Tight Bounds for Strategyproof Classification

Reshef Meir
reshef.meir@mail.huji.ac.il

Shaull Almagor
shaull.almagor@gmail.com

Assaf Michaely
assafmichaely@gmail.com

Jeffrey S. Rosenschein
jeff@cs.huji.ac.il
School of Computer Science and Engineering
The Hebrew University of Jerusalem

ABSTRACT

Strategyproof (SP) classification considers situations in which a decision-maker must classify a set of input points with binary labels, minimizing expected error. Labels of input points are reported by self-interested agents, who may lie so as to obtain a classifier more closely matching their own labels. These lies would create a bias in the data, and thus motivate the design of *truthful* mechanisms that discourage false reporting.

We here answer questions left open by previous research on strategyproof classification [12, 13, 14], in particular regarding the best approximation ratio (in terms of social welfare) that an SP mechanism can guarantee for n agents. Our primary result is a lower bound of $3-\frac{2}{n}$ on the approximation ratio of SP mechanisms under the shared inputs assumption; this shows that the previously known upper bound (for uniform weights) is tight. The proof relies on a result from Social Choice theory, showing that any SP mechanism must select a dictator at random, according to some fixed distribution. We then show how different randomizations can improve the best known mechanism when agents are weighted, matching the lower bound with a tight upper bound. These results contribute both to a better understanding of the limits of SP classification, as well as to the development of similar tools in other, related domains such as SP facility location.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence—*Multiagent Systems*

General Terms

Theory, Algorithms, Economics

Keywords

Mechanism design, Classification, Game theory

1. INTRODUCTION

Approximate mechanism design without money (AMDw/oM) is a rapidly growing area of research in game theory and multiagent systems, whose goal is the design of mechanisms for multiagent optimization problems (without the mechanisms' use of payments).

Cite as: Tight Bounds for Strategyproof Classification, Reshef Meir, Shaull Almagor, Assaf Michaely and Jeffrey S. Rosenschein, *Proc. of 10th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2011)*, Yolum, Tumer, Stone and Sonenberg (eds.), May, 2–6, 2011, Taipei, Taiwan, pp. XXX–XXX.

While the underlying problems (e.g., finding the median, or finding the optimal classifier) typically have efficient algorithms, these algorithms may fail in the presence of strategic behavior. Therefore we seek mechanisms that have additional game-theoretic properties (usually strategyproofness) at the expense of a suboptimal, i.e., approximate, behavior.

One particularly interesting AMDw/oM problem is the design of truthful learning algorithms, which incentivize experts to reveal their true opinions, even in cases where they disagree with one another. Within this framework, we focus on binary classification—that is, there is a set of (known) data points that our mechanism needs to classify as positive/negative. Data points can represent, for example, medical records of tumors that an expert-system has to classify as either *malignant* or *benign*. Following the standard classification literature, the classifier is selected from a predefined set of classifiers (e.g., linear separators in some space) known as the *concept class*.

Our mechanism outputs a classifier based on labels collected from n distinct experts. The goal of the mechanism is to maximize social welfare, by selecting a classifier that is close *on average* to the opinions of all experts. However, experts may disagree as to the correct label of a specific point. Furthermore, they may behave strategically, i.e., report false labels if this will bias the resulting classifier to be closer to their opinion. We are therefore interested in *strategyproof* (SP) classification mechanisms, where no agent (expert) can "gain" by lying. As a result, the outcome is just an *approximation* of the optimal classifier, i.e., the selected classifier makes more errors than the optimal one. We seek the best possible approximation ratio that can be guaranteed using SP mechanisms.

1.1 Motivation

Note that the restriction to a predefined concept class is an important part of the problem. Without it, we could simply classify each data point separately. However, as rigorously demonstrated in the machine learning literature, it is precisely this restriction that enables us to generalize, i.e., to apply the outcome classifier on new, unseen, cases. Previous papers on SP classification and learning (see the next section) cover real-world examples where the need to generalize justifies this restriction.

Nevertheless, SP classification might be required also for one-time *decision making*. The following is an example showing how concept class restrictions can be derived from external constraints.

An example.

Consider a situation in which two or more parties (the agents of our scenario) are in a conflict regarding the ownership of a certain piece of land. 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 between them in a way that will maximize the average utility of all the involved parties. It is reasonable 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.

1.2 Related Work

Strategyproof classification.

The first paper on SP classification was by Meir, Procaccia, and Rosenschein [12], who studied a highly restricted case in which only two classifiers are available. The authors proposed a simple deterministic 3-approximation mechanism, and proved that no better (deterministic) SP mechanisms exist. They further demonstrated a randomized SP mechanism that guarantees an approximation ratio of 2, and that this bound is also tight.

We follow an extension of this model outlined by the same authors in [13], where arbitrary concept classes can be used, but the same set of data points is still *shared* by all agents. Notably, no bounded approximation ratio can be guaranteed by deterministic SP mechanisms, but the authors show how selecting a random agent as a dictator guarantees an approximation ratio of 3, and one that is even better ($3 - \frac{2}{n}$) when agents are non-weighted. However, it is unknown whether better randomized mechanisms exist.

A similar model without the shared inputs assumption has also been studied, showing mainly negative results [14]. Using results from social choice theory, the authors showed that deterministic SP mechanisms cannot guarantee any useful approximation ratio. They further conjectured that a similar reduction can be used to supply a lower bound for randomized mechanisms, but failed to supply one that does not require further technical assumptions.

Approximate mechanism design without money.

Mechanisms that deal with strategic behavior of agents have been proposed recently for a large range of applications. While certain restrictions may allow the design of optimal SP mechanisms [19], often this is not the case, and approximation is a must. Outside the classification domain, SP learning algorithms were studied for both clustering [17] and regression [16, 4]. Other mechanisms have been proposed for facility location (see e.g., [1, 11], and [18], which also provides a clear overview of the field), matching [2, 6], resource allocation [8, 9] and more. As our motivating example shows, problems in one domain can sometimes be formalized in other domains as well. There are also interesting similarities between some of the results and techniques in those various domains.

Other related work.

A closely related, yet different, challenge is *adversarial classification* [10, 3, 5]. Here the underlying assumption is that labels are chosen intentionally to hamper the mechanism (for example to avoid spam detection), whereas in our setting the agents are rational, rather than adversarial. Another difference is that the goal of SP classification is to preclude untruthful behavior in the first place, and not to cope with it.

1.3 Our Contribution

We close the gap left open by [13], matching their $3 - \frac{2}{n}$ upper bound for the non-weighted case with an equal lower bound, thus proving its tightness. The proof relies on the fact that every SP mechanism must be (randomly) dictatorial on a subdomain, thereby showing that the technical assumptions in [14] can be eliminated.

We then consider the weighted case, giving three different SP mechanisms for two agents that beat the known upper bound of 3. While the approximation ratio of the first mechanism is still suboptimal ($\sqrt{5}$), it is based on simple heuristics, and shows an interesting relation to the golden ratio. The other two mechanisms guarantee 2-approximation, thereby matching both the upper and lower bounds for two non-weighted agents. Finally, we present a new mechanism for any set of weighted agents, with a guaranteed approximation ratio of $3 - \frac{2}{n}$, thereby improving the previously known upper bound and matching it with the lower bound.

2. MODEL AND NOTATIONS

2.1 Classification

We adopt the shared input model presented in [13], being consistent where possible with their notations. We refer the reader to previous work on SP classification [12, 13, 14] for more details.

We typically denote sets and their elements by $A = \{a_1, a_2, \ldots\}$, and vectors by $\mathbf{a} = (a(1), a(2), \ldots)$. $\Delta(A)$ contains all probability distribution vectors over the set A. $[\![E]\!]$ denotes the indicator variable of the expression E. To facilitate reading, subscripts are sometimes omitted when clear from the context.

Classifiers.

A *classification setting* is a pair $\langle \mathcal{X}, \mathcal{C} \rangle$, where $\mathcal{X}$ (the input space) is some finite set, and $\mathcal{C}$ (the concept class) contains functions of the form $c : \mathcal{X} \to \{-, +\}$. In the land-ownership problem for example, $\mathcal{C}$ contains all the allowed partitions of the territory.

An *instance* of the setting $\langle \mathcal{X}, \mathcal{C} \rangle$ is a tuple defined as $S = \langle X, I, \{Y_i\}_{i \in I}, \mathbf{w} \rangle$, where $X \in \mathcal{X}^k$ is the (public) set of data points to be classified, I is the set of $n \geq 2$ agents, $Y_i : X \to \{-, +\}$ is the "correct" labeling according to agent i, and $w_i \in \mathbb{R}$ is her weight ($\sum_{i \in I} w_i = 1$). Y_i is referred to as agent i's *type*, and it is private information. We denote the partial dataset of agent i by $S_i = \langle X, Y_i \rangle$. $\mathcal{S}$ contains all possible datasets over the input space $\mathcal{X}$. Let $\mathcal{S}_{n,k}$ be the set of all possible datasets S such that $|I| = n$, $|X| = k$. We also allow the limit case $k = \infty$, in which case $Y_i : \mathcal{X} \to [0, 1]_{\mathbb{Q}}$ states the (rational) positive fraction on each input point. $\mathcal{S}$ contains all datasets (finite and infinite).

