Incentive Incompatibility of Logistic Regression

Abstract

We study the incentive compatibility of multi-class logistic regression. We provide a numerical example in which a strategic data provider has the incentive to misreport her private label to increase the classification probability of her true label. In particular, the model trained given her true label classifies her data point incorrectly, whereas the model trained given her misreported label classifies her data point correctly. We show that this incentive incompatibility disappears for classifiers that satisfy a monotonicity condition and independence of irrelevant alternatives condition. Examples of such classifiers include Bayes classifiers, kernel density estimators, and empirical risk minimization classifiers with zero-one loss.

1 Introduction

Consider an insurance company that makes its pricing decisions based on the customers' public observable characteristics, but the decision models are built using the information on private unobservable characteristics the customers' report. If the insurance company is transparent about its models, they might worry that customers have incentives to misreport their private information to get the contract that is the most beneficial to them. Similar examples include other rating systems that depend on the report of private information, such as loan applications, school grades, and employee screening.

In a general mechanism design problem, each of many strategic agents owns one public data point and reports her private label to the principal. The principal is the learner and builds a classifier based on the labels provided by the agents. Each agent chooses a label to report, not necessarily her true label, to maximize the probability that her data point is classified correctly by the principal. We say that a dataset is incentive-incompatible for the classifier if at least one of the agents has the incentive to misreport, and we characterize classifiers that are incentive-compatible with all possible datasets.

We start with an example dataset that is incentive-incompatible for the multi-class logistic regression classifier. In the dataset, each of the 18 agents owns a two-dimensional data point and a private label with one of three values: "red", "green", or "blue".

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

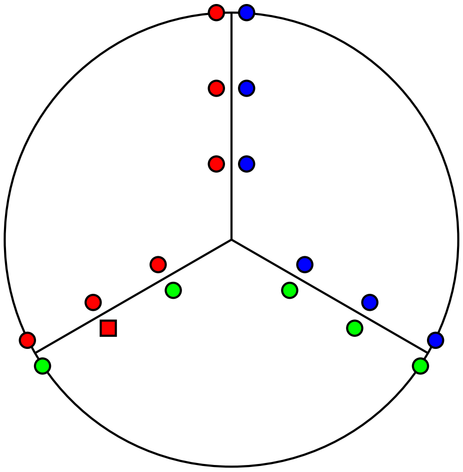

Figure 1: Incentive-incompatible Example

The 18 points are located inside a unit circle, and each point is 0.004 away from the three line segments through the origin that forms angles of 120 degrees between them. There is one red point, drawn as a square in the diagram, that is on the "incorrect" side of the boundary. For the agent represented by the red square, truthfully reporting her label will lead to a multi-class logistic regression model that classifies her point as "green". The probability that this model classifies her point as "red" is 0.3290. However, if the agent misreports her label as "blue", the resulting model classifies her point as "red", and with a probability of 0.4966. By lying about her label, the agent can make the principal learn an incorrect model that classifies her point correctly and with a higher probability.

The example provides insight into the general incentive incompatibility issue of many classification models used in machine learning. The "red" agent that gets incorrectly classified as "green", does not want to misreport her label as "green" but instead has the incentive to misreport her label as the third alternative "blue" to influence the classifier in a way that changes the decision between "red" and "green". Intuitively, if there are only two classes, then the agent would not be willing to misreport her label if the classifier is monotonic, in the sense that adding a point from one class increases the probability that this point is classified as a member of that class; and if there are more than two classes, then the agent would not be willing to misreport her label if the classifier is independent of irrelevant alternatives, in the sense that adding a point from a third class would not affect the decisions between the two classes.

Previous work on mechanism design for machine learning with strategic data sources focuses on designing robust algorithms to incentivize the data providers to report their private data truthfully. Their models mainly differ in the objective and the possible actions of the data providers (agents) and the learner (principal).

The first group of papers focuses on principal-agent problems similar to our paper in which each agent's private data point is the agent's type that the agent cannot change. The only action the agents can take is whether to report their private information truthfully.

1. Some models assume the agents' data points (or feature vectors) are public, but their labels are private. Perote and Perote-Pena (2004), Chen, Podimata, Procaccia, and Shah (2018), and Gast, Ioannidis, Loiseau, and Roussillon (2013) focus on strategy-proof linear regression algorithms and introduced clockwise repeated median estimators, generalized resistant hyperplane estimators, and modified generalized linear squares estimators. Dekel, Fischer, and Procaccia (2010) investigates the general regression problem with empirical risk minimization and absolute value loss. All the previously mentioned papers assume the labels are continuous variables (regression problems), and Meir, Procaccia, and Rosenschein (2012) assumes the labels are discrete variables (classification problems) and proposes a class of random dictator mechanisms.

2. Some models assume the agents' data points are also private. Chen, Liu, and Podimata (2019) investigates such problems for linear regressions.
3. Other models do not involve labels. Each agent has a private valuation. These problems are usually modeled as facility location problems and the solution involves some variant of the Vickrey-Clarke-Groves or Meyerson auction. They include Dütting, Feng, Narasimhan, Parkes, and Ravindranath (2017), Golowich, Narasimhan, and Parkes (2018), Epasto, Mahdian, Mirrokni, and Zuo (2018), and Procaccia and Tennenholtz (2009).

The second group papers focus on moral-hazard problems in which each agent does not have a type but they can choose an action (with a cost) that affects the probability of obtaining the correct label. Richardson, Rokvic, Filos-Ratsikas, and Faltings (2019) focuses on the linear regression problem in this scenario, and Cai, Daskalakis, and Papadimitriou (2015) and Shah and Zhou (2016) investigates the problem for more general machine learning problems. Mihailescu and Teo (2010) also discusses a similar problem for general machine learning algorithms.

The last group of papers uses machine learning or robust statistics techniques without game-theoretic models. This group of papers include Dekel and Shamir (2009b), Dekel and Shamir (2009a).

