y' \in \mathcal{Y}\left(\Pr(f_i(x) = y, f_j(x) = y'|x) = \Pr(f_i(x) = y|x)\Pr(f_j(x) = y'|x)\right). \tag{C.10}$$

Definition C.2.1. 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:

$$d(f, f') = d_{F_X}(f, f') = \mathbb{E}_{x \sim F_X} \left[\Pr(f(x) \neq f'(x)|x) \right]. \tag{C.11}$$

Let $\mathcal{C} \subseteq \mathcal{H}$ any concept class.

Lemma C.2.2. $\forall c \in \mathcal{C}, \forall j \in I$,

$$d(f_j, c) = risk_j(c, F). \tag{C.12}$$

Proof.

$$\begin{aligned}
risk_j(c, F) \equiv & \mathbb{E}_{\langle x,y \rangle \sim F_j} \left[[\![c(x) \neq y]\!] \right] = \mathbb{E}_{x \sim F_X} \left[\sum_{y \in \{-,+\}} \Pr_{y \sim F_j}(y|x) [\![c(x) \neq y]\!] \right] \\
= & \mathbb{E}_{F_X} \left[\underline{f}_j(x) [\![c(x) \neq -]\!] + \overline{f}_j(x) [\![c(x) \neq +]\!] \right] \\
= & \mathbb{E}_{F_X} \left[\Pr(f_j(x) = -|x) [\![c(x) \neq -]\!] + \Pr(f_j(x) = +|x) [\![c(x) \neq +]\!] \right] \\
= & \mathbb{E}_{F_X} \left[\Pr(f_j(x) = -, c(x) = +|x) + \Pr(f_j(x) = +, c(x) = -|x) \right] \\
= & \mathbb{E}_{F_X} \left[\Pr(f_j(x) \neq c(x)|x) \right] = d(c, f_j) \qquad \text{(from eq.(C.11))}
\end{aligned}$$

□

Recall that $c_i = argmin_{c \in \mathcal{C}} risk_i(c, F)$ and $c^* = argmin_{c \in \mathcal{C}} risk_I(c, F)$.
As a special case of equasion C.12, we get that

$$\forall i, j \left(d(c_i, f_j) = risk_j(c_i, F) \right). \tag{C.13}$$

We will also need the following assertions:

Lemma C.2.3. *d is reflexive, non-negative, symmetric and holds triangle inequality.*

Lemma C.2.4.

$$\forall i \in I \left(c_i = argmin_{c \in \mathcal{C}} d(c, f_i) \right)$$

Lemma C.2.5.

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

Lemma C.2.6.

$$2r^* \geq \sum_i \sum_j w_i w_j d(f_i, f_j)$$

From the lemmas, the theorem follows:

$$\begin{aligned}
\sum_{i\in I} w_i risk_I(c_i, F) =& \sum_i \sum_j w_i w_j d(f_i, c_j) \\
\leq& \sum_i \sum_j w_i w_j (d(f_i, f_j) + d(f_j, c_j)) && \text{(Triangle Inequality)} \\
\leq& \sum_i \sum_j w_i w_j (d(f_i, f_j) + d(f_j, c^*)) && \text{(lemma C.2.4)} \\
=& \sum_i \sum_j w_i w_j d(f_i, f_j) + \sum_j w_j d(f_j, c^*) \sum_i w_i \\
\leq& 2r^* + \sum_j w_j d(f_j, c^*) && \text{(lemma C.2.6)} \\
=& 2r^* + \sum_j w_j risk_j(c^*, F) && \text{(from eq. (C.12))} \\
=& 2r^* + risk(c^*, F) = 3r^*
\end{aligned}$$

It remains to prove the correctness of the lemmas.

proof of lemma C.2.3. Non-negativity and symmetry are trivial. $d(f,f) = \mathbb{E}_{x\sim F_X}[\Pr(f(x) \neq f(x)|x)] = \mathbb{E}_{x\sim F_X}[0] = 0$, 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 $x \in \mathcal{X}$

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

and therefore

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

□

proof of lemma C.2.4. Let any $c \in \mathcal{C}$, then from equation (C.12)

$$d(c_i, f_i) = risk_i(c_i, F) \leq risk_i(c, F) = d(c, f_i)$$

□

proof of lemma C.2.5.

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

□

proof of lemma C.2.6.

$$\begin{aligned}
\sum_i \sum_j w_i w_j d(f_i, f_j) &\leq \sum_i \sum_j w_i w_j (d(f_i, c^*) + d(c^*, f_j)) && \text{(Triangle Inequality)} \\
&= \sum_i w_i d(f_i, c^*) \sum_j w_j + \sum_j w_j d(f_j, c^*) \sum_i w_i \\
&= 2 \sum_i w_i d(f_i, c^*) = 2 \sum_i w_i risk_i(c^*, F) && \text{(from eq. (C.12))} \\
&= 2 risk_I(c^*, F) \leq 2r^*
\end{aligned}$$

□

Thus the proof of theorem 5.1.2 is complete. □

Proof of theorem 5.1.1

We first find a lower bound on r^*:

$$\begin{aligned}
r^* &= risk_I(c^*, F) = \sum_{i\in I} w_i risk_i(c^*, F) = \sum_{i\in I} w_i d(c^*, f_i) && \text{(C.14)} \\
&\geq w_j d(c^*, f_j) \geq \frac{1}{n} d(c^*, f_j) && \text{(since } j \text{ is heaviest)}
\end{aligned}$$

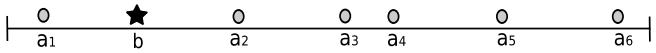


Figure C.1: Consider the interval $[0,1]$. Each point a_i can be thought of as representing the probability that f_i maps x to "+", i.e. $a_i = \overline{f}_i(x)$. Note that the average distance from b only *increases* if it is placed on one of the edges, thus w.l.o.g. $b = \overline{c}(x)$ is placed on either 0 (if $c(x) = -$) or 1 (if $c(x) = +$).

Then we upper bound the risk of c_j:

$$\begin{aligned}
risk_I(c_j, F) =& \sum_{i \in I} w_i d(c_j, f_i) = w_j d(c_j, f_j) + \sum_{i \neq j} w_i d(c_j, f_i) \\
\leq& w_j d(c^*, f_j) + \sum_{i \neq j} w_i (d(c_j, c^*) + d(c^*, f_i)) && \text{(from triangle inequality)} \\
=& d(c_j, c^*) \sum_{i \neq j} w_i + \sum_{i \in I} w_i d(c^*, f_i) = d(c_j, c^*) \sum_{i \neq j} w_i + r^* \\
\leq& d(c_j, c^*) \frac{n-1}{n} + r^* && (w_j \geq \tfrac{1}{n}) \\
\leq& r^* + \frac{n-1}{n} (d(c_j, f_j) + d(f_j, c^*)) && \text{(triagle inequality)} \\
\leq& r^* + \frac{n-1}{n} 2 d(c^*, f_j) && \text{(from lemma C.2.4)} \\
\leq& r^* + \frac{n-1}{n} 2n \cdot r^* && \text{(from eq. (C.14))} \\
=& r^* + (n-1) 2 r^* = (2n-1) r^*
\end{aligned}$$

□

C.3 proof of theorem 5.1.3

Lemma C.3.1. *For all $x \in \mathcal{X}$ and any $c \in \mathcal{C}$,*

$$\sum_i \sum_j \Pr(f_i(x) \neq f_j(x) | x) \leq 2(n-1) \sum_i \Pr(f_i(x) \neq c(x) | x).$$

Remark C.3.2. Let n points $a_1, \ldots, a_n$ scattered across some interval, and let b any other point in this interval, w.l.o.g one of its edges. See figure C.1 for example. A nice geometric interpretation of the lemma (or, more precisely, of the proof), is that the sum of the pairwise distances between the points $\{a_i\}_{i \leq n}$, is bounded by $2(n-1)$ times the sum of distances between these points and b.