The *private risk* of a classifier $c \in \mathcal{C}$ is defined as the fraction of agent i's dataset that is misclassified by c, i.e.,

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

As $\mathbf{R}_i(c, S)$ can be seen as a measure of *dissatisfaction* that i suffers due to outcome c, the *global risk* $\mathbf{R}_I(c, S)$ measures the social welfare, i.e. the (dis)satisfaction of the entire society. It is defined as a weighted average over all agents,

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

Let $\mathbf{p} \in \Delta(\mathcal{C})$ be a lottery over the concept class $\mathcal{C}$, that assigns the probability $p(\mathrm{w})$ to the concept c_{w}. For simplicity we treat $\mathbf{p}$ as if it is a classifier, and extend the risk to lotteries linearly, i.e., $\mathbf{R}(\mathbf{p}, S) = \sum_{\mathrm{w} \in \mathcal{X}} p(\mathrm{w}) \cdot \mathbf{R}(c_{\mathrm{w}}, S)$.

We denote by $ERM(S) \in \mathcal{C}$ (for Empirical Risk Minimizer) the concept that makes the smallest number of errors on S. c_i is a shorthand for $ERM(S_i)$ when S is clear from the context.

Mechanisms.

A *randomized mechanism* is a function $\mathcal{M} : \mathcal{S} \to \Delta(\mathcal{C})$, i.e., that for every input dataset of any size, outputs a lottery over classifiers. We denote by $\mathcal{M}(S)$ or $\mathbf{p}_{\mathcal{M}(S)}$ (or just $\mathbf{p}$ when $\mathcal{M}, S$ are clear from the context) the outcome of the randomized mechanism $\mathcal{M}$ on the input dataset S.

Note that we can define a mechanism using a lottery $\mathbf{d}$ over several other mechanisms $\mathcal{M}_1, \mathcal{M}_2, \ldots$, where $\mathbf{p}_{\mathcal{M}(S)}(c)$ equals $\sum d(j)\mathbf{p}_{\mathcal{M}_j(S)}(c)$. We define the following properties:

A *dictator* mechanism is identified with a single agent i. For any S, $\mathcal{M}$ returns $c_i(S)$ with probability 1.

A *duple* is a mechanism that assigns probability 0 to all concepts, except (at most) two.

A *random-dictator* (RD) mechanism is identified with a lottery $\mathbf{d} \in \Delta(I)$ over dictator mechanisms. This distribution may depend on agent weights, if relevant. The two following RD mechanisms are notable special cases:

- The *weighted random dictator* (WRD) mechanism returns $c_i(S)$ w.p. w_i.
- The *heaviest dictator* (HD) mechanism always returns $c_h(S)$, where $h = \text{argmax}_{i \in I} w_i$. Ties are broken in favor of the agent with the higher index, thus h is uniquely defined.

A *random-dictator-duple* (RDD) mechanism is a lottery over dictators and duples.

A mechanism is said to be an *L-approximation* mechanism if its expected risk is at most L times the optimal risk. Formally, for every dataset S

$$\mathbf{R}_I(\mathcal{M}(S), S) \leq L \cdot \mathbf{R}_I(c^*(S), S).$$

A mechanism is said to be *strategyproof* (SP), if no agent can gain (in expectation) by lying. Formally, for every dataset S, agent i, and alternative labels $\overline{S}_i = \langle X, \overline{Y}_i \rangle$,

$$\mathbf{R}_i(\mathcal{M}(S), S) \leq \mathbf{R}_i(\mathcal{M}(S_{-i}, \overline{S}_i), S).$$

Note that duples and dictator mechanisms are always SP. Moreover, RDs and RDDs are also SP.[1]

Intuitively, good mechanisms are both SP and have a low approximation ratio; thus, we are interested in the best possible approximation ratio that can be achieved by randomized SP mechanisms. The following bounds are known:

THEOREM 1 (MEIR, PROCACCIA AND ROSENSCHEIN [12]). *If $|\mathcal{C}| = 2$, then there is a randomized SP mechanism that guarantees a 2-approximation ratio. Furthermore, no SP mechanism can do better.*

Thus for classes of two functions, SP mechanisms are thoroughly understood. For general concept classes, there are upper bounds:

THEOREM 2 (MEIR, PROCACCIA AND ROSENSCHEIN [13]). *For any concept class $\mathcal{C}$, the WRD mechanism guarantees a 3-approximation ratio. If all agents have equal weight, then the approximation ratio is $3 - \frac{2}{n}$.*

There are examples showing that these are the best approximation ratios that WRD can guarantee. However, it has been unknown whether there are *other* SP mechanisms that are better. Our work comes to answer this question. We make use of two additional properties of classification mechanisms.

[1]This is since duples and dictators are SP in *dominant strategies*, not just in expectation, and therefore any combination of them (as long as it does not depend on labels) is still SP.

Let $a \cdot S$ be a *duplication* of S, i.e., every data point in S appears exactly a times in $a \cdot S$, with the same labels. A mechanism is *consistent* if for all $a \in \mathbb{N}, S \in \mathcal{S}$, $\mathcal{M}(S) = \mathcal{M}(a \cdot S)$.

A probability distribution $\mathbf{p}$ is *μ-granular* if all probabilities $p(c)$ are multiples of μ, i.e., if there is some integer vector $\mathbf{q}$ such that $\mathbf{q} \cdot \mu = \mathbf{p}$. A mechanism is said to be *μ-granular* if for all S, $\mathcal{M}(S)$ is μ-granular. Note that when we deal with mechanisms that are implemented on digital computers, it is useful to assume that they will be μ-granular for some μ.

2.2 Voting

Our proofs make extensive use of voting functions and their relations with classification mechanisms. We bring here the definitions relevant to our needs. For a more detailed background on voting, see e.g., [15].

In a voting scenario there is a set of voters (agents) I, and a finite set of candidates $\mathcal{C}$. Each voter has a strict preference order R_i over all candidates. We denote by $c \succ_i c'$ the fact that voter i prefers c over c'. A *preference profile* $R = (R_1, \ldots, R_n)$ contains the preference order of each voter (agent). Let $\mathcal{R}^n$ be the set of all possible preference profiles for n voters, $\mathcal{R} = \bigcup_{n \geq 2} \mathcal{R}^n$.

A *randomized voting rule* is a function $f : \mathcal{R} \to \Delta(\mathcal{C})$. Note that preferences are private, thus the voting rule must use the orders reported by the agents. The definitions of a duple, RD and RDD also apply to voting rules. While the definition of manipulation in deterministic voting rules is straightforward (i.e., there is an agent that can gain by reporting false preferences), it does not apply as-is to randomized rules. This is since the preferences of agent i over lotteries of candidates are not uniquely defined by R_i. To that end, we must introduce cardinal (dis)utilities.[2]

A utility scale $u_i \in \mathbb{R}^{|\mathcal{C}|}$ *fits* order R_i if for all $c, c' \in \mathcal{C}$,

$$u_i(c) < u_i(c') \iff c \succ_i c'.$$

We adopt the same notation to classification settings, meaning that the risk of c is higher than the risk of c'.

A *manipulation* in f (by Gibbard) consists of a profile R, a utility scale u_i that fits R_i, and an alternative order R'_i, such that i gains according to u_i (formally, that $u_i(f(R)) > u_i(f(R_{-i}, R'_i))$). A voting rule is *strategyproof* (SP) if there are no manipulations in f.

THEOREM 3 (GIBBARD '77 [7]). *Let f be a randomized voting rule. If f is SP, then it is a lottery over duples and dictatorial rules.*

3. RESULTS

3.1 Multiple Agents with Uniform Weights

In this section we match the upper bound of $3 - \frac{2}{n}$ with a lower bound, thus proving it is tight.

We use a simple input space with three input points $\mathcal{X} = \{\mathrm{x}, \mathrm{y}, \mathrm{z}\}$. There are 3 classifiers, $\mathcal{C} = \{c_\mathrm{x}, c_\mathrm{y}, c_\mathrm{z}\}$, where $c_\mathrm{w}(\mathrm{w}')$ ="+" for $\mathrm{w} = \mathrm{w}'$ and "-" otherwise. When both the agent and the dataset are clear from the context, we use the shorthand $r(\mathrm{w}) = \mathbf{R}_i(c_\mathrm{w}, S)$.