2 Logistic Regression

2.1 Model

In this section, we introduce the model using logistic regression as an example. We assume the principal is training a multi-class logistic (softmax) regression. There are n strategic agents each providing the label of one data point to the principal. An agent, i, with public $x_i \in \mathbb{R}^m$, and private discrete $y_i \in \{1, 2, ..., k\}$, has the objective of maximizing the probability that her data point is labeled correctly by the principal's classifier, parameterized by the $m \times (k+1)$ weights (and bias) matrix w. The agent can choose to report $y_i^\dagger$ to achieve the objective, with possibly $y_i^\dagger \neq y_i$. Denoting the weights of the model resulting from the false report from agent i by $w^\star\left(y_i^\dagger\right)$ (and $w^\dagger$ when the identity and the report of the agent is unambiguous), the agent's objective can be written as,

$$\max_{y^\dagger \in \{1,2,...,k\}} \mathbb{P}\left\{Y = y_i | x_i; w^\star\left(y_i^\dagger\right)\right\},$$

where,

$$\mathbb{P}\left\{Y = c | x_i; w\right\} = \frac{e^{z_{i,c}}}{\sum\limits_{c'=1}^{k} e^{z_{i,c'}}},$$

$$z_{i,c} = \sum_{j=1}^{m} w_{j,c} x_{i,j} + b_c, \text{ for } c \in \{1, 2, ..., k\}.$$

The principal is not strategic and he maximizes the likelihood of the data,

$$\max_w \sum_{i=1}^{n} \log\left(\mathbb{P}\left\{Y = y_i^\dagger | x_i; w\right\}\right).$$

We consider the case without a coalition of a group of agents, so only one agent is misreporting at a time, and use the following notations,

$$w^\star = \arg\max_w \sum_{i=1}^{n} \log\left(\mathbb{P}\left\{Y = y_i | x_i; w\right\}\right), \text{ and}$$

$$w^\dagger = w^\star\left(y_i^\dagger\right) = \arg\max_w \log\left(\mathbb{P}\left\{Y = y_i^\dagger | x_i; w\right) + \sum_{i'=0, i' \neq i}^{n} \log\left(\mathbb{P}\left\{Y = y_{i'} | x_{i'}; w\right\}\right).$$

Definition 1. A dataset is incentive-incompatible for a classifier if there exists at least one agent i, and some $y_i^\dagger \neq y_i$ such that,

$$\mathbb{P}\left\{Y = y_i | x_i; w^\star\right\} < \mathbb{P}\left\{Y = y_i | x_i; w^\star\left(y_i^\dagger\right)\right\}.$$

A classifier is incentive-compatible if there does not exist a dataset that is incentive-incompatible for the classifier.

Conjecture 1. *Multi-class logistic regression is not incentive-compatible.*

The example described in Figure 1 is a dataset that is numerically incentive-incompatible.
In this example, agent i reports $x_i \in \mathbb{R}^2$ and y_i is one of 1 (red), 2 (green), or 3 (blue). Suppose the red square point corresponds to agent 1 with $x_1 = (-1.63, -1.17)$ and $y_1 = 1$, then

$$\mathbb{P}\left\{Y = 1 | x_1; w^\star\right\} = 0.3290,$$
$$\mathbb{P}\left\{Y = 1 | x_1; w^\star\left(y_1^\dagger = 3\right)\right\} = 0.4966.$$

Here, parameter estimation is done using maximum likelihood estimation with BFGS, and $w^\star$ is given by, with class 1 weights normalized to 0,

Class	(Intercept)	x1	x2
1	0	0	0
2	-0.6053178	104.9925	-181.3391914
3	-0.2852057	209.4190	0.3656777

and $w^\star\left(y_1^\dagger = 3\right)$ is given by,

Class	(Intercept)	x1	x2
1	0	0	0
2	-0.1915645	3.473426	-5.507418
3	0.8273350	4.309293	-1.200060

Currently, there is no formal proof that the result is not due to numerical instability, therefore, @(logit) is stated as a conjecture. Numerical experiments indicate that incentive-incompatible datasets are rare. If the data points are two-dimensional and standard normally distributed, and the labels are created using randomly generated weights with a small probability of error, then such a dataset is incentive-incompatible with a probability of 0.005.

2.2 Loss Functions

It is, however, possible to change the loss function so that logistic regression is incentive-compatible. Changing the loss function to absolute value L^1 loss is one possibility, due to Dekel, Fischer, and Procaccia (2010). Their result on incentive compatibility of empirical risk minimization in the regression setting is applicable in our model. In addition to absolute value loss, empirical risk minimizers with zero-one loss are also always incentive-compatible.

Proposition 1. *Multi-class probabilistic empirical risk minimization classifiers with absolute value L^1 loss are incentive-compatible.*

Proof. The proof is adapted from the proof of Theorem 4.1 in Dekel, Fischer, and Procaccia (2010). For any dataset $\left\{(x_i, y_i)\right\}_{i=1}^n$, and the convex hypothesis class $\mathcal{H}$, with the notation $h_k(x)$ representing the probability that the point x is classified as k, let the optimal classifier be,

$$
\begin{aligned}
h^{\star} &= \arg\min_{h\in\mathcal{H}} \sum_{i'=1}^{n} \ell\left(y_{i'}; h\right) \\
&= \arg\min_{h\in\mathcal{H}} \sum_{i'=1}^{n} \left|1 - h_{y_{i'}}\left(x_{i'}\right)\right|,
\end{aligned}
$$

and for any fixed agent i,

$$
h^{\dagger} = \arg\min_{h\in\mathcal{H}} \sum_{i'\neq i}^{n} \ell\left(y_{i'}; h\right) + \ell\left(y_i^{\dagger}; h\right).
$$

For a contradiction, assume $\ell\left(y_i; h^{\dagger}\right) \leqslant \ell\left(y_i; h^{\star}\right)$, we show that $h^{\dagger} = h^{\star}$.
Since ℓ is the absolute value loss,

$$
h_{y_i}^{\star}\left(x_i\right) \leqslant h_{y_i}^{\dagger}\left(x_i\right) \leqslant 1.
$$

Now define $\alpha \in (0, 1]$ by,

$$
\alpha = \frac{1 - h_{y_i}^{\star}\left(x_i\right)}{h_{y_i}^{\dagger}\left(x_i\right) - h_{y_i}^{\star}\left(x_i\right)},
$$

and define $h^{\alpha} \in \mathcal{H}$, since $\mathcal{H}$ is convex, by,

$$
h^{\alpha} = \alpha h^{\dagger} + (1-\alpha) h^{\star}.
$$

Then, we have,

$$
\ell\left(y_i^{\dagger}; h^{\star}\right) - \ell\left(y_i; h^{\star}\right) = \ell\left(y_i^{\dagger}; h^{\alpha}\right) - \ell\left(y_i; h^{\alpha}\right).
$$

Note that the above equality holds for other agents too since they report truthfully, so the equality holds for the objective too. Therefore, $h^{\star}$ and $h^{\dagger}$ are both empirical risk minimizers on the original dataset.