Proof. W.l.o.g $c(x) = +$.

$$
\begin{aligned}
&\sum_i \sum_j \Pr(f_i(x) \neq f_j(x)|x) = \sum_i \sum_{j \neq i} \Pr(f_i(x) \neq f_j(x)|x) \\
=&\sum_i \sum_{j \neq i} \Pr(f_i(x) = c(x), f_j(x) \neq c(x)|x) + \Pr(f_i(x) \neq c(x), f_j(x) = c(x)|x) \\
=&\sum_i \sum_{j \neq i} \Pr(f_i(x) = c(x)|x) \Pr(f_j(x) \neq c(x)|x) + \Pr(f_i(x) \neq c(x)|x) \Pr(f_j(x) = c(x)|x) \\
& \qquad \text{(from (C.10))} \\
=&\sum_i \sum_{j \neq i} \Pr(f_i(x) = +|x) \Pr(f_j(x) = -|x) + \Pr(f_i(x) = -|x) \Pr(f_j(x) = +|x) \\
=&\sum_i \sum_{j \neq i} \overline{f}_i(x)\underline{f}_j(x) + \underline{f}_i(x)\overline{f}_j(x) = \sum_i \sum_{j \neq i} (1 - \underline{f}_i(x))\underline{f}_j(x) + \underline{f}_i(x)(1 - \underline{f}_j(x)) \\
=&\sum_i \sum_{j \neq i} \underline{f}_i(x) + \underline{f}_j(x) - 2\underline{f}_i(x)\underline{f}_j(x) \leq \sum_i \sum_{j \neq i} \underline{f}_i(x) + \sum_i \sum_{j \neq i} \underline{f}_j(x) \\
=&2 \sum_i \underline{f}_i(x) \sum_{j \neq i} 1 = 2(n-1) \sum_i \overline{f}_i(x) \\
=&2(n-1) \sum_i \Pr(f_i(x) = -|x) = 2(n-1) \sum_i \Pr(f_i(x) \neq c(x)|x)
\end{aligned}
$$

□

We use this bound to recompute the pairwise differences:

$$
\begin{aligned}
\sum_i \sum_j w_i w_j d(f_i, f_j) =& \frac{1}{n^2} \sum_i \sum_{j \neq i} d(f_i, f_j) \\
=& \frac{1}{n^2} \sum_i \sum_{j \neq i} \mathbb{E}_{x \sim F_X} \left[\Pr(f_i(x) \neq f_j(x)|x) \right] \\
=& \frac{1}{n^2} \mathbb{E}_{F_X} \left[\sum_i \sum_{j \neq i} \Pr(f_i(x) \neq f_j(x)|x) \right] \\
\leq& \frac{1}{n^2} \mathbb{E}_{F_X} \left[2(n-1) \sum_i \Pr(f_i(x) \neq c^*(x)|x) \right] \qquad \text{(from lemma C.3.1)} \\
=& \frac{2(n-1)}{n^2} \sum_i \mathbb{E}_{F_X} \left[\Pr(f_i(x) \neq c^*(x)|x) \right] \\
=& \frac{2(n-1)}{n} \sum_i w_i risk_i(c^*, F) = \frac{2(n-1)}{n} risk_I(c^*, F) = 2\left(1 - \frac{1}{n}\right) r^*
\end{aligned}
$$

Finally, we re-construct the proof of theorem 5.1.2 using the new bounds:

$$\begin{aligned}\sum_{j\in I} w_j risk_I(c_j, F) =& \sum_i \sum_j w_i w_j d(f_i, c_j) \\ =& \sum_i \sum_j w_i w_j d(f_i, f_j) + \sum_j w_j d(f_j, c^*) \sum_i w_i \\ \leq& 2\left(1 - \frac{1}{n}\right) r^* + r^* = \left(3 - \frac{2}{n}\right) r^*\end{aligned}$$

□

C.4 Proof of theorem 6.3.6

Let S a dataset, and let $c^* = c^*(S)$ any optimal classifier w.r.t. S. We refer to agent i as "good" if it agrees with c^* completely, and is "bad" otherwise. Formally, $I_G = \{i \in I \text{ s.t. } risk_i(c^*, S) = 0\}$, and $I_B = I \setminus I_G$. An example $\langle x, y\rangle \in S$ is "good" if it is controlled by a good agent, and "bad" if it is misclassified by c^*. Let G, B the sets of good and bad examples, in accordance. Denote by $G = G_P \cup G_N$ the set of all good examples, divided to positive and negative examples. Note that $G = \bigcup_{i\in I_G} S_i$ and $B \subseteq \bigcup_{i\in I_B} S_i$, i.e. all bad examples are controlled by bad agents, but not all examples are either good or bad.

W.l.o.g the c^* is of the form "$x > c$", i.e. all the positive good examples are on the right side of c, and all the negative ones are on the left. Also, w.l.o.g $|G_P| \geq |G_N|$.

For each $j \in I_B$ denote by $B_j \subseteq S_j$ the set of all bad examples of j that disagree with c^*. Clearly $B = \bigcup_j B_j$. Note that $r^* = \frac{|B|}{m} \geq \frac{|I_B|}{m}$. Define $F_N, (F_P)$ the events that when the first bad agent is selected, at least one negative (positive) good example had already been selected. $F = F_P \wedge F_N$ means that the current range of classifiers is closed from both sides.

We emphasize that the events F_N, F_P are *not* properties of the dataset. Their occurrence (for a given dataset S) depends on the random order in which mechanism 9 selects the agents.

Lemma C.4.1. $\Pr(\neg F_P) \leq \frac{2k|B|}{m}$

Lemma C.4.2. $\Pr(\neg F_N) \leq \frac{k|B|}{|G_N|}$

Lemma C.4.3. $risk(\mathcal{M}_{IR}, S|F_P \wedge F_N) \leq \frac{5k^2|B|}{m}$

Lemma C.4.4. $risk(\mathcal{M}_{IR}, S|F_P \wedge \neg F_N) \leq \frac{5k^2|B|+|G_N|}{m}$

If $|B| = 0$ then all agents agree and the problem is trivial.

Assume $|B| \geq 1$. From the above lemmas:

$$\begin{aligned}
risk(\mathcal{M}_{IR}, S) = &\Pr(F_N \wedge F_P) risk(\mathcal{M}_{IR}|F_N \wedge F_P) + \Pr(F_P \wedge \neg F_N) risk(\mathcal{M}_{IR}|F_P \wedge \neg F_N) \\
&+ \Pr(\neg F_P) risk(\mathcal{M}_{IR}|\neg F_P) \\
\leq &risk(\mathcal{M}_{IR}|F_N \wedge F_P) + \Pr(F_P \wedge \neg F_N) risk(\mathcal{M}_{IR}|\neg F_P \wedge \neg F_N) + \Pr(\neg F_P) \\
\leq &\frac{5k^2|B|}{m} + \Pr(\neg F_N) risk(\mathcal{M}_{IR}|F_P \wedge \neg F_N) + \Pr(\neg F_P) && \text{(from lemma C.4.3)} \\
\leq &\frac{5k^2|B|}{m} + \Pr(\neg F_N) \frac{5k^2|B| + |G_N|}{m} + \frac{2k|B|}{m} && \text{(from lemmas C.4.1,C.4.4)} \\
\leq &\frac{(5k^2 + 2k)|B|}{m} + \frac{5k^2|B|}{m} + \Pr(\neg F_N) \frac{|G_N|}{m} \\
\leq &\frac{(10k^2 + 2k)|B|}{m} + \frac{k|B|}{|G_N|} \frac{|G_N|}{m} && \text{(from lemma C.4.2)} \\
= &\frac{(10k^2 + 2k)|B|}{m} + \frac{k|B|}{m} \\
= &\frac{(10k^2 + 3k)|B|}{m} = (10k^2 + 3k) r^*.
\end{aligned}$$