THEOREM 4. *Let $\mathcal{M}$ be an SP mechanism for the scenario $\langle \mathcal{X}, \mathcal{C} \rangle$. Then for any $\tilde{\epsilon} > 0$ and any $|I| = n \geq 2$, there is an instance S with uniform weights such that*

$$\boldsymbol{R}_I(\mathcal{M}(S), S) > \left(3 - \frac{2}{n} - \tilde{\epsilon}\right) \boldsymbol{R}_I(c^*(S), S).$$

Also, if $\mathcal{M}$ is either *consistent* or *μ-granular, then we can find such a dataset which is finite, and has $k = O\left(\frac{1}{\tilde{\epsilon}}, \frac{1}{\mu}\right)$ data points.*

[2]For consistency with the risk, we treat lower utility as *better*.

We will restrict the allowed datasets as follows. First, X contains exactly k data points on each input point, i.e., $3k$ data points in total. We denote by $\overline{k}_i(\mathrm{w}), \underline{k}_i(\mathrm{w})$ the number of positive and negative labels for each point. We further restrict the labels of each agent, such that: one input point of $\mathcal{X}$ is all negative (i.e., $\overline{k}_i(\cdot) = 0$); one is all positive (i.e., $\overline{k}_i(\cdot) = k$); and the third has at least one label of each (i.e., $1 \leq \overline{k}_i(\cdot) \leq k-1$).

We refer to this third point as the *contingent point*.[3] Clearly, $\mathcal{M}$ is still SP w.r.t. the restricted case.

The risk of each classifier can be simply written (e.g., for c_{x}) as

$$r(\mathrm{x}) = \mathbf{R}_i(c_{\mathrm{x}}, S) = \frac{1}{3k}\left(\underline{k}_i(\mathrm{x}) + \overline{k}_i(\mathrm{y}) + \overline{k}_i(\mathrm{z})\right).$$

Note that every partial dataset S_i is now identified with a strict preference order R_i over $\mathcal{C}$ (for ease of exposition, assume $R_i = (c_{\mathrm{x}} \succ_i c_{\mathrm{y}} \succ_i c_{\mathrm{z}})$), and a rational number $\alpha_i \in (0,1)$ which is the fraction of negative labels on the contingent point y.

To see this, observe that

$$r(\mathrm{x}) = \frac{1-\alpha_i}{3};\ r(\mathrm{y}) = \frac{1+\alpha_i}{3};\ r(\mathrm{z}) = \frac{3-\alpha_i}{3}. \qquad (1)$$

Consequently, $c_{\mathrm{x}}, c_{\mathrm{z}}$ classify the contingent point (which is y in this case) as negative, and c_{y} classifies it as positive.

We can therefore write each S_i as $\langle R_i, \alpha_i \rangle$.

Our proof sketch can be summarized as follows:

1. Give a simpler, normalized presentation of the risk scale.
2. Show that $\mathcal{M}$ is monotonic.
3. Show that any (monotonic) SP mechanism must ignore the value of α.
4. Thus $\mathcal{M}$ is actually a randomized voting rule over $\mathcal{C}$.
5. Since $\mathcal{M}$ is SP, it is an RDD.
6. Duples are bad, so $\mathcal{M}$ is almost entirely an RD.
7. We show a dataset S on which RD mechanisms have a close to $3 - \frac{2}{n}$ approximation ratio.

Crucially, all steps except the last one (Lemma 11) are independent of agent weights.

Proof of Theorem 4. The preference order of agent i over lotteries in a given setting S, is completely defined by her risk scale, i.e., by the vector $\mathbf{r} = (r(\mathrm{x}), r(\mathrm{y}), r(\mathrm{z}))$. Note that the risk of lottery $\mathbf{p}$ according to risk scale $\mathbf{r}$ is the inner product $\mathbf{R}_i(\mathbf{p}, S) = \mathbf{r} \cdot \mathbf{p}$.

DEFINITION 1. *Two risk scales $\mathbf{r}, \mathbf{t}$ are* equivalent*, if for any two outcomes $\mathbf{p}, \mathbf{p}' \in \Delta(\mathcal{C})$,*

$$\mathbf{r} \cdot \mathbf{p} < \mathbf{r} \cdot \mathbf{p}' \iff \mathbf{t} \cdot \mathbf{p} < \mathbf{t} \cdot \mathbf{p}',$$

i.e., if they induce the same order over outcomes.

LEMMA 5 (NORMALIZATION). *Let $S_i = \langle R_i, \alpha_i \rangle$, then the risk scales $\mathbf{r} = (r(\mathrm{x}), r(\mathrm{y}), r(\mathrm{z}))$ and $\mathbf{t} = (0, \alpha_i, 1)$ are equivalent.*

Proof. We denote by $\delta(\mathrm{w}) = p(\mathrm{w}) - p'(\mathrm{w})$. Note that

$$\delta(\mathrm{x}) + \delta(\mathrm{y}) + \delta(\mathrm{z}) = 0. \qquad (2)$$

[3]For infinite datasets with $k = \infty$ this means that the contingent point must have a non-zero fraction of each sign.

In addition, it holds from (1) that

$$\frac{r(\mathrm{y}) - r(\mathrm{x})}{r(\mathrm{z}) - r(\mathrm{x})} = \frac{1 + \alpha_i - (1-\alpha_i)}{3 - \alpha_i - (1-\alpha_i)} = \frac{2\alpha_i}{2} = \alpha_i. \qquad (3)$$

$$\begin{aligned}
&\mathbf{p} \cdot \mathbf{r} < \mathbf{p}' \cdot \mathbf{r} && \iff \\
0 >& p(\mathrm{x})r(\mathrm{x}) + p(\mathrm{y})r(\mathrm{y}) + p(\mathrm{z})r(\mathrm{z}) \\
&- (p'(\mathrm{x})r(\mathrm{x}) + p'(\mathrm{y})r(\mathrm{y}) + p'(\mathrm{z})r(\mathrm{z})) \\
=& \delta(\mathrm{x})r(\mathrm{x}) + \delta(\mathrm{y})r(\mathrm{y}) + \delta(\mathrm{z})r(\mathrm{z}) \\
=& \delta(\mathrm{x})r(\mathrm{x}) + \delta(\mathrm{y})r(\mathrm{y}) + \delta(\mathrm{z})r(\mathrm{z}) \\
&- (\delta(\mathrm{x}) + \delta(\mathrm{y}) + \delta(\mathrm{z}))r(\mathrm{x}) && \text{(from (2))} \\
=& \delta(\mathrm{y})\left(r(\mathrm{y}) - r(\mathrm{x})\right) + \delta(\mathrm{z})\left(r(\mathrm{z}) - r(\mathrm{x})\right) && \iff \\
0 >& \delta(\mathrm{y})\frac{r(\mathrm{y}) - r(\mathrm{x})}{r(\mathrm{z}) - r(\mathrm{x})} + \delta(\mathrm{z}) && \text{(division by a positive number)} \\
=& \delta(\mathrm{y})\alpha_i + \delta(\mathrm{z}) && \text{(from (3))} \\
=& \delta(\mathrm{x})t(\mathrm{x}) + \delta(\mathrm{y})t(\mathrm{y}) + \delta(\mathrm{z})t(\mathrm{z}) = \mathbf{p} \cdot \mathbf{t} - \mathbf{p}' \cdot \mathbf{t},
\end{aligned}$$

thus $\mathbf{p} \cdot \mathbf{t} < \mathbf{p}' \cdot \mathbf{t}$, as required. □

Due to Lemma 5, we can work with the normalized risk scale $\mathbf{t}$ instead of $\mathbf{r}$. This also holds for utility scales of voting functions.

REMARK 1. *Normalization only works for a fixed scale $\mathbf{r}$. If $\mathbf{t}$ is the normalized scale of $\mathbf{r}$, it is not true for example that $\mathbf{p} \cdot \mathbf{t} > \mathbf{p} \cdot \mathbf{t}'$ derives $\mathbf{p} \cdot \mathbf{r} > \mathbf{p} \cdot \mathbf{r}'$.*

The following notations are used in our next two lemmas. Let $S_i = \langle R_i, \alpha \rangle$, $S_i' = \langle R_i, \alpha' \rangle$. Assume w.l.o.g. that $R_i = (\mathrm{x} \succ_i \mathrm{y} \succ_i \mathrm{z})$ (i.e., x has the lowest risk for i). Let $\mathbf{p} = \mathcal{M}(S)$ and $\mathbf{p}' = \mathcal{M}(S')$ denote the outcome of the mechanism on both datasets. Let $\mathbf{t}$ and $\delta(\mathrm{w})$ as in Lemma 5.