□

Proposition 2. *Multi-class deterministic empirical risk minimization classifiers with zero-one loss are incentive-compatible.*

Proof. For any dataset $\left\{\left(x_i, y_i\right)\right\}_{i=1}^{n}$, and the hypothesis class $\mathcal{H}$, let the optimal classifier be,

$$
h^{\star} = \arg\min_{h\in\mathcal{H}} \sum_{i'=1}^{n} \mathbb{1}_{\left\{y_{i'} \neq h\left(x_{i'}\right)\right\}}.
$$

Fix an agent i with x_i, and fix the other agents' reports, $\left(x_{-i}, y_{-i}\right)$, define the loss function given the classifier h and report of agent $i, y_i^{\dagger}$, as,

$$
\ell\left(y_i^{\dagger}; h\right) = \sum_{i'\neq i} \mathbb{1}_{\left\{y_{i'} \neq h\left(x_{i'}\right)\right\}} + \mathbb{1}_{\left\{y_i^{\dagger} \neq h\left(x_i\right)\right\}}.
$$

If $y_i = h^{\star}\left(x_i\right)$, then the classifier is already classifying x_i correctly, misreporting will not improve the outcome for i. Assume the prediction is $h^{\star}\left(x_i\right) = y^{\star} \neq y_i$, and suppose $h^{\star}$ is making q mistakes, meaning,

$$
q = \min_{h\in\mathcal{H}} \ell\left(y_i; h^{\star}\right).
$$

Agent i can misreport in the following two ways: □

1. If agent i reports $y_i^{\dagger} = y^{\star}$, let the new classifier be $h^{\dagger}$, note that we must have,

$$
\ell\left(y^{\star}; h^{\dagger}\right) \leqslant q - 1,
$$

since $\ell\left(y^{\star}; h^{\dagger}\right) > q - 1 = \ell\left(y^{\star}; h^{\star}\right)$ contradicts the optimality of $h^{\dagger \cdot}$
Suppose that agent i could get her true label with $h^{\dagger}$, meaning $h^{\dagger}\left(x_i\right) = y_i$, then,

$$\begin{aligned} \ell\left(y_i; h^{\dagger}\right) &= \ell\left(y^{\star}; h^{\dagger}\right) - 1 \\ &\leqslant k - 2 \\ &< \ell\left(y_i; h^{\star}\right), \end{aligned}$$

which contradicts the optimality of $h^{\star}$. Therefore, agent i cannot improve the outcome by misreporting $y^{\star}$.

2. If agent i reports $y_i^{\dagger} = y' \neq y^{\star}$, let the new classifier be $h^{\dagger}$, note that we must have,

$$\ell\left(y'; h^{\dagger}\right) \leqslant q,$$

since $\ell\left(y'; h^{\dagger}\right) > q = \ell\left(y'; h^{\star}\right)$ contradicts the optimality of $h^{\dagger \cdot}$
Suppose that agent i could get her true label with $h^{\dagger}$, then,

$$\begin{aligned} \ell\left(y_i; h^{\dagger}\right) &= \ell\left(y_i; h^{\dagger}\right) - 1 \\ &\leqslant q - 1 \\ &< \ell\left(y_i; h^{\star}\right), \end{aligned}$$

which contradicts the optimality of $h^{\star}$. Therefore, agent i cannot improve the outcome by misreporting y'.
Therefore, no agent can improve the outcome and the dataset is incentive-compatible.

2.3 Binary Logistic Regression

Binary logistic regression is always incentive-compatible, and agents with one label do not want to pretend to have the other label. This is not true in general for binary empirical risk minimization classifiers, and additional normalization condition on the loss function is required. The result is discussed in the next section. In the special case of logistic regression, since $\ell\left(0; h\right) + \ell\left(1; h\right) = 1$ holds for any $h \in \mathcal{H}$, the following Proposition holds.

Proposition 3. *Binary logistic classifiers are incentive-compatible.*

Proof. For any dataset $\left\{\left(x_i, y_i\right)\right\}_{i=1}^{n}$, and the logistic hypothesis class $\mathcal{H}$, let the optimal classifier in the case every agent reports truthfully be,

$$h^{\star} = \arg\min_{h \in \mathcal{H}} \sum_{i'=1}^{n} \ell\left(y_{i'}; h\right).$$

Fix an agent i with x_i, and fix other agents' reports, $\left(x_{-i}, y_{-i}\right)$, define the optimal classifier given the misreport of agent $i, y_i^{\dagger} = 1 - y_i$ as,

$$h^{\dagger} = h^{\star}\left(y_i^{\dagger}\right) = \arg\min_{h \in \mathcal{H}} \sum_{i'=1, i' \neq i}^{n} \ell\left(y_{i'}; h\right) + \ell\left(y_i^{\dagger}; h\right).$$

Now suppose, for a contradiction, that agent i prefers misreporting,

$$\ell\left(y_i; h^{\star}\right) > \ell\left(y_i; h^{\dagger}\right),$$

which implies, since $y_i^{\dagger} = 1 - y_i$, and the special property of the logistic loss function $\ell\left(0; h\right) + \ell\left(1; h\right) = 1$,

$$\ell\left(y_i^{\dagger}; h^{\star}\right) < \ell\left(y_i^{\dagger}; h^{\dagger}\right).$$

Note that the above implication only works for binary classification.
Due to the optimality of $h^\dagger$,

$$\sum_{i'=1, i' \neq i}^{n} \ell\left(y_{i'}; h^\dagger\right) + \ell\left(y_i^\dagger; h^\dagger\right) \leqslant \sum_{i'=1, i' \neq i}^{n} \ell\left(y_{i'}; h^\star\right) + \ell\left(y_i^\dagger; h^\star\right),$$

using the above inequalities, the comparison can be simplified to,

$$\sum_{i'=1, i' \neq i}^{n} \ell\left(y_{i'}; h^\dagger\right) \leqslant \sum_{i'=1, i' \neq i}^{n} \ell\left(y_{i'}; h^\star\right),$$

$$\sum_{i'=1}^{n} \ell\left(y_{i'}; h^\star\right) \leqslant \sum_{i'=1}^{n} \ell\left(y_{i'}; h^\star\right),$$

which is a contradiction to the optimality of $h^\star$.

□

3 Other Classifiers

In this section, we show that classifiers, including Bayes classifiers and kernel density estimators, that satisfy some separability conditions are always incentive-compatible. One intuition behind why some classifiers are incentive-incompatible is that one-vs-one classification decisions are not made independently. Logistic regression has highly interdependent one-vs-one decisions. The following example is one in which 1-vs-2 decisions are completely determined by the 2-vs-3 decisions, and as a result, a class-1 point that is misclassified as class-2 could misreport as class-3 to influence the 2-vs-3 decision boundary and indirectly change the 1-vs-2 decision boundary in its favor.