Thus in order to prove the approximation bound, it suffices to prove the four lemmas. To that matter, we define the following order relation on agents:

Definition C.4.5. An agent $j_1 \in I$ *precedes* $j_2 \in I$, if it is selected first by the mechanism, i.e. if $\pi(i) < \pi(j)$. In that case, we write $j_1 \prec j_2$.[1] We extend this definition to *disjoint sets* of agents. Let $J_1, J_2 \subseteq I$, then $J_1 \prec J_2$ iff

$$\exists j_1 \in J_1 \forall j_2 \in J_2 \left(j_1 \prec j_2\right).$$

Lemma C.4.6. *For any 2 disjoint sets of agents* $J_1, J_2 \subseteq I$, $\Pr(J_1 \prec J_2) = \frac{|J_1|}{|J_1|+|J_2|}$.

Proof. Suppose there are M agents left in the beginning of time t. Denote $J = J_1 \uplus J_2$, and let t^* the round in which the first agent is selected from J.

$$\Pr(J_1 \prec J_2) = \Pr(\pi(t^*) \in J_1) = \frac{|J_1|}{|J|} = \frac{|J_1|}{|J_1| + |J_2|}$$

□

For any set of examples $H \subseteq S$, we denote by $J(H) \subseteq I$ all the agents that control examples in H. Note that the following statements trivially hold:

1. $I_B = J(B)$,

[1] The meaning of the $\prec$ symbol here is unrelated to the same symbol used in section 6.1.

2. $I_G = J(G)$, and

3. $\frac{1}{k}|H| \leq |J(H)| \leq |H|$.

Since good and bad examples are controlled by distinct agents, using lemma C.4.6 on this definition, we get that for any $G' \subseteq G,\ B' \subseteq B$, it holds that

$$\Pr(B' \prec G') = \frac{|J(B')|}{|J(G')| + |J(B')|} \leq \frac{|B'|}{\frac{1}{k}|G'| + |B'|} \leq \frac{k|B'|}{|G'| + k|B'|}. \tag{C.15}$$

Once we rephrase the events F_P, F_N in terms of precedence, the first two lemmas follow directly from equation (C.15):

Proof of lemma C.4.1.

$$\Pr(\neg F_P) \equiv \Pr(B \prec G_P) \leq \frac{k|B|}{|G_P| + k|B|} \leq \frac{k|B|}{|G|/2 + k|B|} = \frac{2k|B|}{|G| + 2k|B|} \leq \frac{2k|B|}{m}$$

□

Proof of lemma C.4.2.

$$\Pr(\neg F_N) \equiv \Pr(B \prec G_N) \leq \frac{k|B|}{|G_N| + k|B|} \leq \frac{k|B|}{|G_N|}$$

□

Proof of lemma C.4.3. Observe that after each iteration of the algorithm, the set of unclassified examples is narrowing, while examples that have already been classified remain untouched. Moreover, under the events F_N, F_P, the set of examples is partitioned to three segments: **negatives** (i.e. examples already classified as negative) on the left, **positives** on the right, and all unclassified examples in the middle segment. The middle segment will be termed as the **active** segment. Examples inside the active segment are *alive*. Examples outside it are *dead*. In each iteration the boundaries of the active segment "move" toward the middle, so once an example dies it can never return to life.

We partition the set of iterations in classification process into *phases*. Each phase ends with the selection of a bad agent that has (at least one) example that *had not been classified* in the previous phases. Note that the examples of the selected agent do not have to be alive in the moment of selection - only at the end of the previous phase. Observe that when a good agent is selected, the right border of the active segment moves (left) to a new location, which is the location of the leftmost live example belongs to this agent. All the examples between the former and the new border of the active segment are "killed", and classified as positive. Similarly, when

a bad agent is selected, the left border moves (right) to the location of the agent's rightmost live example, this time killing all the examples on the left by classifying them as negative. A good example that is killed in this way, i.e. classified as negative, is *ruined*.

Informally, we count how many good examples have been "ruined" (i.e. misclassified) in each phase, in expectation, and show that it is linear in k and in the number of bad examples that survived the previous phase.

Some intuition to this claim may be gained by observing that the number of ruined examples in a phase is affected by the location of the bad examples controlled by the bad agent that is selected in this phase: if a bad (negative) example is relatively on the *right* side, i.e. there are more good (positive) examples on its left, then all these examples may be ruined. On the other hand, being on the right side also increases the chance that by the time this bad example is selected (the end of the phase), it is already dead and cannot ruin any good example at all.

By showing that the number of unclassified bad examples decreases exponentially in each phase, we get that the risk of the algorithm is linear in k^2 and $|B|$, which gives us a good approximation.

Formally, denote by Z_t the number of good examples that are ruined in phase t. Clearly $risk(\mathcal{M}_{IR}) = \frac{\mathbb{E}[Z]+\#(\text{ruined bad examples})}{|S|}$, where $Z = \sum_{t=1}^{|I_B|} Z_t$.

Lemma C.4.7.

$$\forall t \left(\mathbb{E}\left[Z_t | F_P \wedge F_N\right] \leq k|B|(1-\frac{1}{2k})^{t-1} \right)$$

From this lemma it follows that

$$\begin{aligned}\mathbb{E}\left[Z|F_P \wedge F_N\right] =& \mathbb{E}\left[\sum_t Z_t | F_P \wedge F_N\right] = \sum_t \mathbb{E}\left[Z_t|F_P \wedge F_N\right] \leq \sum_t k|B| \left(1-\frac{1}{2k}\right)^{t-1} \\ \leq& k|B| \sum_{t=0}^{\infty} \left(1-\frac{1}{2k}\right)^{t} = k|B|2(1-(1-2k)) = 4k^2|B|\end{aligned} \tag{C.16}$$

Equation C.16 bounds the number of misclassified *good* examples. All other examples belong to bad agents, so they can add at most $k|J(B)| \leq k|B|$ errors. In total:

$$risk(\mathcal{M}_{IR}|F_P \wedge F_N) \leq \frac{4k^2|B| + k|B|}{|S|} = \frac{(4k^2+k)|B|}{m} \leq \frac{5k^2|B|}{m}$$

Proof of lemma C.4.7. We now put forward some formal definitions and notations that will be used in the proof. Recall that for each bad agent $j \in I_B$, B_j is the set of all bad examples that

belong to j. W.l.o.g. we may assume that all of B_j are on the right side of c^*, and therefore negative.[2] Throughout the proof we will use t to count phases, rather than iterations. Denote:

- B^t, G^t are the sets of bad and good points that are not classified at the end of phase t. Note that $B^0 = B$, $G^0 = G$, and $B^t \cup G^t \subseteq S_t$.
- j_t is the bad agent whose selection marks the *end* of phase t.
- $B_t = B_{j_t} \cap B^{t-1}$ are all the examples that are controlled by agent j_t, and have not been classified in previous phases. Assume the examples in $B_t = \{b_{t,1}, b_{t,2}, \ldots, b_{t,|B_t|}\}$ are sorted from left to right, i.e. $b_{t,i}$ denotes the i'th example of the bad agent j_t, from the left side of the active segment.
- $E_{t,i}$ is the event that when agent j_t is selected, $b_{t,i}$ is still unclassified. We denote its indicator variable by $\hat{E}_{t,i}$.
- $A_{t,i} = \{g \in G^{t-1} : x(g) < x(b_{t,i})\}$, i.e. all the good points that are on the left side of $b_{t,i}$, and are still not classified.