Since $\mathcal{M}$ is SP, we have the following constraints:

1. $\mathbf{R}_i(\mathbf{p}, S) \leq \mathbf{R}_i(\mathbf{p}', S)$ (otherwise, i can easily gain by reporting S_i' instead of S_i).
2. $\mathbf{R}_i(\mathbf{p}, S') \geq \mathbf{R}_i(\mathbf{p}', S')$ (otherwise, i can gain by reporting S_i instead of S_i').

We use $r(\mathrm{w})$ and $r'(\mathrm{w})$ as shorthand for $\mathbf{R}_i(\mathrm{w}, S)$ and $\mathbf{R}_i(\mathrm{w}, S')$, respectively.

The next lemma shows that SP mechanisms must be "monotone", i.e., adding more positive labels to a point can only increase the probability that it will be classified as positive.

LEMMA 6 (MONOTONICITY). *If $\alpha < \alpha'$, then $p(\mathrm{y}) \geq p'(\mathrm{y})$.*

Proof. From the first constraint we have that $\mathbf{p} \cdot \mathbf{r} \leq \mathbf{p}' \cdot \mathbf{r}$. From Lemma 5 we can replace $\mathbf{r}$ with the normalized risk $\mathbf{t}$, and thus

$$\begin{aligned}
&\mathbf{p} \cdot \mathbf{t} \leq \mathbf{p}' \cdot \mathbf{t} && \Rightarrow \\
&p(\mathrm{y})\alpha + p(\mathrm{z}) \leq p'(\mathrm{y})\alpha + p'(\mathrm{z}) && \Rightarrow \\
&\delta(\mathrm{y})\alpha \leq -\delta(\mathrm{z}) && (4)
\end{aligned}$$

Similarly, from the second constraint we have that

$$\delta(\mathrm{y})\alpha' \geq -\delta(\mathrm{z}) \qquad (5)$$

Taking the two inequalities together,

$$\begin{aligned}
&\delta(\mathrm{y})\alpha \leq -\delta(\mathrm{z}) \leq \delta(\mathrm{y})\alpha' && \Rightarrow \\
&\alpha\delta(\mathrm{y}) \leq \alpha'\delta(\mathrm{y}) && \Rightarrow \\
&\delta(\mathrm{y}) \leq \frac{\alpha'}{\alpha}\delta(\mathrm{y}) && \Rightarrow \quad (\text{since } \tfrac{\alpha'}{\alpha} > 1) \\
&\delta(\mathrm{y}) \geq 0 \ \Rightarrow\ p(\mathrm{y}) \geq p'(\mathrm{y}) && \square
\end{aligned}$$

OBSERVATION 7. *If there is a manipulation under utility scale* $(0, \alpha, 1)$*, the same manipulation must work* either *for any* $1 > t > \alpha$*,* or *for any* $0 < t < \alpha$*. This follows directly from* (4)*, since the inequality must hold as we change* α *in one of the directions.*

Our next lemma shows that the size of the positive fraction on the contingent point is irrelevant, as long as the preference order R_i is kept.

LEMMA 8 (INVARIANCE OF LABELS).

$$\mathcal{M}(S_{-i}, S_i) = \mathcal{M}(S_{-i}, S_i').$$

Proof. We need to show that the constraints induced by strategyproofness become inconsistent unless the outcomes $\mathbf{p}$ and $\mathbf{p}'$ coincide. Unfortunately, the constraints that follow from α and α' will not suffice, and it is in fact possible to find a pair of outcomes that hold them. The crux lies in adding a *third* point β between the first two, showing that new constraints reach a contradiction.

We rename α' to γ, so that we have $\alpha < \beta < \gamma$. We denote the outcome of $\mathcal{M}$ on each dataset as $\mathbf{p}_\alpha$, $\mathbf{p}_\beta$, and $\mathbf{p}_\gamma$, where $\mathbf{p}_\alpha = \mathcal{M}(S_{-i}, \langle R_i, \alpha \rangle)$, etc. Rewriting (4) and reversing p_α, p_γ,

$$(p_\gamma(\mathrm{y}) - p_\alpha(\mathrm{y}))\alpha \geq p_\alpha(\mathrm{z}) - p_\gamma(\mathrm{z}) \tag{6}$$

Using β, we similarly derive the constraints:

$$(p_\beta(\mathrm{y}) - p_\alpha(\mathrm{y}))\beta \leq p_\alpha(\mathrm{z}) - p_\beta(\mathrm{z}) \tag{7}$$

(otherwise reporting (R_i, α) is a manipulation in β), and

$$(p_\gamma(\mathrm{y}) - p_\beta(\mathrm{y}))\gamma \leq p_\beta(\mathrm{z}) - p_\gamma(\mathrm{z}) \tag{8}$$

(otherwise reporting $\langle R_i, \beta \rangle$ is a manipulation in γ).

Now, assume (towards a contradiction) that $p_\alpha(\mathrm{y}) \neq p_\gamma(\mathrm{y})$. From monotonicity we have that $p_\alpha(\mathrm{y}) > p_\gamma(\mathrm{y})$, and strict inequality also holds for at least one of the subintervals, i.e., either $p_\alpha(\mathrm{y}) > p_\beta(\mathrm{y})$ or $p_\beta(\mathrm{y}) > p_\gamma(\mathrm{y})$.

$$\begin{aligned}
&(p_\gamma(\mathrm{y}) - p_\alpha(\mathrm{y}))\alpha \geq p_\alpha(\mathrm{z}) - p_\gamma(\mathrm{z}) && \text{(from (6))}\\
&\quad = (p_\alpha(\mathrm{z}) - p_\beta(\mathrm{z})) + (p_\beta(\mathrm{z}) - p_\gamma(\mathrm{z}))\\
&\quad \geq (p_\beta(\mathrm{y}) - p_\alpha(\mathrm{y}))\beta + (p_\gamma(\mathrm{y}) - p_\beta(\mathrm{y}))\gamma && \text{(from (7),(8))}\\
&\quad > (p_\beta(\mathrm{y}) - p_\alpha(\mathrm{y}))\alpha + (p_\gamma(\mathrm{y}) - p_\beta(\mathrm{y}))\alpha\\
&\qquad\qquad \text{(from monotonicity and } \alpha < \beta, \gamma)\\
&\quad = (p_\beta(\mathrm{y}) - p_\alpha(\mathrm{y}) + p_\gamma(\mathrm{y}) - p_\beta(\mathrm{y}))\alpha\\
&\quad = (p_\gamma(\mathrm{y}) - p_\alpha(\mathrm{y}))\alpha, \quad \text{which is a contradiction.}
\end{aligned}$$

Thus $p_\alpha(\mathrm{y}) = p_\gamma(\mathrm{y})$, i.e., $\delta(\mathrm{y}) = 0$. From (4) and (5) it follows that $\delta(\mathrm{z}) = 0$. Finally, from (2) we have that $\delta(\mathrm{x}) = 0$ as well, and therefore $\mathcal{M}(S_{-i}, S_i) = \mathbf{p} = \mathbf{p}' = \mathcal{M}(S_{-i}, S_i')$.

A subtle issue lies in the finite k case, since the proof works only for pairs α, γ that differ by at least 2 points (so there is β between them). However, for $k \geq 5$, take any $\alpha < \alpha' < \gamma < \gamma'$. We then have that $\mathbf{p}_\gamma = \mathbf{p}_\alpha = \mathbf{p}'_\gamma = \mathbf{p}'_\alpha$, i.e., the same distribution must be used at every point. □

LEMMA 9 (REDUCTION). *$\mathcal{M}$ is an RDD.*

Proof. This lemma completes the argument that $\mathcal{M}$ is effectively a voting rule, and therefore subject to the known limitations of SP voting rules. It must use our assumptions on $\mathcal{M}$ in order to bound the sample size; however, we first prove the lemma *without* these assumptions, for the limit case of $k = \infty$.

We define a voting rule f as follows. For any profile R, construct the corresponding dataset S by setting $S_i = \langle R_i, \alpha_i \rangle$ for some arbitrary $\alpha_i \in (0, 1)$. The (randomized) outcome of f is defined to be $\mathcal{M}(S)$. From Lemma 8, the choice of α_i does not affect the outcome of f.

Assume (towards a contradiction) that there is a collection of datasets $\hat{\mathcal{S}}$ on which $\mathcal{M}$ is not an RDD. Let $\hat{\mathcal{R}}$ be the corresponding preference profiles to $\hat{\mathcal{S}}$; thus f is not an RDD on these profiles. From Theorem 3, f is not SP, and thus has a manipulation.

W.l.o.g., there is a manipulation (in f) for voter i, such that $\mathrm{x} \succ_i \mathrm{y} \succ_i \mathrm{z}$. By scaling u_i, we can further assume that $u_i(\mathrm{x}) = 0,\ u_i(\mathrm{y}) = \beta,\ u_i(\mathrm{z}) = 1$.[4]

From Observation 7 we can assume that the same manipulation works with $\beta = \frac{1}{k'}$ for some $k' \in \mathbb{N}$ (or $\beta = 1 - \frac{1}{k'}$, which is the symmetric case).