3.1 A Good Example

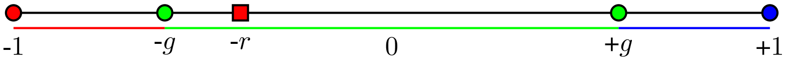

Figure 2: $1D$ Artificial Incentive-incompatible Example 1 (Truthful)

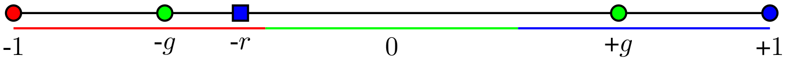

Figure 3: $1D$ Artificial Incentive-incompatible Example 1 (Misreport)

Consider a 3-way classifier $h_t : \mathbb{R} \to \{ \text{ red }, \text{ green }, \text{ blue } \}$ parametrized by $t \geqslant 0$:

$$h_t(x) = \begin{cases} \text{red} & \text{if } x < t \\ \text{green} & \text{if } -t \leqslant x \leqslant t \\ \text{blue} & \text{if } x > t. \end{cases} \tag{1}$$

Let the hypothesis space be

$$\mathcal{H} = \{h_t : t \geqslant 0\}. \tag{2}$$

Equivalently, a hypothesis h_t partitions $\mathbb{R}$ into three sets: $X_t^{\text{red}} = (-\infty, -t)$, $X_t^{\text{green}} = [-t, t]$, $X_t^{\text{blue}} = (t, \infty)$. Given a labeled point (x, y) with $y \in \{ \text{ red }, \text{ green }, \text{ blue } \}$, it could be outside the "color region" suggested by h_t. Accordingly, we define a loss function ℓ based on the distance it takes to move the point to the corresponding color region suggested by h_t. Concretely,

$$\ell(x, y, h_t) = f(d(x, X_t^y)) \tag{3}$$

where

$$d(x, X_t^y) = \min_{x' \in X_t^y} \|x - x'\| \tag{4}$$

is the shortest distance from the point x to the set (color region) X_t^y. $f \geqslant 0$ is strictly convex and continuously differentiable with a minimum of 0 at 0: $f(0) = 0$. For example, f can be the square function $f(z) = z^2$. To be concrete, for a fixed t,

$$\ell\left(x, y = \text{ red }, h_t\right) = \begin{cases} 0 & \text{if } x \leqslant -t \\ f(x+t) & \text{if } x > -t \end{cases}, \tag{5}$$

Similarly, we have,

$$\ell\left(x, y = \text{ green }, h_t\right) = \begin{cases} f(t-x) & \text{if } x \leqslant -t \\ 0 & \text{if } -t < x < t \\ f(x-t) & \text{if } x \geqslant t \end{cases}, \tag{6}$$

and,

$$\ell\left(x, y = \text{ blue }, h_t\right) = \begin{cases} f(t-x) & \text{if } x \leqslant t \\ 0 & \text{if } x > t \end{cases}. \tag{7}$$

Also note that strict convexity implies continuity but not differentiability. For example, $f(t) = t^2 + |t|$ is strictly convex but not differentiable at $t = 0$. Given a training set $S = \{(x_1, y_1), \ldots (x_n, y_n)\}$, consider the Empiricial Risk Minimizer (ERM)

$$\hat{h} \in \text{argmin}_{h \in \mathcal{H}} \sum_{i=1}^{n} \ell(x_i, y_i, h). \tag{8}$$

We define R as the risk, the objective function in the above minimization,

$$R\left(S, h\right) = \sum_{i=1}^{n} \ell\left(x_i, y_i, h\right).$$

We now exhibit a family of IIC datasets S parameterized by g and r, see Figure 2. S consists of five labeled points:

$$S\left(g, r\right) = \left\{(x_i, y_i)\right\}_{i=1}^{5} = \left\{\left(-1, \text{ red }\right), \left(-g, \text{ green }\right), \left(-r, \text{ red }\right), \left(g, \text{ green }\right), \left(1, \text{ blue }\right)\right\}, \tag{9}$$

with $0 < r < g < 1$.

Proposition 4. *For any $g \in (0, 1]$, there exists an $r \in (0, g)$ such that the dataset $S\left(g, r\right)$ is incentive incompatible with respect to ERM on $\mathcal{H}$ and ℓ.*

Proof. The optimal threshold for the original dataset S is,

$$\hat{t} = \arg\min_{t \geqslant 0} R\left(S, h_t\right) = \arg\min_{t \geqslant 0} \begin{cases} 2f\left(g - t\right) & \text{if } t \leqslant r \\ 2f\left(g - t\right) + f\left(t - r\right) & \text{if } r < t < g \\ f(t - r) & \text{if } t \geqslant g \end{cases}, \tag{10}$$

Since f is strictly convex thus continuous in t on [-1, 1], any linear combination with positive coefficients is also strictly convex and continuous in t. In addition, due to the assumption that $f\left(0\right) = 0$,

$$R\left(S, h_t | t = r\right) = 2f\left(g - r\right) = \lim_{t \to r^+} R\left(s, h_t\right), \tag{11}$$

and,

$$R\left(S, h_t | t = g\right) = f\left(g - r\right) = \lim_{t \to g^-} R\left(s, h_t\right), \tag{12}$$

implying that $R\left(S, h_t\right)$ is continuous in t.
Given the assumption that f is minimized at 0, we have $2f\left(t - r\right)$ is minimized at $t = r < g$ and $f\left(t - g\right)$ is minimized at $t = g > r$. Therefore, the minimum of $R\left(S, h_t\right)$ occurs in the region $r < t < g$, meaning,

$$\hat{t} \in \left(r, g\right), \tag{13}$$

which classifies (-r, red) incorrectly as a green point.

Now, let $S^\dagger$ be the dataset in which all agents except for (-r, red) report truthfully, and the agent (-r, red) misreports her label as blue. In this case, the optimal threshold for $S^\dagger$ is,

$$\hat{t}\dagger = \arg\min_{t\geqslant 0} R\left(S^\dagger, h_t\right) = \arg\min_{t\geqslant 0} \begin{cases} 2f\left(g-t\right)+f\left(t+r\right) & \text{if } t<r \\ 2f\left(g-t\right)+f\left(t+r\right) & \text{if } r\leqslant t<g \\ f(t+r) & \text{if } t\geqslant g \end{cases}, \tag{14}$$

Since $f\left(t+r\right)$ is minimized at $t=-r<g$, the minimum of $R\left(S^\dagger, h_t\right)$ occurs in the region $r<g$,

$$\hat{t}\dagger \in \left(0, g\right), \tag{15}$$

and we have, in this range,

$$R\left(S^\dagger, h_t\right) = 2f\left(g-t\right)+f\left(t+r\right). \tag{16}$$

Due to strict convexity of f, we have $f'\left(a\right)$ is strictly increasing in a. Given that $f'\left(0\right)=0$ at the global minimum, we can find small $\delta>0$ such that the following holds

$$2f'\left(\delta\right) < f'\left(2r\right). \tag{17}$$

Now, take $r=g-\delta$, we have, for $t\in\left(0,g\right)$,

$$\frac{\partial R\left(S, h_t\right)}{\partial t} = -2f'\left(g-r\right)+f'\left(2r\right) > 0. \tag{18}$$

Therefore, the objective is decreasing at $t=r$, and given that it is the sum of two strictly convex functions thus strictly convex itself, we have,

$$\hat{t}\dagger \in \left(0, r\right), \tag{19}$$

which classifies (-r, red) correctly as a red point.