The last definition is the most important one. Note that $A_{t,i-1} \subseteq A_{t,i}$. Denote $a_{t,i} = |A_{t,i}|$.

Lemma C.4.8.

$$Z_t \leq \sum_{i=1}^{|B_t|} \hat{E}_{t,i} a_{t,i}.$$

Proof. If $E_{t,i}$ is true (i.e. $b_{t,i}$ is already classified at the time j_t is selected), then it does not have any effect and cannot ruin any good example. If it is not classified, then everything on its left will be classified as negative, which means all the examples in $A_{t,i}$ are ruined. If j_t has more than one bad example, then all the sets of examples that are ruined are overlapping. Hence, only the rightmost bad example (that is not yet classified) will have the effect:

$$Z_t = max_{i:b_{t,i} \in B_t}\{\hat{E}_{t,i} a_{t,i}\} \leq \sum_{i=1}^{|B_t|} \hat{E}_{t,i} a_{t,i}$$

□

Figure C.2 illustrates a single phase.

Lemma C.4.9.

$$\mathbb{E}\left[\hat{E}_{t,i} | B^{t-1}\right] \leq \frac{k|J(B^{t-1})|}{a_{t,i}}$$

[2]Every bad example that is positive is already dead by the time the first bad agent is selected. Thus in the worst case, they are all negative.

By conditioning on the random set B^{t-1}, we are taking the expected value of $\hat{E}_{t,i}$, conditional on the event that B^{t-1} is the set of active bad examples at the beginning of phase t.

Proof. Let any bad example $b_{j,i} \in B^{t-1}$. By definition $E_{j,i}$ is true iff the first example from B^{t-1} is selected before the first example from $A_{j,i}$, i.e. there is an agent controlling an example in B^{t-1} that precedes all agents controlling $A_{j,i}$ in π. Thus, from Lemma C.4.6

$$\mathbb{E}\left[\hat{E}_{j,i}|B^{t-1}\right] = \Pr(E_{j,i}|B^{t-1}) = \Pr(B^{t-1} \prec A_{j,i}) = \frac{|J(B^{t-1})|}{|J(B^{t-1})| + |J(A_{j,i})|} \leq \frac{|J(B^{t-1})|}{\frac{1}{k}|A_{j,i}|} = \frac{k|J(B^{t-1})|}{a_{j,i}}$$

In particular it holds for $j = j_t$, Thus

$$\mathbb{E}\left[\hat{E}_{t,i}|B^{t-1}\right] \leq \frac{k|J(B^{t-1})|}{a_{t,i}}$$

□

Lemma C.4.10.

$$\mathbb{E}\left[Z_t|B^{t-1}\right] \leq k|B^{t-1}|$$

This lemma shows the strong dependency between the number of bad examples that remain active at time t, and the number of good examples that will be ruined in this phase.

Proof.

$$\begin{aligned}
\mathbb{E}\left[Z_t|B^{t-1}\right] &= \sum_{j\in J(B^{t-1})} \Pr(j_t=j)\mathbb{E}\left[Z_t|B^{t-1},j_t=j\right] = \sum_{j\in J(B^{t-1})} \frac{1}{|J(B^{t-1})|}\mathbb{E}\left[Z_t|B^{t-1},j_t=j\right] \\
&\leq \frac{1}{|J(B^{t-1})|}\sum_{j\in J(B^{t-1})} \mathbb{E}\left[\sum_{i=1}^{|B_t|}\hat{E}_{j,i}a_{j,i} \mid B^{t-1},j_t=j\right] && \text{(from lemma C.4.8)} \\
&= \frac{1}{|J(B^{t-1})|}\sum_{j\in J(B^{t-1})} \sum_{i=1}^{|B_j\cap B^{t-1}|}\mathbb{E}\left[\hat{E}_{j,i}a_{t,i}|B^{t-1},j_t=j\right] \\
& && \text{(from the definition of } B_t\text{)} \\
&= \frac{1}{|J(B^{t-1})|}\sum_{j\in J(B^{t-1})} \sum_{i=1}^{|B_j\cap B^{t-1}|}\mathbb{E}\left[\hat{E}_{t,i}|B^{t-1}\right]a_{t,i} \\
&\leq \frac{1}{|J(B^{t-1})|}\sum_{j\in J(B^{t-1})} \sum_{i=1}^{|B_j\cap B^{t-1}|}\frac{k|J(B^{t-1})|}{a_{t,i}}a_{t,i} && \text{(from lemma C.4.9)} \\
&= \frac{1}{|J(B^{t-1})|}\sum_{j\in J(B^{t-1})} \sum_{i=1}^{|B_j\cap B^{t-1}|}k|J(B^{t-1})| \\
&= k\sum_{j\in J(B^{t-1})} \sum_{i=1}^{|B_j\cap B^{t-1}|}1 = k\sum_{j\in J(B^{t-1})}|B_j\cap B^{t-1}| \\
&= k|\biguplus_{j\in J(B^{t-1})}(B_j\cap B^{t-1})| = k|B^{t-1}|
\end{aligned}$$

□

It remains to show that the number of bad examples indeed drops exponentially fast.

Lemma C.4.11. $\mathbb{E}\left[|B_t|\right] \leq |B|(1-\frac{1}{2k})^t$

Proof. Denote by $B_L \uplus B_D = B^{t-1}$ the live and dead bad examples just before j_t is selected. Also denote by $\alpha = \frac{|B_D|}{|B^{t-1}|}$ the fraction of dead examples. For each example $b \in B^{t-1}$ denote by $\#b$ the number of living bad examples on its left, i.e. $\#b = |\{b' \in B^{t-1} : x(b') \leq x(b)\}|$. Clearly the examples that will not continue to the next phase are B_D (which are classified as positive) *plus* all the bad examples that will be classified as negative by the selection of j_t. Note that if $B_t \subseteq B_D$ then no bad examples will be classified as negative in this phase, and otherwise, the rightmost example of B_t that is still alive will classify all the examples from B_L that are on

its left. Denote this set of examples (which j_t kills) by $B_{L'}$, formally:

$$\mathbb{E}\left[|B_{L'}| \mid B_t \subseteq B_D\right] = 0, \tag{C.17}$$

and

$$\begin{aligned}
\mathbb{E}\left[|B_{L'}| \mid B_t \nsubseteq B_D\right] =& \sum_{j \in J(B_L)} \Pr(j_t = j) max_{i \in B_j \cap B_L}\{\#b_{j,i}\} \\
\geq& \sum_{j \in J(B_L)} \Pr(j_t = j) \frac{1}{|B_j \cap B_L|} \sum_{i \in B_j \cap B_L} \#b_{j,i} \\
\geq& \frac{1}{|J(B_L)|} \frac{1}{k} \sum_{j \in J(B_L)} \sum_{i \in B_j \cap B_L} \#b_{j,i} = \frac{1}{k|J(B_L)|} \sum_{b \in B_L} \#b \\
=& \frac{1}{k|J(B_L)|} \sum_{i=1}^{|B_L|} i = \frac{1}{k|J(B_L)|} \frac{|B_L|^2 + |B_L|}{2} \\
\geq& \frac{1}{k|B_L|} \frac{|B_L|^2}{2} \\
=& \frac{|B_L|}{2k} = \frac{|B^{t-1}| - |B_D|}{2k} = |B^{t-1}| \frac{1-\alpha}{2k}.
\end{aligned} \tag{C.18}$$