It is easy to see that if $S_i = \langle R_i, \beta \rangle$, then reporting the false labeling $S_i' = \langle R_i', \alpha_i \rangle$ is a manipulation for agent i in $\mathcal{M}$:

$$\begin{aligned}
u_i(f(R)) > u_i(f(R_{-i}, R_i')) \Rightarrow\\
\mathbf{R}_i(\mathcal{M}(S), S) > \mathbf{R}_i(\mathcal{M}(S_{-i}, S_i'), S),
\end{aligned}$$

since u_i is also the normalized risk scale for S_i. This is in contradiction to $\mathcal{M}$ being SP; therefore, $\mathcal{M}$ is an RDD.

Since $\frac{1}{\beta}$ is not bounded, we allow $\underline{k}_i(\mathrm{y})/k$ to take arbitrarily small values, which is the limit case $\mathcal{S}_{k=\infty}$.

Bounding k under the consistency assumption.

We next show how the lemma still holds for *any* k, provided that $\mathcal{M}$ is consistent. It holds from the previous paragraph that $\mathcal{M}$ behaves as an RDD for all datasets of size k' or more. Let $k'' \geq k$ such that $k'' = a \cdot k$ for some integer a. Now consider all a duplications of datasets of size k, i.e., all duplicated datasets $a \cdot S$ s.t. $S \in \mathcal{S}_k$. Since $\mathcal{M}$ is an RDD for $\mathcal{S}_{k''}$, it is in particular an RDD for the duplicated datasets $a \cdot \mathcal{S}_k \subseteq \mathcal{S}_{k''}$, and from consistency also for $\mathcal{S}_k$.

Bounding k under the μ-granularity assumption.

We show that under this assumption, $\mathcal{M}$ is RDD for all datasets of size $k' \geq \frac{2}{\mu}$. Denote by $\mathbf{p}, \mathbf{p}'$ the output of $\mathcal{M}$ on the sets S_i and S_i', respectively, and let $\boldsymbol{\delta} = \mathbf{p} - \mathbf{p}'$. Recall that the normalized utility scale of i is $(0, \beta, 1)$. Since R' is a manipulation, we have that

$$u_i(f(R)) - u_i(f(R_{-i}, R_i')) = \beta\delta(\mathrm{y}) + \delta(\mathrm{z}) > 0. \tag{9}$$

We wish to show that there exists $\beta' \in \left[\frac{\mu}{2}, 1 - \frac{\mu}{2}\right]$ such that if we take $\beta = \beta'$, then R' remains a manipulation (and then k' samples suffice).

Case 1 If $\delta(\mathrm{z}) = 0$, then from (9) we have $\delta(\mathrm{y}) > 0$. Thus, taking $\beta = \mu$ still ensures that R' is a manipulation, since $\mu\delta(\mathrm{y}) + \delta(\mathrm{z}) = \mu\delta(\mathrm{y}) > 0$.

Case 2 If $\delta(\mathrm{z}) > 0$, then by the assumption of μ-granularity we have that $\delta(\mathrm{z}) \geq \mu$. Also, we have the naïve bound of $\delta(\mathrm{y}) \geq -1$. By setting $\beta = \frac{\mu}{2}$ we get $\frac{\mu}{2}\delta(\mathrm{y}) + \delta(\mathrm{z}) \geq -\frac{\mu}{2} + \mu = \frac{\mu}{2} > 0$.

Case 3 If $\delta(\mathrm{z}) < 0$ then by (9) we get $\delta(\mathrm{y}) \geq \beta\delta(\mathrm{y}) > -\delta(\mathrm{z})$. Thus, we can write $-\delta(\mathrm{z}) = a\mu$ and $\delta(\mathrm{y}) = b\mu$ for integers

[4] More formally, if there is a manipulation according to u_i, then from Lemma 5 the same manipulation works with the utility scale $u' = (0, \beta, 1)$, where $\beta = \frac{u_i(\mathrm{z}) - u_i(\mathrm{y})}{u_i(\mathrm{z}) - u_i(\mathrm{x})}$.

	S_1				$S_j,\ j \neq 1$			
	x	y	z	$\mathbf{R}_1(c)$	x	y	z	$\mathbf{R}_j(c)$
$\overline{k}_i(\cdot)/k$	$1-\epsilon$	1	0		1	ϵ	0	
err of c_x	ϵ	1	0	$1+\epsilon$	0	ϵ	0	ϵ
err of c_y	$1-\epsilon$	0	0	$1-\epsilon$	1	$1-\epsilon$	0	$2-\epsilon$

Table 1: The first row shows the positive fraction on each point in S. The next rows describe the errors that each classifier makes on each point. $\mathbf{R}_i(c,S)$ is the sum of error fractions of c over the three points in S_i.

$\frac{1}{\mu} \geq b > a \geq 0$. From this we get

$$\frac{-\delta(\text{z})}{\delta(\text{y})} = \frac{a\mu}{b\mu} \leq \frac{a\mu}{(a+1)\mu} = \frac{a}{a+1} = 1 - \frac{1}{a+1}$$
$$\leq 1 - \frac{1}{\frac{1}{\mu}+1} = 1 - \frac{\mu}{1+\mu} < 1 - \frac{\mu}{2}.$$

Thus, we have $\left(1 - \frac{\mu}{2}\right)\delta(\text{y}) + \delta(\text{z}) > 0$. □

We introduce a small constant $\epsilon > 0$, whose value will be determined later. For now it is sufficient to require that the number of samples k would be at least $\frac{1}{\epsilon}$, so that the contingent point can have a positive fraction of ϵ or less.

LEMMA 10. *If $\mathcal{M}$ returns a duple with some probability greater than 3ϵ, then its approximation ratio is at least 3.*

Proof. Suppose that with probability of at least 3ϵ, $\mathcal{M}$ returns a duple over $\{c_x, c_y\}$. We define a dataset S, in which all agents label z as positive ,x as negative, and y with a positive fraction of ϵ (i.e., $\overline{k}_i(\text{z}) = k$, $\overline{k}_i(\text{x}) = 0$, and $\overline{k}_i(\text{y}) = 1$).[5] The optimal classifier $c^*(S)$ is of course c_z, with a global risk of $r^* = \frac{1}{3k}$.

However, $\mathcal{M}$ must return c_y (or c_x) w.p. of at least 3ϵ; thus its risk is at least $3\epsilon \cdot \mathbf{R}_I(c_y, S) = 3\epsilon\left(\frac{1}{3}(1+\frac{1}{k})\right) > \epsilon \geq 3 \cdot r^*$. □

We can therefore assume that $\mathcal{M}$ returns a random dictator w.p. of at least $1 - 18\epsilon$ (there are 6 different duples, and each one has a probability of at most 3ϵ).

LEMMA 11. *Assume all n agents have the same weight. If $\mathcal{M}$ returns a random dictator (i.e., some lottery $\mathbf{d}$ over agents), then the approximation ratio of $\mathcal{M}$ is at least $3 - \frac{2}{n} - \epsilon''$, where $\epsilon'' = 2n\epsilon + 96\epsilon > 0$.*

Proof. Let i (w.l.o.g. $i = 1$) be the agent selected with the highest probability (i.e., $d(1) \geq \frac{1}{n}$). We define the dataset S as follows: $S_1 = \langle(\text{y} \succ \text{x} \succ \text{z}), 1-\epsilon\rangle$, and for all $j \neq 1$, $S_j = \langle(\text{x} \succ \text{y} \succ \text{z}), \epsilon\rangle$. Thus the selected concept of agent 1 is $c_1 = c_y$, and the selected concept of any other agent is $c_j = c_x$ (which is also the optimal concept). The construction of S is given in Table 1. To simplify computations, we do not divide the risk by the number of points and agents, and thus the global risk is in the range $[0, 3n]$. Thus,

$$r^*(S) = \mathbf{R}_I(c_x, S) = \mathbf{R}_1(c_x, S_1) + (n-1)\mathbf{R}_j(c_x, S_j) \quad (10)$$
$$= 1 + \epsilon + (n-1)\epsilon = 1 + n\epsilon, \text{ whereas}$$
$$\mathbf{R}_I(c_y, S) = \mathbf{R}_1(c_y, S_1) + (n-1)\mathbf{R}_j(c_y, S_j) \quad (11)$$
$$= 1 - \epsilon + (n-1)(2-\epsilon) = 2n - 1 - n\epsilon.$$

[5] In the limit case replace $\frac{1}{k}$ with ϵ, as any fraction is allowed.