Therefore, the point has the incentive to misreport and the dataset is incentive incompatible. □

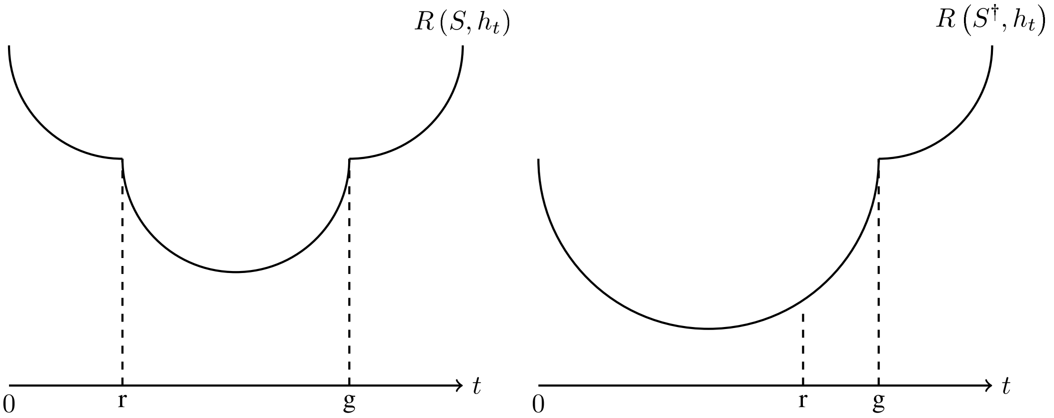

Figure 4: Risk on S vs $S^\dagger$

Example 1. When f is the square function, the loss is the square of the distance it takes to move the point to the correct region. In this case, $S\left(g,r\right)$ is IIC as long as $g<2r$, which is consistent with the condition in Proposition 4.

In particular, the optimal threshold for the original dataset is,

$$\hat{t} = \frac{2}{3}g+\frac{1}{3}r \in \left(r,g\right),$$

which classifies (-r, red) incorrectly as a green point, but the optimal threshold if the agent misreports her label as blue is,

$$\hat{t}\dagger = \frac{2}{3}g-\frac{1}{3}r \in \left(0,r\right),$$

which classifies (-r, red) correctly as a red point.

3.2 Another Example

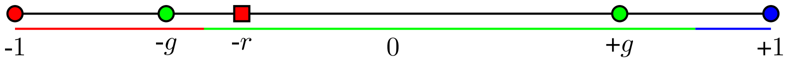

Figure 5: $1D$ Artificial Incentive-incompatible Example 2 (Truthful)

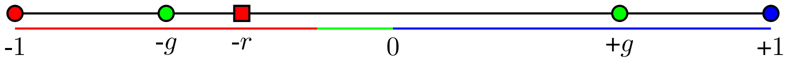

Figure 6: $1D$ Artificial Incentive-incompatible Example 2 (Misreport)

Intuitively, in the previous example, when the misclassified red point pretends to be a blue point with a large loss, the classifier modifies the green-vs-blue decision boundary to minimize that loss, and the point benefits from the interdependence between the green-vs-blue decision and the green-vs-red decision.

The decision boundaries can be related in more complicated ways. Consider another 3-way classifier $h_{a,b} : \mathbb{R} \to \{ \text{ red }, \text{ green }, \text{ blue } \}$ parameterized by two thresholds a and b.

$$h_{a,b}(x) = \begin{cases} \text{red} & \text{if } x < a \\ \text{green} & \text{if } a \leqslant x \leqslant b \\ \text{blue} & \text{if } x > b \end{cases} \tag{20}$$

Let the hypothesis space be,

$$\mathcal{H}' = \{h_{a,b} : -1 < a < b < 1\} \tag{21}$$

Here, a hypothesis $h'_{a,b}$ partitions [-1, 1] into three sets: $X^{\text{red}}_{a,b} = [-1, a), X^{\text{green}}_{a,b} = [a, b], X^{\text{blue}}_{a,b} = (b, 1]$. This time, we define a loss function ℓ based on the distance from the point to the center of the corresponding color region suggested by $h_{a,b}$. This leads to a classifier similar to k-means clustering procedure,

$$\ell'(x, y, h_{a,b}) = \left\| x - \bar{X}^{y}_{a,b} \right\|^{2,} \tag{22}$$

where $\bar{X}^{\text{red}}_{a,b} = \frac{-1+a}{2}, \bar{X}^{\text{green}}_{a,b} = \frac{a+b}{2}$, and $\bar{X}^{\text{blue}}_{a,b} = \frac{b+1}{2}$ are the centers of the decision regions.

Proposition 5. *For any $g > \frac{1}{3}$, there exists $r \in \left(\frac{1}{3}, g\right)$ such that the dataset $S(g, r)$ is incentive incompatible with respect to ERM on $\mathcal{H}'$ and ℓ'.*

Proof. The optimal thresholds for the original dataset is, □

$$a^\star = -\frac{3}{4}r - \frac{1}{4}, b^\star = \frac{1}{2}r + \frac{1}{2}, \tag{23}$$

and since $r > \frac{1}{3}$, we have,

$$a^\star < -\frac{1}{2} < -\frac{1}{3}, b^\star > 0, \tag{24}$$

which classifies (-r, red) incorrectly as a green point.
Similarly, the optimal threshold if the agent misreports her label as blue is,

$$a^\star = \frac{1}{2}r - \frac{1}{2}, b^\star = -\frac{3}{4}r + \frac{1}{4}, \tag{25}$$

and since $r > \frac{1}{3}$, we have,

$$a^{\star} > -\frac{1}{3}, b^{\star} > 0, \tag{26}$$

which classifies (-r, red) correctly as a red point.
Therefore, the point has the incentive to misreport and the dataset is incentive incompatible.

In this case with $\mathcal{H}'$ and ℓ', when the misclassified red point pretends to be a blue point with a large loss, the classifier modifies the green-vs-blue decision boundary to reduce the loss from the new blue point, but it increases the loss from the green point on the right at the same time. As a result, the classifier then modifies the green-vs-red decision boundary to reduce the loss from the green points. Intuitively, the misclassified red point could only directly affect the green-vs-blue decision boundary, however, it could indirectly affect the green-vs-red decision boundary by getting help from the green point.

4 Generalization

To generalize the above observations, suppose the learner is a probabilistic classifier with parameters estimated by maximum likelihood, and the agents report their labels to maximize the classification probability of their true labels. Then the following two conditions guarantee that the classification is incentive-compatible.

Definition 2. (Monotonic Condition) A multi-class probabilistic classifier is monotonic if, given a training set S, for any point x with labels a and b,

$$\frac{\mathbb{P}\left\{Y = a | x; w^{\star}\left(S\right)\right\}}{\mathbb{P}\left\{Y = b | x; w^{\star}\left(S\right)\right\}} \geqslant \frac{\mathbb{P}\left\{Y = a | x; w^{\star}\left(S \cup \left\{\left(x, y = a\right)\right\}\right)\right\}}{\mathbb{P}\left\{Y = b | x; w^{\star}\left(S \cup \left\{\left(x, y = a\right)\right\}\right)\right\}}.$$

The assumption says that the probability that x is classified as a increases when there is an additional point (x, a) in the training set.