Also, from (C.15)

$$\begin{aligned}
\Pr(B_t \subseteq B_D) =& \Pr(B_D \prec B_L) = \frac{|J(B_D) \setminus J(B_L)|}{|J(B^{t-1})|} \\
\leq& \frac{|J(B_D)|}{|J(B^{t-1})|} \leq \frac{k|B_D|}{|B^{t-1}|} = k\alpha,
\end{aligned}$$

and therefore

$$\Pr(B_t \nsubseteq B_D) \geq 1 - k\alpha. \tag{C.19}$$

From last three equations:

$$\begin{aligned}
\mathbb{E}\left[|B^t||B^{t-1}\right] =& |B^{t-1}| - |B_D| - \mathbb{E}\left[|B_{L'}| \mid B^{t-1}\right] \\
=& |B^{t-1}|(1-\alpha) - \Pr(B_t \nsubseteq B_D)\mathbb{E}\left[|B_{L'}| \mid B_t \nsubseteq B_D\right] && \text{(from eq. C.17)} \\
\leq& |B^{t-1}|(1-\alpha) - (1-k\alpha)|B^{t-1}|\frac{1-\alpha}{2k} && \text{(from eq. (C.18),(C.19))} \\
=& |B^{t-1}|\left((1-\alpha) - (1-k\alpha)\frac{1-\alpha}{2k}\right) \\
=& |B^{t-1}|\left((1-\alpha) - \frac{1-\alpha-k\alpha+k\alpha^2}{2k}\right) \\
=& |B^{t-1}|\left(1 - \frac{1}{2k} - \alpha\left(1 - \frac{1-\alpha}{2} - \frac{1}{2k}\right)\right) \leq |B^{t-1}|\left(1-\frac{1}{2k}\right).
\end{aligned}$$

Finally, by induction:

$$\begin{aligned}
\mathbb{E}\left[|B^t|\right] =& \mathbb{E}\left[\mathbb{E}\left[|B^t||B^{t-1}\right]\right] \leq \mathbb{E}\left[|B^{t-1}|\left(1-\frac{1}{2k}\right)\right] \\
=& \left(1-\frac{1}{2k}\right)\mathbb{E}\left[|B^{t-1}|\right] \leq \ldots \leq \left(1-\frac{1}{2k}\right)^t |B^0| = \left(1-\frac{1}{2k}\right)^t |B|
\end{aligned}$$

□

To conclude the proof of lemma C.4.7, observe that $|B^t|$ only *decreases* when F_N, F_P occur, since these events mean that good agents were selected earlier. Thus:

$$\begin{aligned}
\mathbb{E}\left[Z_t|F_N,F_P\right] =& \mathbb{E}\left[\mathbb{E}\left[Z_t|B^{t-1}\right]|F_N,F_P\right] \leq \mathbb{E}\left[k|B^{t-1}||F_N,F_P\right] && \text{(from lemma C.4.10)} \\
=& k\mathbb{E}\left[|B^{t-1}||F_N,F_P\right] \leq k\mathbb{E}\left[|B^{t-1}|\right] \leq k|B|\left(1-\frac{1}{2k}\right)^{t-1}
\end{aligned}$$

□

Thereby completing the proof of lemma C.4.3 as well. □

Proof of lemma C.4.4. The main idea of the proof is in understanding that the situation for G_P is the same as in lemma C.4.3.

Consider the special case of $|G_N| = 0$, i.e. all good examples are positive, and some bad examples (possibly negative) are scattered between them. Since, as in lemma C.4.3, the first selected agent controls only good examples (in G_P), the first examples are positive. We now have a middle segment classified as positive, and unclassified examples on either side of it. Once the first bad example (that is alive) is selected, we have the familiar situation we faced in lemma C.4.3. That is, an active segment (i.e. unclassified) between the closest pair of negative

and positive examples, with positive classification on one side, and negative on the other. For all purposes, this is equivalent to the occurrence of $F_P \wedge F_N$. By directly applying lemma C.4.3, we get that no more than $5k^2|B|$ examples are misclassified (in expectation).

Now, adding more examples to $|G_N|$ does not change this scenario, only the probability that the classifier will face left/right. In either case we still have that at most $5k^2|B|$ examples from $S \setminus G_N$ are ruined in expectation. Even if *all* the examples of G_N are ruined, we have that

$$risk(\mathcal{M}_{IR}|F_P \wedge \neg F_N) \leq \frac{5k^2|B| + |G_N|}{m},$$

as needed. □

□

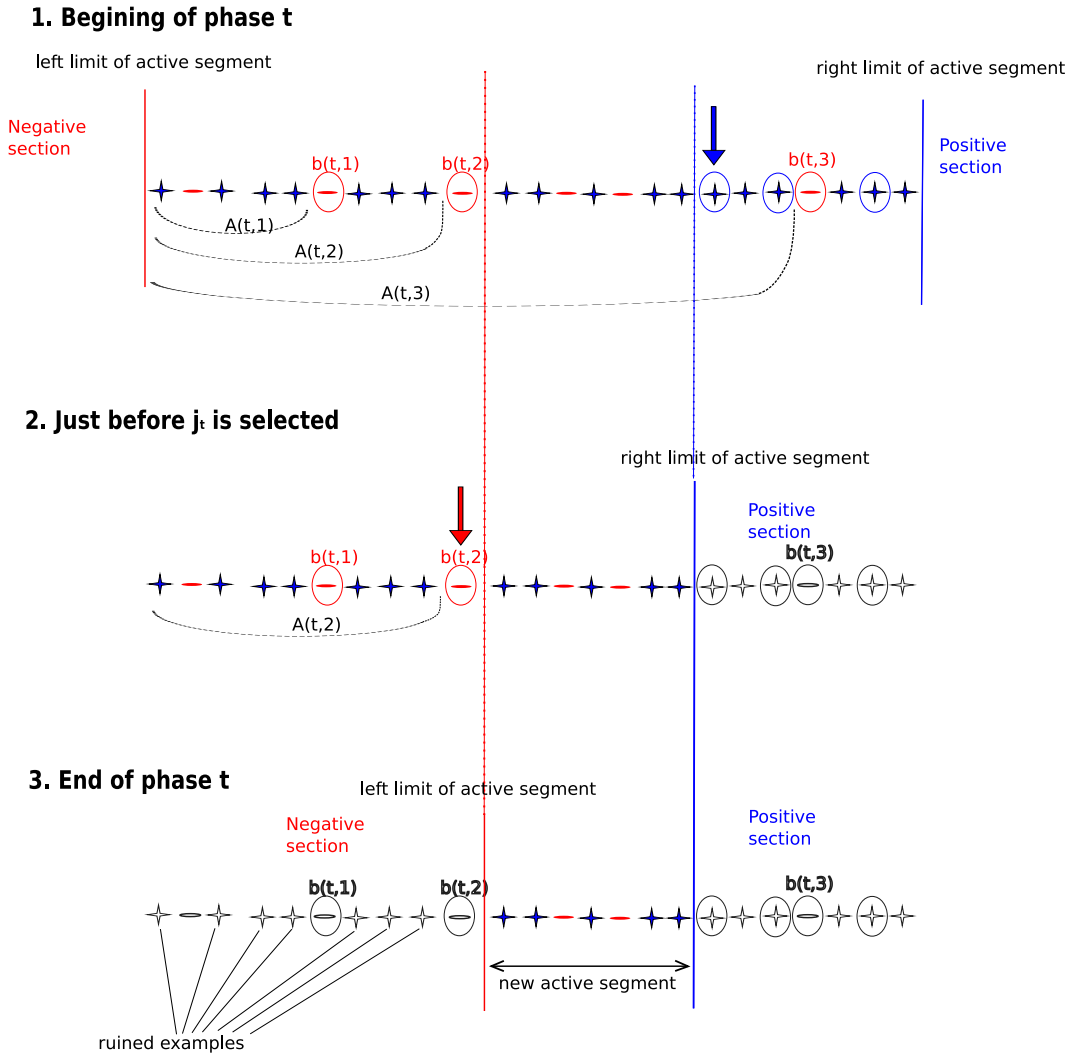


Figure C.2: An illustration of a single phase.
1. The bad examples of agent j_t are circled in red. In this figure $a_{t,1} = 4$, $a_{t,2} = 7$ and $a_{t,3} = 14$. All the good examples selected during the phase (i.e. prior to the selection of j_t) are circled in blue.
2. The leftmost example that was selected before j_t (blue arrow) sets the new limit for the active segments, and "kills" all the examples to its right including $b_{t,3}$. Thus, $E_{t,3} = FALSE$ and $Z_t = max\{\hat{E}_{t,i} \cdot a_{t,i}\} = a_{t,2} = 7$.
3. The rightmost example of j_t that was alive on selection (red arrow) sets the new left limit of the active segment. Since $Z_t = 7$, 7 good examples are ruined in this phase.