Our RD mechanism returns $c_1 = c_y$ w.p. of $d(1) \geq \frac{1}{n}$, and the best thing it can do is return $c^* = c_x$ w.p. of $1 - \frac{1}{n}$. The risk of the mechanism can be lower-bounded as follows:

$$\begin{aligned}
\mathbf{R}_I(\mathcal{M}) &\geq \frac{1}{n}\mathbf{R}_I(c_y, S) + \frac{n-1}{n}r^* \\
&\geq \frac{1}{n}(2n - 1 - n\epsilon) + \frac{n-1}{n}(1+n\epsilon) \qquad \text{(from (10),(11))} \\
&= 2 - \frac{1}{n} - \epsilon + 1 + n\epsilon - \frac{1}{n} - \epsilon \\
&= 3 - \frac{2}{n} + (n-2)\epsilon = 3 - \frac{2}{n} + (\epsilon'' - \epsilon'') + (n-2)\epsilon \\
&= 3 - \frac{2}{n} - \epsilon'' + (2n\epsilon + 96\epsilon) + n\epsilon - 2\epsilon \\
&> 3 - \frac{2}{n} - \epsilon'' + \left(3 - \frac{2}{n} - \epsilon''\right)n\epsilon \\
&= (3 - \frac{2}{n} - \epsilon'')(1+n\epsilon) = (3 - \frac{2}{n} - \epsilon'')r^*.
\end{aligned}$$

□

Finally, we bound the total risk of $\mathcal{M}$. Due to Lemma 9, the outcome of $\mathcal{M}$ is an RDD, i.e., a lottery over all 6 possible duples, and n possible dictators. We denote by RD the event that $\mathcal{M}$ selected any of the dictators. Note that due to Lemma 10, either $Pr(RD) \geq 1 - 18\epsilon$, or the approximation ratio of $\mathcal{M}$ is at least 3 (and thus we are done).

Assume therefore that $Pr(RD) \geq 1 - 18\epsilon$. From Lemma 11 we have that $\mathbf{R}_I(\mathcal{M}(S), S|RD) \geq (3 - \frac{2}{n} - \epsilon'')r^*(S)$ (for S as defined in the lemma). Denote $\epsilon' = 18\epsilon$, $\tilde{\epsilon} = \epsilon'' + 6\epsilon' = (2n+200)\epsilon$.

$$\begin{aligned}
\mathbf{R}_I(\mathcal{M}(S), S) &= Pr(RD)\mathbf{R}_I(\mathcal{M}(S), S|RD) \\
&\quad + Pr(\neg RD)\mathbf{R}_I(\mathcal{M}(S), S|\neg RD) \\
&\geq Pr(RD)\mathbf{R}_I(\mathcal{M}(S), S|RD) \\
&\geq (1-\epsilon')\left(3 - \frac{2}{n} - \epsilon''\right)r^*(S) \qquad \text{(from Lemmas 10,11)} \\
&> (1-\epsilon')\left(3 - \frac{2}{n} - \tilde{\epsilon} + 6\epsilon' - \frac{4}{n}\epsilon' - 2\tilde{\epsilon}\epsilon'\right)r^*(S) \\
&= (1-\epsilon')\left(3 - \frac{2}{n} - \tilde{\epsilon}\right)(1+2\epsilon')r^*(S) \\
&= \left(1 + \epsilon' - 2(\epsilon')^2\right)\left(3 - \frac{2}{n} - \tilde{\epsilon}\right)r^*(S) \\
&> \left(3 - \frac{2}{n} - \tilde{\epsilon}\right)r^*(S).
\end{aligned}$$

This concludes our proof, as for any $\tilde{\epsilon}$, we only need to set ϵ small enough (i.e., k large enough). Specifically, $k \geq \frac{1}{\epsilon} = \frac{2n+200}{\tilde{\epsilon}}$ will suffice. ■

3.2 Two Weighted Agents

In this section, we restrict our analysis to datasets that are composed of just two partial datasets. Due to [13] we know that the WRD mechanism guarantees a 3-approximation ratio in the worst-case. Moreover, we know that for this mechanism the analysis is tight when the smaller weight approaches 0. As for a lower bound, we know from [12] that it is at least 2. Theorem 4 does not contribute anything in this case, both because weights are non-uniform, and because $3 - \frac{2}{n}$ for $n = 2$ is still 2.

Due to Lemmas 9 and 10, we know that in this case too, any SP mechanism must be an RD (with high probability), but we still have the freedom to define the probability of selecting each of the

two dictators, according to their weights.

Unless explicitly stated otherwise, we assume w.l.o.g. that $w_1 \leq \frac{1}{2} \leq w_2$, and denote $w = w_1$. We consider the HD and WRD mechanisms, as described in Section 2.1. Clearly both mechanisms are SP.

Consider Theorem 2. A slight variation of its proof reveals a more accurate bound. Let $w_{\min} = \min_{i \in I} w_i$ be the weight of the lightest agents (in the two agent case, $w_{\min} = w$).

THEOREM 12. *WRD has an approximation ratio of* $3 - 2w_{\min}$, *and this bound is tight.*

The following lemma will be useful in the analysis of our proposed mechanisms. The proof is omitted due to space constraints.

LEMMA 13. *Let* $S = \langle X, I, \{Y_i\}_{i \in I}, \mathbf{w} \rangle$ *be some instance with* n *agents. Suppose we remove an agent (w.l.o.g. agent 1), thereby creating an instance* $S' = \langle X, I', \{Y_i\}_{i \in I'}, \mathbf{w}' = (w_2, \ldots, w_n) \rangle$. *Let* $c' = c^*(S')$ *be the optimal classifier for* S'; *then*

$$\boldsymbol{R}_I(c', S) \leq \frac{1 + w_1}{1 - w_1} \boldsymbol{R}_I(c^*(S), S).$$

THEOREM 14. *HD has an approximation ratio of* $\frac{1+w}{1-w}$, *and this bound is tight.*

Proof. The upper bound follows immediately from Lemma 13, as c' is selected by the remaining, heavier, agent. For tightness, consider the following scenario. Let $w \leq \frac{1}{2}$. There are 2 samples: $X = \{x, y\}$. Agent 1 classifies both as "-", and agent 2 classifies x as "+" and y as "-". There are two classifiers, $\mathcal{C} = \{c_+, c_-\}$, that classify both samples as "+" and "-", respectively. The optimal classifier is obviously c_-, whose risk is $1 - w$. However, the heaviest dictator is agent 2, who chooses c_+ (we assume a bias for tie-breaking). The risk of c_+ is $2w + 1 - w = 1 + w$. Thus, the approximation ratio in this case is $\frac{1+w}{1-w}$. ■

Next, we combine HD and WRD into a better SP mechanism. Let $T = \frac{3-\sqrt{5}}{2}$. We define the *threshold dictator* (TD) as follows.

- The TD mechanism behaves like WRD when $w > T$ and like HD otherwise.

COROLLARY 15. *TD has a worst-case approximation ratio of* $\sqrt{5}$, *and this bound is tight.*

Proof. Suppose $w \leq T$. Then from Theorem 14 the approximation ratio of TD is $\frac{1+w}{1-w} \leq \frac{1+T}{1-T} = \sqrt{5}$. Now suppose $w > T$; then from Theorem 12 the approximation ratio of TD is $3 - 2w \leq 3 - 2T = \sqrt{5}$. The lower bound is achieved for $w = T$. ■

Curiously, the optimal threshold T is such that the ratio between agents' weights is exactly Φ, the golden ratio.

A natural question is whether *even better* SP mechanisms exist, and in particular mechanisms that match the lower bound of $3 - \frac{2}{2} = 2$. Interestingly, the answer is *yes*, and we now give two examples of such mechanisms.

- The *square-weight random dictator* (SRD) mechanism returns c_i w.p. $\frac{w_i^2}{\sum_{j \in I} w_j^2}$.

THEOREM 16. *For two agents, the SRD mechanism has a worst-case approximation ratio of* 2.

Proof. We will use the following lemma, showing a reduction to a simpler problem (proof omitted).

LEMMA 17. *Consider a setting with only two concepts that disagree on all points* $\{c_-, c_+\}$, *and let* $\mathcal{M}$ *be an RD mechanism for two agents. If* $\mathcal{M}$ *guarantees* L*-approximation in this restricted setting (for* $L \geq 2$*), then* $\mathcal{M}$ *is an* L*-approximation mechanism.*

Due to Lemma 17, we can assume that c_1, c_2 completely disagree, and that one of them is the optimal classifier c^*. Assume w.l.o.g. that $c^* = c_1$, and denote the optimal risk by r^*.