Definition 3. (Independence of Irrelevant Alternatives (IIA) Condition) A multi-class classifier is independent of irrelevant alternatives if, given a training set S, for any point x and any pair of labels a and b,

$$\frac{\mathbb{P}\left\{Y = a | x; w^{\star}\left(S\right)\right\}}{\mathbb{P}\left\{Y = b | x; w^{\star}\left(S\right)\right\}} = \frac{\mathbb{P}\left\{Y = a | x; w^{\star}\left(S \cup \left\{\left(x', y' \notin \left\{a, b\right\}\right)\right\}\right)\right\}}{\mathbb{P}\left\{Y = b | x; w^{\star}\left(S \cup \left\{\left(x', y' \notin \left\{a, b\right\}\right)\right\}\right)\right\}}.$$

The assumption says that the ratio between the classification probabilities of x of any two classes is not changed by adding a point at x with a third class.

Combining the two assumptions MC and IIA, we have that an agent with label a cannot change the decision of a vs b by misreporting its label as a third class c. This observation is formalized in the following proposition.

Theorem 1. *A multi-class probabilistic classifier estimated by maximum likelihood is incentive-compatible if it is monotonic and independent of irrelevant alternatives.*

Proof. Fix a dataset $\left\{(x_i, y_i)\right\}_{i=1}^{n}$, let the maximum likelihood estimates in the case every agent report truthfully be,

$$w^{\star} = \arg\max_{w} \sum_{i'=1}^{n} \log\left(\mathbb{P}\left\{Y = y_{i'} | x_{i'}; w\right\}\right).$$

Fix an agent i, her feature vector x_i, and fix other agents' reports, (x_{-i}, y_{-i}), define the maximum likelihood estimate given the misreport of agent $i, y_i^\dagger$ as,

$$w^\dagger = \arg\max_w \sum_{i'=1, i' \neq i}^n \log\left(\mathbb{P}\left\{Y = y_{i'} | x_{i'}, w\right\}\right) + \log\left(\mathbb{P}\left\{Y = y_i^\dagger | x_i; w\right\}\right).$$

Now suppose, for a contradiction, that agent i prefers misreporting, assume the following incentive inequality,

$$\mathbb{P}\left\{Y = y_i | x_i; w^\star\right\} > \mathbb{P}\left\{Y = y_i | x_i; w^\dagger\right\}.$$

If there are only two classes, then by symmetry,

$$\mathbb{P}\left\{Y = y_i^\dagger | x_i; w^\star\right\} < \mathbb{P}\left\{Y = y_i^\dagger | x_i; w^\dagger\right\}.$$

If there are more than two classes, fix a third $y_i' \notin \left\{y_i, y_i^\dagger\right\}$, and define an intermediate maximum likelihood estimate from removing the point (x_i, y_i),

$$w' = \arg\max_w \sum_{i'=1, i' \neq i}^n \log\left(\mathbb{P}\left\{Y = y_{i'} | x_{i'}; w\right\}\right),$$

then the Monotonic Condition implies,

$$\frac{\mathbb{P}\left\{Y = y_i | x_i; w^\star\right\}}{\mathbb{P}\left\{Y = y_i' | x_i; w^\star\right\}} \leqslant \frac{\mathbb{P}\left\{Y = y_i | x_i; w'\right\}}{\mathbb{P}\left\{Y = y_i' | x_i; w'\right\}},$$

and the IIA Condition implies,

$$\frac{\mathbb{P}\left\{Y = y_i | x_i; w'\right\}}{\mathbb{P}\left\{Y = y_i' | x_i; w'\right\}} = \frac{\mathbb{P}\left\{Y = y_i | x_i; w^\dagger\right\}}{\mathbb{P}\left\{Y = y_i' | x_i; w^\dagger\right\}}.$$

Combining the above two inequalities with the incentive inequality, we have,

$$\mathbb{P}\left\{Y = y_i' | x_i; w^\star\right\} > \mathbb{P}\left\{Y = y_i' | x_i; w^\dagger\right\}.$$

Note that the above inequality is true for all $y_i' \notin \left\{y_i, y_i^\dagger\right\}$, summing over all such y_i' results in,

$$\sum_{y \notin \left\{y_i, y_i^\dagger\right\}} \mathbb{P}\left\{Y = y_i' | x_i; w^\star\right\} > \sum_{y \notin \left\{y_i, y_i^\dagger\right\}} \mathbb{P}\left\{Y = y_i' | x_i; w^\dagger\right\},$$

given that the class probabilities sum up to 1,

$$1 - \mathbb{P}\left\{Y = y_i | x_i; w^\star\right\} - \mathbb{P}\left\{Y = y_i^\dagger | x_i; w^\star\right\} > 1 - \mathbb{P}\left\{Y = y_i | x_i; w^\dagger\right\} - \mathbb{P}\left\{Y = y_i^\dagger | x_i; w^\dagger\right\},$$

and using the incentive inequality again,

$$\mathbb{P}\left\{Y = y_i^\dagger | x_i; w^\star\right\} < \mathbb{P}\left\{Y = y_i^\dagger | x_i; w^\dagger\right\}.$$

Now, due to the optimality of $h^\dagger$,

$$\sum_{i'=1, i' \neq i}^n \log\left(\mathbb{P}\left\{Y = y_{i'} | x_{i'}; w^\dagger\right\}\right) + \log\left(\mathbb{P}\left\{Y = y_i^\dagger | x_i; w^\dagger\right\}\right)$$
$$\leqslant \sum_{i'=1, i' \neq i}^n \log\left(\mathbb{P}\left\{Y = y_{i'} | x_{i'}; w^\star\right\}\right) + \log\left(\mathbb{P}\left\{Y = y_i^\dagger | x_i; w^\star\right\}\right),$$

using the above inequalities, the comparison can be simplified to,

$$\sum_{i'=1, i' \neq i}^n \log\left(\mathbb{P}\left\{Y = y_{i'} | x_{i'}; w^\dagger\right\}\right) < \sum_{i'=1, i' \neq i}^n \log\left(\mathbb{P}\left\{Y = y_{i'} | x_{i'}; w^\star\right\}\right),$$
$$\sum_{i'=1}^n \log\left(\mathbb{P}\left\{Y = y_{i'} | x_{i'}; w^\dagger\right\}\right) < \sum_{i'=1}^n \log\left(\mathbb{P}\left\{Y = y_{i'} | x_{i'}; w^\star\right\}\right),$$

which is a contradiction to the optimality of $w^\star$.

□

Corollary 1. *Binary probabilistic classifiers estimated by maximum likelihood are incentive-compatible.*

Proof. MC holds due to the optimality conditions and IIA holds since there are only two classes.

□

The assumptions Definition 2 (MC) and Definition 3 (IIA) can be significantly simplified for a separable class of the classifiers.