Appendix D

Literal Conjunctions

A common concept class used over $\mathcal{X} = \{T, F\}^d$ is the set $\mathcal{C}_d$ of all *literal conjunctions* over d boolean variables. We denote the events $x_i = T$ and $x_i = F$ by $\underline{x}_i, \overline{x}_i$, respectively.

A literal conjunction $c : \{T, F\}^d \to \{+, -\}$ is defined by two sets $pos_c, neg_c \subseteq [d]$. For any input vector $x \in \{T, F\}^d$, $c(x) =$ "+" if

$$(\forall i \in pos_c(\underline{x}_i)) \wedge (\forall i \in neg_c(\overline{x}_i)) \,,$$

and $c(x) =$ "-" otherwise. I.e., each c classifies a hypercube in $\{T, F\}^d$ as positive, and everything else as negative.

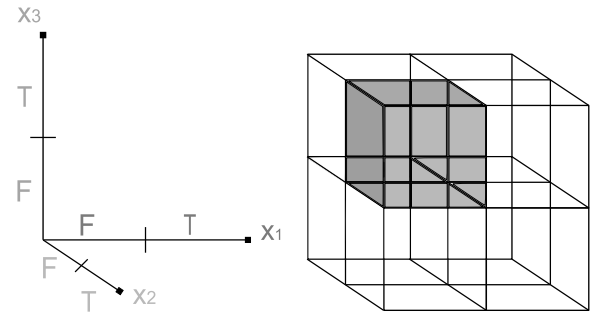
Figure D demonstrates some conjuctions and their spatial interpretation. We refer to the problem of learning a conjuction in $\{T, F\}^d$ as CONJ$_d$.

It is notable that

- if $pos_c = neg_c = \emptyset$ then $c \equiv +$.
- if $pos_c \cap neg_c \neq \emptyset$ then $c \equiv -$.
- $\mathcal{C}_d$ is not symmetric, i.e. in the general case $(-c) \notin \mathcal{C}_d$.

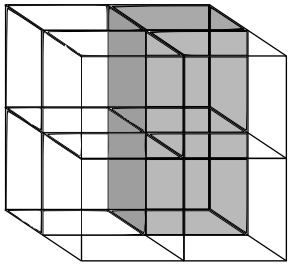
Theorem D.0.12. *There is a* 3*-approximation SP deterministic mechanism and a* 2*-approximation SP randomized mechanism for CONJ*$_1$

Proof. $\{T, F\}^1$ contains exactly 2 different points, i.e. $\mathcal{X} = \{T, F\}$. $\mathcal{C}_1$ contains all 4 possible dichotomies of $\mathcal{X}$. Consequently, we can run any of the mechanisms we used for the constant functions setting (PLUS-MINUS) independetly on $x = T$ and on $x = F$, and approximate the best label for each of them. □

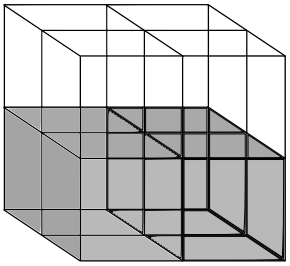


(a)

(b) $c_1 = \overline{x}_1\underline{x}_2\underline{x}_3$, i.e. $pos_{c_1} = \{2,3\}, neg_{c_1} = \{1\}$



(c) $c_2 = \underline{x}_1\overline{x}_2$



(d) $c_3 = \overline{x}_3$

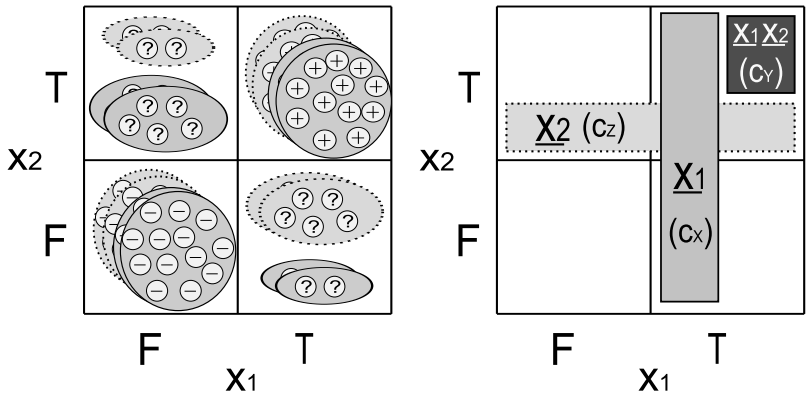
Figure D.1: Conjunctions over the input space $\mathcal{X} = \{T, F\}^3$, which contains 8 different data points. Figure (a) shows the axes of the space. Figures (b)-(d) are examples of different classifiers. The grayed area is the positive part of space.

D.1 Lower bounds

Remark D.1.1. In the synthetic problems, we defined our instances such that c_X labeled x as *negative*, and so on. We could of course define it symmetrically by replacing everywhere the word "negative" with "positive" and vice versa. Assume that this is indeed the case (i.e. that c_X labels x as *positive* and so on).

Theorem D.1.2. *For any $d \geq 2$, there is no deterministic mechanism for CONJ$_d$ that is both SP and β-approximating, for any $\beta = o(\sqrt{m})$.*

Proof. We use are a reduction to $DET{-}SYNTHETIC$. For each agent in $DET{-}SYNTHETIC$ we add one agent to the new instance. Whereas originally each agent controls $2k+3$ examples, it will now control $k' = 2n(2k+3)^2$ examples. Our proof will handle the case $d = 2$, as it can easily be extended to spaces of higher dimension. The total number of examples in the original instance is $m = n(2k+3) = \Theta(nk)$. The total number of examples is thus $m' = nk' = 2(n(2k+3))^2 = \Theta((nk)^2) = \Theta(m^2)$. We translate any example originally on x to the location $a_1 = (F, T)$, and each examples on z to the location $a_2 = (T, F)$. We also flip the label of all examples (i.e. any example which was originally negative will be positive in



(a) The samples of type 1 agents are circled in a bold line. Those of type 2 agents are circled in a dashed line.

(b) The 3 viable classifiers are shown.

Figure D.2: Construction of the lower bound example for the CONJ_2 scenario.

the constructed instance and vice versa). In addition, the remainder of examples of each agent are evenly split between $b_1 = (F, F)$ and $b_2 = (T, T)$. each example on b_1 or b_2 is assigned a negative or positive label respectively. Figure D.1(a) shows how examples are arranged in the constructed instance.