Suppose first that $w > 1 - w$. This is the easy case, as it implies that the better classifier is selected with greater probability. Assume therefore that $w \leq 1 - w$, and consider mechanism HD. In the latter case, we have that $\mathbf{R}_I(HD(S), S) = 1 - r^*$. From Theorem 14 we have that $1 - r^* \leq \frac{1+w}{1-w} r^*$, therefore

$$\begin{aligned}
\mathbf{R}_I(SRD(S), S) &= \frac{w^2 \mathbf{R}_I(c_1, S) + (1-w)^2 \mathbf{R}_I(c_2, S)}{w^2 + (1-w)^2} \\
&= \frac{w^2 r^* + (1-w)^2 (1 - r^*)}{w^2 + (1-w)^2} \leq \frac{w^2 r^* + (1-w)^2 \frac{1+w}{1-w} r^*}{w^2 + (1-w)^2} \\
&= \frac{w^2 r^* + (1-w)(1+w) r^*}{w^2 + (1-w)^2} = \frac{1}{2w^2 - 2w + 1} r^*. \\
&\leq \frac{1}{1/2} r^* = 2r^*,
\end{aligned}$$

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

By considering Lemma 17 together with Theorem 1, it follows directly that there is another 2-approximation mechanism, using the same randomization suggested by Meir, Procaccia and Rosenschein for the two-function setting [12]. We refer to this mechanism as MPR8.[6]

3.3 More than Two Weighted Agents

In this final section we extend our results beyond the two-agent setting, describing a worst-case optimal SP mechanism for any set of weighted agents.

We first try the threshold approach. Theorem 12 supplies us with an approximation ratio of $3 - 2w_{\min}$ for the WRD mechanism. Suppose we have some SP d_{n-1}-approximation mechanism $\mathcal{M}_{n-1}$ for $n - 1$ agents, where $d_{n-1} < 3$. We can derive an SP mechanism $\mathcal{M}_n$ for n agents as follows: set a threshold $T_n \in (0, 1)$. If all agents weigh more than T_n, use WRD. Otherwise, remove the lightest agent and run $\mathcal{M}_n$ on the remaining data.

THEOREM 18. *Mechanism* $\mathcal{M}_n$ *is SP, and has an approximation ratio of* $\max\left\{3 - 2T_n, \frac{1+T_n}{1-T_n} d_{n-1}\right\}$.

The proof follows directly from Lemma 13 and Theorem 12.

We can bound the worst-case approximation then, by setting T_n such that $3 - 2T_n = \frac{1+T_n}{1-T_n} d_{n-1}$. As a special case for $n = 2$, we get the TD mechanism with $\sqrt{5}$ approximation (Theorem 15). Also, we know that $d_2 = 2$ (from Theorem 16), and thus by setting the threshold for three agents to $T_3 \cong \frac{3}{20}$, we get a (roughly) $3 - \frac{6}{20} = 2\frac{7}{10}$ approximation mechanism for three weighted agents. Similar threshold mechanisms can be iteratively derived for any number of agents. While this mechanism already beats the upper bound of 3, it does not match the lower bound of $3 - \frac{2}{n}$.

We finally turn to describing our final mechanism, which either generalizes or beats all previous mechanisms for SP classification with shared inputs. Let $p'_i = \frac{w_i}{2(1-w_i)}$, and $\alpha_{\mathbf{w}} = \frac{1}{\sum_{i \in I} p'_i}$.

[6]The mechanism, applied to our scenario, would select the lighter and heavier agents w.p. of $\frac{w}{2-2w}$ and $\frac{2-3w}{2-2w}$, respectively.

- The *convex-weight random dictator* (CRD) mechanism, returns c_i w.p. $p_i = \alpha_{\mathbf{w}} p'_i$.

THEOREM 19. *The CRD mechanism has an approximation ratio of* $\alpha_{\mathbf{w}} + 1$, *which is at most* $3 - \frac{2}{n}$.

We omit the proof due to space constraints. However, we note that it is based on the convexity of the weight function, giving rise to the name of the mechanism. When applied to two agents, the CRD mechanism is similar (but not identical) to the MPR8 mechanism, and can therefore be seen as a generalization of it. Moreover, all the upper bounds in [12, 13], as well as the ones in this paper, follow as special cases from Theorem 19.

4. DISCUSSION

Our results have two primary implications on strategyproof classification. On the negative side, we have shown that the use of dictators is necessary if one wants to maintain truthfulness in learning algorithms, even when randomization is allowed. This means in particular that the previously known bounds for SP classification with uniform weights are tight.

On the positive side, we show that while dictators play a key role in SP classification, non-trivial selection of the dictator can lead to improvements in the approximation ratio of the mechanism. We demonstrated how simple threshold heuristics can be used to safely discard low-weight agents, thus improving the worst-case approximation ratio (although it is still suboptimal). Our main positive result is the CRD mechanism, which matches the lower bound for SP classification and therefore cannot be further improved. In addition to generalizing all previously known upper bounds for the shared input setting (from [12, 13]), our result shows that the uniform weight case is also the most difficult, and a better approximation ratio can be achieved as weights become more biased in favor of some agents.

The learning-theoretic setting.

An important issue is the possibility to generalize from sampled data, and apply the result classifier on unseen data from the same distribution (a task known as *supervised learning*). It is shown in Section 3 in [13] how the WRD mechanism can be extended in such a way to a learning-theoretic setting. We note that all of our mechanisms can be applied directly to the learning-theoretic setting, making the same strategic assumptions described in [13].

Future research.

Perhaps more important than the specific bounds we proved, our results and techniques may aid in improving the understanding of randomized approximation mechanisms in other domains. Some mechanisms for facility location [1] are based on ideas similar to the WRD mechanism; our insights can be used to improve their weighted versions. Also, our impossibility proof tackles rather general issues, such as continuity and private information.[7] This may also help in the study of lower bounds in other domains.

Other future directions may include the study of new types of strategic behaviors in learning problems, and providing a more formal picture of the relations between seemingly unrelated AMDw/oM problems.

[7]This is in contrast to [14], for example, where the lower-bound proof relies on the intricate details of the reduction.

Acknowledgments

This work was partially supported by Israel Science Foundation grant #898/05, and Israel Ministry of Science and Technology grant #3-6797.

5. REFERENCES

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

[2] I. Ashlagi, F. Fischer, I. Kash, and A. D. Procaccia. Mix and match. In *Proc. of 11th EC*, pages 305–314, 2010.

[3] N. Dalvi, P. Domingos, Mausam, S. Sanghai, and D. Verma. Adversarial classification. In *Proc. of 10th KDD*, pages 99–108, 2004.

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

[5] O. Dekel and O. Shamir. Good learners for evil teachers. In *Proc. of 26th ICML*, pages 216–223, 2009.

[6] S. Dughmi and A. Ghosh. Truthful assignment without money. In *Proc. of 11th EC*, pages 325–334, 2010.

[7] A. Gibbard. Manipulation of schemes that mix voting with chance. *Econometrica*, 45:665–681, 1977.

[8] M. Guo and V. Conitzer. Strategy-proof allocation of multiple items between two agents without payments or priors. In *Proc. of 9th AAMAS*, pages 881–888, 2010.

[9] M. Guo, V. Conitzer, and D. Reeves. Competitive repeated allocation without payments. In *Proc. of 5th WINE*, pages 244–255, 2009.

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

[11] P. Lu, X. Sun, Y. Wang, and Z. A. Zhu. Asymptotically optimal strategy-proof mechanisms for two-facility games. In *Proc. of 11th EC*, pages 315–324, 2010.

[12] R. Meir, A. D. Procaccia, and J. S. Rosenschein. Strategyproof classification under constant hypotheses: A tale of two functions. In *Proc. of 23rd AAAI*, pages 126–131, 2008.

[13] R. Meir, A. D. Procaccia, and J. S. Rosenschein. Strategyproof classification with shared inputs. In *Proc. of 22nd IJCAI*, pages 220–225, 2009.

[14] R. Meir, A. D. Procaccia, and J. S. Rosenschein. On the limits of dictatorial classification. In *Proc. of 9th AAMAS*, pages 609–616, 2010.

[15] B. Peleg. Game-theoretic analysis of voting in committees. In K. Arrow, A. Sen, and K. Suzumura, editors, *Handbook of Social Choice and Welfare, Vol. 1*, chapter 8. Elsevier Science B., 2002.

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

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

[18] A. D. Procaccia and M. Tennenholtz. Approximate mechanism design without money. In *Proc. of 10th EC*, pages 177–186, 2009.

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

APPENDIX

Proof of Theorem 12, for $n = 2$. For ease of exposition, we assume that $w = \frac{1}{m}$ for some integer $m \geq 2$. Let $S = (S_1, S_2)$ be a two-agent dataset. We now define a similar dataset S' for $n = m$ agents with uniform weights: $S'_1 = S_1$, and for all $i > 1$, $S'_i = S_2$. Now suppose we use the (n-agent) WRD mechanism on S'. Clearly, WRD returns c_1 w.p. of $\frac{1}{n} = w$, and c_2 w.p. $\frac{n-1}{n} = 1 - w$, thus $\text{WRD}(S') = \text{WRD}(S)$.