Definition 4. A probabilistic classifier is separable if the parameters $w^\star$ can be partitioned into k classes, $w_1^\star, w_2^\star, ..., w_k^\star$, one set of parameters for each class, such that for given training sets S and S' and any label a if,

$$\{(x_i, y_i) \in S : y_i = a\} = \left\{(x_i, y_i) \in S' : y_i = a\right\},$$

then,

$$w_a^\star(S) = w_a^\star\left(S'\right),$$

and if there is a value function $v_a\left(x; w_a^\star\right)$ that are independent of $w_b^\star, b \neq a$ such that,

$$\mathbb{P}\{Y = y|x; w^\star\} = \frac{v_a\left(x; w_a^\star\right)}{\sum\limits_{b=1}^{K} v_b\left(x; w_b^\star\right)},$$

then the classifier is separable.

Logistic regression satisfies the value function requirement but fails the separability condition since training $w_a^\star$ uses data with labels that are not a. On the other hand, Bayes-type classifiers are separable. For separable classifiers, Definition 2 (MC) and Definition 3 (IIA) are always satisfied.

Corollary 2. *A separable multi-class probabilistic classifier estimated by maximum likelihood is incentive-compatible.*

Proof. Due to separability,

$$v_a\left(x; w^\star(S)\right) = v_a\left(x; w^\star\left(S \cup \left\{\left(x', y' \notin \{a, b\}\right)\right\}\right)\right), \text{ and}$$
$$v_b\left(x; w^\star(S)\right) = v_b\left(x; w^\star\left(S \cup \{(x, y = a)\}\right)\right) = v_b\left(x; w^\star\left(S \cup \left\{\left(x', y' \notin \{a, b\}\right)\right\}\right)\right).$$

MC follows from the optimality condition of $w^\star(S)$ and IIA follows immediately.

□

Corollary 3. *Bayes classifiers estimated by maximum likelihood are incentive-compatible.*

Proof. Follows from Corollary 2.

□

Kernel density estimators are not estimated by maximum likelihood, so the previous results do not hold, although the proof is similar. There are two general approaches to use kernel densities for classification, the first is to use all the points to estimate the density and the second is to estimate the densities for each class separately (see Taylor (1997)). The second approach is similar to a separable classifier. K-Nearest Neighbor is a special case of this with a uniform kernel.

Corollary 4. *Kernel density estimators are incentive-compatible.*

Proof. The first approach suggests that,

$$\mathbb{P}\left\{X = x\right\} = \frac{1}{nh^D} \sum_{i'=1}^{n} w_{i'}, \text{ where } w_{i'} = K\left(\frac{x - x_{i'}}{h}\right).$$

Define $w^\star$ as the weight function when all agents report truthfully, and $w^\dagger = w^\star\left(y_i^\dagger\right)$ as the weight if agent i misreports, then the classification probabilities for agent i from dividing up the sum based on the class is,

$$\begin{aligned}
\mathbb{P}\left\{Y = y_i | x_i; w^\star\right\} &= \frac{1}{n} \sum_{i'=1}^{n} w_{i'}^\star \mathbb{1}_{\hat{y}_{i'} = y_i} \\
&= \frac{1}{n} \sum_{i'=1}^{n} w_{i'}^\star \mathbb{1}_{\hat{y}_{i'} = y_i, i' \neq i} + \frac{1}{n} w_i^\star \\
&= \mathbb{P}\left\{Y = y_i | x_i; w^\dagger\right\} + \frac{1}{n} K\left(0\right), \text{ since } y_i^\dagger \neq y_i \\
&\leqslant \mathbb{P}\left\{Y = y_i | x = x_i; w^\star \dagger\right\},
\end{aligned}$$

meaning reporting truthfully results in a larger probability compared to reporting $y_i^\dagger$ instead. Alternatively, the second approach suggests that, if the classification probabilities are computed based on Taylor (1997),

$$\mathbb{P}\left\{X = x | Y = y\right\} = \frac{1}{n_y h^D} \sum_{i'=1}^{n} w_{i'} \mathbb{1}_{\hat{y}_{i'} = y}, n_y = \sum_{i'=1}^{n} \mathbb{1}_{\hat{y}_{i'} = y}$$

Similar to the above derivation (and also as a special case of a Bayes estimator),

$$\begin{aligned}
\mathbb{P}\left\{Y = y_i | x_i; w^\star\right\} &= \frac{1}{n_y} \sum_{i'=1}^{n} w_{i'}^\star \mathbb{1}_{\hat{y}_{i'} = y_i} \\
&= \mathbb{P}\left\{Y = y_i | x_i; w^\star \dagger\right\} + \frac{1}{n_y} K\left(0\right) \\
&\leqslant \mathbb{P}\left\{Y = y_i | x_i; w^\star \dagger\right\}.
\end{aligned}$$

□

4.1 Empirical Risk Minimization

A similar result can be obtained for empirical risk minimization. We could either add an assumption that the loss function can be normalized so that the sum is constant and it behaves the same way as a probabilistic classifier, or we could use stronger Monotonic and IIA Conditions. Here, we state the additional normalization condition.

Definition 5. (Normalized Loss) A loss function ℓ is normalized if given a hypothesis h, for any point,

$$\sum_{y} \ell\left(y; h\right) = C, \text{ constant}.$$

Definition 6. (Monotonic Condition for ERM) Multi-class empirical risk minimization classifiers are monotonic if, given a training set S, for any point x with labels a and b,

$$\frac{\ell\left(y = a; h^\star\left(S\right)\right)}{\ell\left(y = b; h^\star\left(S\right)\right)} \geqslant \frac{\ell\left(y = a; h^\star\left(S \cup \left\{\left(x, y = a\right)\right\}\right)\right)}{\ell\left(y = b; h^\star\left(S \cup \left\{\left(x, y = a\right)\right\}\right)\right)}.$$

Definition 7. (IIA Condition for ERM) Multi-class empirical risk minimization classifiers are independent of irrelevant alternatives if, given a training set S, for any point x and any pair of labels a and b,

$$\frac{\ell\left(y=a; h^{\star}\left(S\right)\right)}{\ell\left(y=b; h^{\star}\left(S\right)\right)} = \frac{\ell\left(y=a; h^{\star}\left(S \cup \left\{\left(x', y' \notin \left\{a, b\right\}\right\}\right)\right)\right)}{\ell\left(y=b; h^{\star}\left(S \cup \left\{\left(x', y' \notin \left\{a, b\right\}\right\}\right)\right)\right)}.$$

Corollary 5. *Multi-class empirical risk minimization classifiers with normalized loss functions are incentive-compatible if it is monotonic and independent of irrelevant alternatives.*

Proof. For a fixed dataset $\left\{\left(x_i, y_i\right)\right\}_{i=1}^{n}$, and the hypothesis class $\mathcal{H}$, let the optimal classifier in the case every agent report truthfully be,

$$h^{\star} = \arg\min_{h \in \mathcal{H}} \sum_{i'=1}^{n} \ell\left(y_{i'}; h\right).$$