Let $\mathcal{M}$ any deterministic SP mechanism for CONJ_2. There are two possible cases:

- There is an instance S for which $\mathcal{M}(S)$ labels b_1 as positive or b_2 as negative.
- For every instance S, $\mathcal{M}(S) \in \{\underline{x}_1, \underline{x}_2, \underline{x}_1\underline{x}_2\}$.

In the first case, the mechanism makes $\Omega(m')$ errors whereas $\underline{x}_1\underline{x}_2$ makes $O(nk)$ errors. Thus

$$risk_I(\mathcal{M}, S) \geq \Omega(\frac{m'}{nk})r^* = \Omega(\frac{m'}{\sqrt{m'}})r^* = \Omega(\sqrt{m'})r^*.$$

In the second case, we map the conjunction $\underline{x}_1$ to c_X, $\underline{x}_2$ to c_Z, and $\underline{x}_1\underline{x}_2$ to c_Y. Let $\beta' = o(\sqrt{m'})$ and assume that $\mathcal{M}$ indeed guarantees an approximation ratio of β'. Let any instance S of the original problem in DET-SYNTHETIC , and let a denote the minimal number of mistakes in it, i.e. $a = m \cdot r^*$. The minimal number of mistakes in the new instance is also a, by using the conjunction that corresponds to $c^* \in \mathcal{C}_s$. From our assumption, $\mathcal{M}$ does not make more than $\beta' a$ mistakes on the new instance. The corresponding classifier from $\mathcal{C}_s$ makes the same number of mistakes, providing us with an approximation ratio of $\beta' = o(\sqrt{m'}) = o(m)$ to the original problem.

Thus, if m is the number of examples, then the existence of any SP mechanism with approximation ratio of $o(\sqrt{m})$ for $\mathcal{C}_2$ will also provide us with a $o(m)$-approximating mechanism for the synthetic scenario, in contradiction to theorem 6.1.5.

If $d > 2$, then we can simply ignore all literals $x_3, \ldots, x_d$. If $\underline{x}_i \overline{x}_i$ appears in c for some i, then we are in case 1, otherwise, the extra literals do not affect the labels of the samples in S.

□

We want to prove a similar lower bound for the randomized case, but as in the LINEAR$_d$ problem, a straight-forward reduction fails. By adding a minor restriction we are able to prove a slightly weaker claim.

We define the CONJ$'_d$ problem as CONJ$_d$, with the restriction that the concepts in $\mathcal{C}'_d$ may not contain a literal and its negation (equivalently, the positive hypercube cannot be empty).

Theorem D.1.3. *Assume agents can attribute weights to their examples. For all $d \geq 3$, there is no randomized mechanism for CONJ$'_d$ that is both SP and β-approximating, for any $\beta = o(k)$.*

Proof. Similarly to the deterministic case, we create a reduction to RAND-SYNTHETIC . This is simply by mapping the locations x, y, z from the RAND-SYNTHETIC problem, to the locations (T, F, F), (F, T, F) and (F, F, T) in the CONJ$'_3$ problem, respectively. We add negative sample for each agent on each of the other 5 locations in $\{T, F\}^3$. each of these samples is assigned a weight of w', which will be defined later.

Naturally, the classifiers c_X, c_Y and c_Z are translated to the conjuctions $\underline{x}_1 \overline{x}_2 \overline{x}_3$, $\overline{x}_1 \underline{x}_2 \overline{x}_3$ and $\overline{x}_1 \overline{x}_2 \underline{x}_3$, respectively (see figure D.3).

Let $\mathcal{M}$ any randomized mechanism for CONJ$'_3$. Suppose there is an instance S for which $\mathcal{M}$ selects any classifier other than the three we earlier defined with some positive probability $p > 0$. By setting w' s.t. the weight of all "real" samples is smaller than $\frac{p}{\beta}$, $\mathcal{M}$ cannot be β-approximating.

This leaves us with a one-to-one reduction to the RAND-SYNTHETIC problem, thus any randomized SP β-approximating mechanism induces a similar mechanism for the RAND-SYNTHETIC problem. Therefore $\beta = \Omega(k)$, otherwise it is a contradition to theorem 6.1.13. for higher dimensions ($d > 3$), we construct a similar dataset on the first three dimensions, only we add a negative sample for each agent to any of the $2^d - 3$ locations that are not x, y or z. □

As in the cases of the other concept classes, Gibbard's theorem and the weights are only required to show that any SP mechanism must be a mixture of duples as dictatorial mechanisms. If we take the last restriction as it is, then a more powerfull result can be proved:

Theorem D.1.4. *Let $d \geq 2$ and let $\mathcal{M}$ a randomized classification mechanism for learning CONJ$_d$. If $\mathcal{M}$ is a probability mixture of duples and dictatorial mechanisms, then it has an approximation ratio of $\Omega(k)$.*

The proof is almost identical to the proof of theorem 6.3.4.

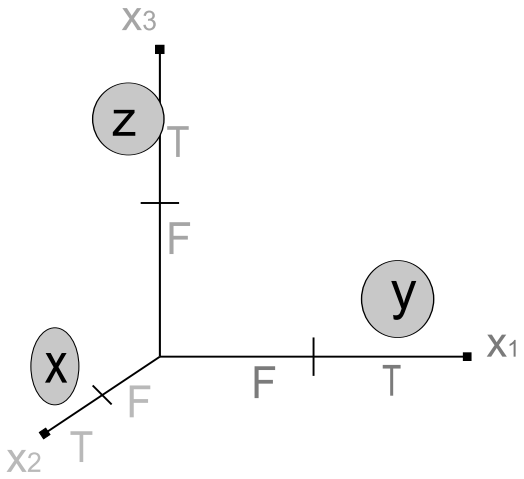


Figure D.3: Construction of the randomized lower bound example for the CONJ$'_3$ scenario.

D.2 Upper bounds

Theorem D.2.1. *Let $\mathcal{X}$ a finite space of size s. For any concept class $\mathcal{C}$ on $\mathcal{X}$, and any dataset S, the IRD mechanism is SP and*

$$risk_I(\mathcal{M}_{IRD}, S) \leq (s \cdot k + 1) r^*(S)$$

which is an $O(k)$-approximation for any $\mathcal{X}$ of a fixed size.

Proof. We first rewrite $\mathcal{X}$ as $(a_1, \ldots, a_s)$. Let $c^* \in \mathcal{C}$ any optimal concept. Denote by I_G all agents that completely agree with c^* (i.e. $risk_i(c^*, S) = 0$). Let $G = \bigcup_{i \in I_G} S_i$ the set of all "good" data points. Denote by B all examples that are inconsistent with c^*. Clearly $r^* = \frac{|B|}{m}$, and $B \subseteq \bigcup_{i \in I_B} S_i$. Note that there may be examples that are not in G nor in B. For each $j \leq s$, we denote by G_j the set of all good examples that are placed on a_j, that is $G_j = \{\langle x, y \rangle \in G | x = a_j\}$.

Lemma D.2.2. *If a good example from G_j (i.e., its controlling agent) is selected prior to* all *bad examples, then all of G_j are labeled correctly in the end of the process.*

Proof. Let $\langle x, y \rangle \in G_j$ the first example selected from G_j (note that $y = c^*(x)$). Suppose that it is controlled by agent $i \in I_G$, and that i is selected in time t before all bad agents. When i is selected, the set of allowed classifiers ($\mathcal{C}_{t-1}$) contains all calssifiers that are consistent with the examples of all prior agents. Since all of these agents are good, they are all consistent with c^*, and thus $c^* \in \mathcal{C}_{t-1}$. for all $t' \geq t$ (and in particular $t' = n$), all concepts in $\mathcal{C}_{t'}$ must agree with $\langle x, c^*(x) \rangle$. This means the final concept has to label as all examples in G_j with the label $c^*(x)$, which is indeed the correct label for good examples. □

Denote by Z_j the number of examples that are labeled correctly in place a_j. Denote by $Z = \sum_j Z_j$ the total number of correct labels.