From Theorem 2.4 in [13], $\mathbf{R}_I(\text{WRD}(S'), S') \leq 3 - \frac{2}{n} = 3 - 2w$, and this bound is tight. For general w, the proof is a minor variation of the proof in [13]. ■

Proof of Lemma 13. We denote by $c^* = c^*(S)$ the optimal classifier for S. If $c^* = c'$ we are done, therefore assume they differ. Let $B \subseteq X$ the points on which c^*, c' disagree. The worst case is when agent 1 completely agrees with c^*, i.e., when $\mathbf{R}_1(c^*, S) = 0$ (and in particular $c_1 = c^*$). Otherwise we can increase $\frac{\mathbf{R}_I(c',S)}{\mathbf{R}_I(c^*,S)}$ by removing all data points on which they do not agree. We can similarly assume that both classifiers make no errors on $X \setminus B$ (since this will only improve the approximation).

Denote by $r(c, A), r'(c, A)$ the fraction of errors on points from $A \subset X$ according to $\mathbf{w}, \mathbf{w}'$ (in particular $r(c, X) = \mathbf{R}_I(c, S)$ and $r'(c, X) = \mathbf{R}_{I'}(c, S')$. Note that c' must also agree with c_1 on all points outside B, thus

$$r(c', X) \leq (1 - w_1)r'(c', X) + w_1 \frac{|B|}{|X|}. \tag{12}$$

Also, since c^*, c' disagree on all B, we have that

$$\frac{|B|}{|X|} r'(c', B) = r'(c', X) \leq r'(c^*, X) = \frac{|B|}{|X|} r'(c^*, B) \tag{13}$$

$$r'(c^*, B) \geq \frac{1}{2} \geq r'(c', B) \tag{14}$$

$$r(c^*, X) = (1 - w_1)r'(c^*, X) \tag{15}$$

Therefore,

$$\begin{aligned}
r(c^*, X) &= (1 - w_1)\frac{|B|}{|X|} r'(c^*, B) && \text{(from (13),(15))} \\
&\geq (1 - w_1)\frac{|B|}{|X|}\frac{1}{2} && \text{(from (14))} \\
r(c', X) &\leq (1 - w_1)r'(c^*, X) + w_1 \frac{|B|}{|X|} && \text{(from (12),(13))} \\
&\leq (1 - w_1)r'(c^*, X) + w_1 \frac{2r(c^*, X)}{1 - w_1} \\
&= r(c^*, X) + \frac{2w_1 r(c^*, X)}{1 - w_1} && \text{(from (15))} \\
&= r(c^*, X)\left(\frac{1 - w_1 + 2w_1}{1 - w_1}\right) = \frac{1 + w_1}{1 - w_1} r(c^*(X)),
\end{aligned}$$

as required. □

Proof of Lemma 17. Suppose at first that $c^* \in \{c_1, c_2\}$. In this case we can effectively narrow our concept class to $\mathcal{C}' = \{c_1, c_2\}$, i.e., it is of size two. Now remove from X all data points on which the two selected concepts agree, i.e., $X' = \{x \in X : c_1(x) \neq c_2(x)\}$. Clearly this can only increase the approximation ratio, as it accentuates the errors caused by selecting the wrong classifier. Note that now both classifiers disagree on all data points; thus, we can take another step in simplifying our scenario, by renaming labels and classifiers so that $c_1(x) = c_+(x) =$"+"; $c_2(x) = c_-(x) =$"-" for all data points $x \in X$, and we are done.

We now turn to the case where $c^* \notin \{c_1, c_2\}$. We will show how to alter S so it would fit into the restricted setting, while the approximation ratio can only increase.

Let $B \subseteq X$ be all data points on which $Y_2(x) \neq c^*(x)$ (recall that $w_2 \geq w_1$). We now create a new dataset $\hat{S}$, in which the labels of agent 2 for B are flipped, i.e., $\hat{Y}_2(x) = c^*(x)$ for all x. In the new dataset $\hat{S}$, c^* is the best concept for agent 2, and thus $c^*(\hat{S}) = c_2(\hat{S})$. From the previous case we have that $\mathbf{R}_I(\mathcal{M}(\hat{S}), \hat{S}) \leq L\mathbf{R}_I(c^*, \hat{S})$. Denote by $r(c), \hat{r}(c)$ the risk of c on $S, \hat{S}$, respectively. Since Y_1 remains unchanged, $\hat{r}_1(c) = r_1(c)$.

Suppose that $\mathcal{M}(S)$ returns c_1, c_2 w.p. p_1, p_2. Then on $\mathcal{M}(\hat{S})$ has the same probabilities (weights are unchanged), except c^* is returned instead of c_2, as this is the best classifier for agent 2 in the new dataset.

Note that Y_2, c^* disagree on at most $|B|$ points, as otherwise agent 2 would have originally preferred c^* over c_2. Thus $r_2(c_2) \leq |B|$. Also, in the new dataset we remove these $|B|$ errors, thus $\hat{r}_2(c^*) = 0, r_2(c^*) = |B|$.

$$\begin{aligned}
\hat{r}(c^*) &= w_1\hat{r}_1(c^*) + w_2 0 && (16) \\
&= w_1 r_1(c^*) + w_2\left(r_2(c^*) - |B|\right) = r(c^*) - w_2|B| \\
r(c_2) &\leq w_1 r_1(c_2) + w_2 r_2(c_2) && (17) \\
&\leq w_1\left(r_1(c^*) + |B|\right) + w_2|B| = |B| + w_1 r_1(c^*) \\
&= |B| + w_1\hat{r}_1(c^*) + w_2\hat{r}_2(c^*) = \hat{r}(c^*) + |B| \\
\hat{r}(c_1) &= w_1\hat{r}_1(c_1) + w_2\hat{r}_2(c_1) && (18) \\
&\geq w_1 r_1(c_1) + w_2\left(r_2(c_1) - |B|\right) = r(c_1) - w_2|B|
\end{aligned}$$

Finally:

$$\begin{aligned}
r(\mathcal{M}(S)) &= p_1 r(c_1) + p_2 r(c_2) && \text{(from (17),(18):)} \\
&\leq p_1\left(\hat{r}(c_1) + w_2|B|\right) + p_2\left(\hat{r}(c^*) + |B|\right) \\
&= p_1\hat{r}(c_1) + p_2\hat{r}(c^*) + |B|(p_1 w_2 + p_2) \\
&= \hat{r}(\mathcal{M}(\hat{S})) + |B|(p_1 w_2 + p_2(w_1 + w_2)) \\
&\leq L\hat{r}(c^*) + 2w_2|B| \leq L(\hat{r}(c^*) + w_2|B|) && (w_2 \geq w_1) \\
&= L \cdot r(c^*), && \text{(from (16))}
\end{aligned}$$

which means $\mathbf{R}_I(\mathcal{M}(S), S) \leq L \cdot \mathbf{R}_I(c^*, S)$, as required. □

Proof of Theorem 19.

LEMMA 20. $\alpha_{\mathbf{w}} \leq 2 - \frac{2}{n}$.

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

$$\frac{1}{n} \leq \sum_{i \in I} w_i^2 \leq 1. \tag{19}$$

$$\begin{aligned}
(\alpha_{\mathbf{w}})^{-1} &= \sum_{i \in I} p'_i = \sum_{i \in I} w_i \frac{1}{2 - 2w_i} = \sum_{i \in I} w_i g(w_i) \\
&\geq g\left(\sum_{i \in I} w_i \cdot w_i\right) = \frac{1}{2 - 2\sum_{i \in I} w_i^2} \\
&\qquad \text{(from Jensen's inequality)} \\
&\geq \frac{1}{2 - 2(1/n)}, && \text{(from (19))}
\end{aligned}$$

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

Let $\mathcal{F}$ be the set of all labeling functions $f : X \to \{-, +\}$. In particular $\mathcal{C} \subseteq \mathcal{F}$. We denote by $d(f, f')$ the number of disagreements between f and f'. d is a pseudo-metric, and thus symmetric and satisfies the triangle inequality (T.I.) (see [13] for more details).

f_i, c_i denote the labels of agent i (i.e., $f_i \equiv Y_i$), and the classifier in $\mathcal{C}$ that is the closest to them (i.e., $c \in \mathcal{C}$ that minimizes $d(c, f_i)$). For any c, it holds that

$$\mathbf{R}_I(c, S) = \sum_{i \in I} w_i \mathbf{R}_i(c, S) = \sum_{i \in I} w_i d(c, f_i)$$

.

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

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

■