Fix an agent i, her feature vector x_i, and fix other agents' reports, $\left(x_{-i}, y_{-i}\right)$, define the optimal classifier given the classifier h and the misreport of agent i, $y_i^{\dagger}$ as,

$$h^{\dagger} = \arg\min_{h \in \mathcal{H}} \sum_{i'=1, i' \neq i}^{n} \ell\left(y_{i'}; h\right) + \ell\left(y_i^{\dagger}; h\right).$$

Now suppose, for a contradiction, that agent i prefers misreporting, assume the following incentive inequality,

$$\ell\left(y_i; h^{\star}\right) > \ell\left(y_i; h^{\dagger}\right).$$

If there are only two classes, then by symmetry,

$$\ell\left(y_i^{\dagger}; h^{\star}\right) < \ell\left(y_i^{\dagger}; h^{\dagger}\right).$$

If there are more than two classes, fix a third $y_i' \notin \left\{y_i, y_i^{\dagger}\right\}$, and define an intermediate maximum likelihood estimate from removing the point $\left(x_i, y_i\right)$,

$$h' = \arg\min_{h \in \mathcal{H}} \sum_{i'=1, i' \neq i}^{n} \ell\left(y_{i'}; h\right),$$

then the Monotonic Condition for ERM implies,

$$\frac{\ell\left(y_i; h^{\star}\right)}{\ell\left(y_i'; h^{\star}\right)} \leqslant \frac{\ell\left(y_i; h'\right)}{\ell\left(y_i'; h'\right)},$$

and the IIA Condition implies,

$$\frac{\ell\left(y_i; h'\right)}{\ell\left(y_i'; h'\right)} = \frac{\ell\left(y_i; h^{\dagger}\right)}{\ell\left(y_i'; h^{\dagger}\right)}.$$

Combining the above two inequalities with the incentive inequality, we have,

$$\ell\left(y_i'; h^{\star}\right) > \ell\left(y_i'; h^{\dagger}\right).$$

Note that the above inequality is true for all $y_i' \notin \left\{y_i, y_i^{\dagger}\right\}$, summing over all such y_i' results in,

$$\sum_{y \notin \left\{y_i, y_i^{\dagger}\right\}} \ell\left(y_i'; h^{\star}\right) > \sum_{y \notin \left\{y_i, y_i^{\dagger}\right\}} \ell\left(y_i'; h^{\dagger}\right),$$

and given the losses are normalized,

$$C - \ell\left(y_i; h^{\star}\right) - \ell\left(y_i^{\dagger}; h^{\star}\right) > C - \ell\left(y_i; h^{\dagger}\right) - \ell\left(y_i^{\dagger}; h^{\dagger}\right),$$

and using the incentive inequality again,

$$\ell\left(y_i^{\dagger}; h^{\star}\right) < \ell\left(y_i^{\dagger}; h^{\dagger}\right).$$

Now, due to the optimality of $h^{\dagger}$,

$$\sum_{i'=1,i'\neq i}^{n} \ell\left(y_{i'}; h^{\dagger}\right) + \ell\left(y_i^{\dagger}; h^{\dagger}\right) \leqslant \sum_{i'=1,i'\neq i}^{n} \ell\left(y_{i'}; h^{\star}\right) + \ell\left(y_i^{\dagger}; h^{\star}\right),$$

using the above inequalities, the comparison can be simplified to,

$$\sum_{i'=1,i'\neq i}^{n} \ell\left(y_{i'}; h^{\dagger}\right) \leqslant \sum_{i'=1,i'\neq i}^{n} \ell\left(y_{i'}; h^{\star}\right),$$

$$\sum_{i'=1}^{n} \ell\left(y_{i'}; h^{\dagger}\right) \leqslant \sum_{i'=1}^{n} \ell\left(y_{i'}; h^{\star}\right),$$

which is a contradiction to the optimality of $h^{\star}$.

□

References

Cai, Yang, Constantinos Daskalakis, and Christos Papadimitriou (2015), "Optimum statistical estimation with strategic data sources." In *Conference on Learning Theory*, 280–296.

Chen, Yiling, Yang Liu, and Chara Podimata (2019), "Grinding the space: Learning to classify against strategic agents." *arXiv preprint arXiv:1911.04004*.

Chen, Yiling, Chara Podimata, Ariel D Procaccia, and Nisarg Shah (2018), "Strategyproof linear regression in high dimensions." In *Proceedings of the 2018 ACM Conference on Economics and Computation*, 9–26.

Dekel, Ofer, Felix Fischer, and Ariel D Procaccia (2010), "Incentive compatible regression learning." *Journal of Computer and System Sciences*, 76, 759–777.

Dekel, Ofer and Ohad Shamir (2009a), "Good learners for evil teachers." In *Proceedings of the 26th annual international conference on machine learning*, 233–240.

Dekel, Ofer and Ohad Shamir (2009b), "Vox populi: Collecting high-quality labels from a crowd." In *COLT*.

Dütting, Paul, Zhe Feng, Harikrishna Narasimhan, David C Parkes, and Sai Srivatsa Ravindranath (2017), "Optimal auctions through deep learning." *arXiv preprint arXiv:1706.03459*.

Epasto, Alessandro, Mohammad Mahdian, Vahab Mirrokni, and Song Zuo (2018), "Incentive-aware learning for large markets." In *Proceedings of the 2018 World Wide Web Conference*, 1369–1378.

Gast, Nicolas, Stratis Ioannidis, Patrick Loiseau, and Benjamin Roussillon (2013), "Linear regression from strategic data sources." *arXiv preprint arXiv:1309.7824*.

Golowich, Noah, Harikrishna Narasimhan, and David C Parkes (2018), "Deep learning for multi-facility location mechanism design." In *IJCAI*, 261–267.

Meir, Reshef, Ariel D Procaccia, and Jeffrey S Rosenschein (2012), "Algorithms for strategyproof classification." *Artificial Intelligence*, 186, 123–156.

Mihailescu, Marian and Yong Meng Teo (2010), "Strategy-proof dynamic resource pricing of multiple resource types on federated clouds." In *International Conference on Algorithms and Architectures for Parallel Processing*, 337–350, Springer.

Perote, Javier and Juan Perote-Pena (2004), "Strategy-proof estimators for simple regression." *Mathematical Social Sciences*, 47, 153–176.

Procaccia, Ariel D and Moshe Tennenholtz (2009), "Approximate mechanism design without money." In *Proceedings of the 10th ACM conference on Electronic commerce*, 177–186.

Richardson, Adam, Ljubomir Rokvic, Aris Filos-Ratsikas, and Boi Faltings (2019), "Privately computing influence in regression models."

Shah, Nihar B and Dengyong Zhou (2016), "Double or nothing: Multiplicative incentive mechanisms for crowdsourcing." *The Journal of Machine Learning Research*, 17, 5725–5776.

Taylor, Charles (1997), "Classification and kernel density estimation." *Vistas in Astronomy*, 41, 411–417.