Denote by e_j the event that $G_j \prec B$ ("there is an agent in G_j that is selected before all agents in B"). From the lemma, we get that if e_j occurs, then $Z_j = |G_j|$. Thus

$$Z_j \geq [\![e_j]\!]|G_j|.$$

Therefore,

$$\begin{aligned}
\mathbb{E}[Z] =& \mathbb{E}\left[\sum_j Z_j\right] \geq \sum_j \mathbb{E}\left[[\![e_j]\!]|G_j|\right] = \sum_j |G_j| \Pr(e_j) \\
=& \sum_j |G_j|(1 - \Pr(B \prec G_j)) = \sum_j |G_j| - \sum_j |G_j| \Pr(B \prec G_j) \\
\geq& \sum_j |G_j| - \sum_j |G_j| \left(\frac{k|B|}{k|B| + |G_j|}\right) \qquad \text{(from lemma D.2.2)} \\
=& |G| - k|B| \sum_j \left(\frac{|G_j|}{k|B| + |G_j|}\right) \geq |G| - k|B| \sum_j \left(\frac{|G_j|}{|G_j|}\right) \\
=& |G| - k|B| \sum_j 1 = |G| - k|B| \cdot s
\end{aligned}$$

The total risk is composed of the misclassified bad examples (at most $|B|$), and the misclassified good examples ($|G| - Z$). Thus

$$risk_I(\mathcal{M}_{IRD}, S) \leq \mathbb{E}\left[\frac{|B| + (|G| - Z)}{m}\right] = \frac{|B|}{m} + \frac{|G| - \mathbb{E}[Z]}{m} \leq r^* + \frac{k|B| \cdot s}{m} = (s \cdot k + 1)r^*.$$

□

Corollary D.2.3. *The IRD mechanism is an SP $(2^d \cdot k + 1)$-approximation mechanism for CONJ$_d$.*

Note that for every fixed d, we get a linear approximation $O(k)$, but the constant rise exponentially with the dimension of the problem. We conjecture that the mechanism is actually stronger:

Conjecture D.2.4. There is a fixed constant q, such that for any $d \geq 1$, and any instance S in the CONJ$_d$ problem, $risk_I(IRD(S), S) \leq q \cdot k(S) \cdot r^*(S)$.

The intuition is similar to the one that guides the proof of the LINEAR$_1$ solution. That is, the "damage" that each bad sample may inflict if selected, is proportional to the probability that it will be eliminated prior to selection.

Bibliography

[1] K. J. Arrow. A difficulty in the concept of social welfare. *The Journal of Political Economy*, 58(4):328–346, 1950.

[2] M.-F. Balcan, A. Blum, J. D. Hartline, and Y. Mansour. Mechanism design via machine learning. In *46th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2005), 23-25 October 2005, Pittsburgh, PA, USA, Proceedings*, pages 605–614. IEEE Computer Society, 2005.

[3] S. Barbera and M. O. Jackson. Strategy-proof exchange. Discussion Papers 1021, Northwestern University, Center for Mathematical Studies in Economics and Management Science, January 1993.

[4] S. Barbera, M.O. Jackson, and A. Neme. Strategy-proof allotment rules. *Games and Economic Behavior*, 18:1–21(21), January 1997.

[5] M. Barreno, B. Nelson, R. Sears, A. D. Joseph, and J. D. Tygar. Can machine learning be secure? In Ferng-Ching Lin, Der-Tsai Lee, Bao-Shuh Lin, Shiuhpyng Shieh, and Sushil Jajodia, editors, *Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security, ASIACCS 2006, Taipei, Taiwan, March 21-24, 2006*, pages 16–25. ACM, 2006.

[6] S. J. Brams, M. A. Jones, and C. Klamler. Better ways to cut a cake. *Notices of the AMS*, 53(11):1314–1321, December 2006.

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

[8] V. Conitzer and T. Sandholm. Complexity of mechanism design. In Adnan Darwiche and Nir Friedman, editors, *UAI '02, Proceedings of the 18th Conference in Uncertainty in Artificial Intelligence, University of Alberta, Edmonton, Alberta, Canada, August 1-4, 2002*, pages 103–110. Morgan Kaufmann, 2002.

[9] V. Conitzer and T. Sandholm. Automated mechanism design for a self-interested designer. In *Proceedings 4th ACM Conference on Electronic Commerce (EC-2003), San Diego, California, USA, June 9-12, 2003*, pages 232–233, New York, NY, USA, 2003. ACM.

[10] N. Dalvi, P. Domingos, Mausam, S. Sanghai, and D. Verma. Adversarial classification. In Won Kim, Ron Kohavi, Johannes Gehrke, and William DuMouchel, editors, *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004*, pages 99–108. ACM, 2004.

[11] O. Dekel, F. Fischer, and A. D. Procaccia. Incentive compatible regression learning. In Shang-Teng Huang, editor, *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2008, San Francisco, California, USA, January 20-22, 2008*, pages 277–286. SIAM, 2008.

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

[13] J. Duggan and T. Schwartz. Strategic manipulability without resoluteness or shared beliefs: Gibbard-Satterthwaite generalized. *Social Choice and Welfare*, 17(1):85–93, 2000.

[14] B. Dutta, H. Peters, and A. Sen. Strategy-proof probabilistic mechanisms in economies with pure public goods. *Journal of Economic Theory*, 106(2):392–416, October 2002.

[15] L. Ehlers, H. Peters, and T. Storcken. Strategy-proof probabilistic decision schemes for one-dimensional single-peaked preferences. *Journal of Economic Theory*, 105(2):408–434, August 2002.

[16] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936.

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

[18] A. Gibbard. Manipulation of schemes that mix voting with chance. *Econometrica*, 45(3):665–681, 1977.

[19] A. Gibbard. Straightforwardness of game forms with lotteries as outcomes. *Econometrica*, 46(3):595–614, 1978.

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

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

[22] D. Lowd and C. Meek. Adversarial learning. In Robert Grossman, Roberto J. Bayardo, and Kristin P. Bennett, editors, *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, Illinois, USA, August 21-24, 2005*, pages 641–647. ACM, 2005.

[23] R. Meir, A. D. Procaccia, and J. S. Rosenschein. Strategyproof classification under constant hypotheses: A tale of two functions. In Dieter Fox and Carla P. Gomes, editors, *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, AAAI 2008, Chicago, Illinois, USA, July 13-17, 2008*, pages 126–131. AAAI Press, 2008.

[24] H. Moulin. On strategy-proofness and single peakedness. *Public Choice*, 35:437–455, 1980.

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

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

[27] N. Nisan and A. Ronen. Algorithmic mechanism design. In *Proceedings of the Thirty-First Annual ACM Symposium on Theory of Computing, May 1-4, 1999, Atlanta, Georgia, USA. ACM, 1999*, pages 129–140. ACM, 1999.

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

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

[30] A. D. Procaccia, A. Zohar, Y. Peleg, and J. S. Rosenschein. Learning voting trees. In *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence, July 22-26, 2007, Vancouver, British Columbia, Canada*, pages 110–115. AAAI Press, 2007.

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

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

[33] 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.

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

[35] Y. Sprumont. Strategyproof collective choice in economic and political environments. *The Canadian Journal of Economics*, 28(1):68–107, 1995.

[36] E. Tardos and V. Vazirani. Basic solution concepts and computational issues. In N. Nisan, T. Roughgarden, É. Tardos, and V. Vazirani, editors, *Algorithmic Game Theory*, chapter 1. Cambridge University Press, 2007.

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

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

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