Incentive Incompatibility of Logistic Regression

Abstract

We study the incentive compatibility problem of multi-class logistic regression. We provide a simple 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 problem disappears if the logistic model is estimated by minimizing the zero-one loss, and if a Bayes classifier estimated with maximum likelihood is used.

1 Introduction

There are n strategic agents each providing the label of one data point to the principal. The principal is the learner and builds a machine learning model based on the data points provided by the agents. An agent, i, has publicly known feature vector, x_i, and a private discrete label, y_i. The objective of the agent is to maximize the probability that her data point is labeled correctly by the principal's model, and the agent can choose to report $y_i^\dagger$ to achieve the objective, with the possibility of misreporting $y_i^\dagger \neq y_i$. We say a dataset is incentive incompatible with respect to the learner, described by a parametric model, if at least one of the n agents has the incentive to misreport.

The following is the diagram showing a dataset that is incentive incompatible with respect to the multi-class logistic regression model. In the dataset, each of the $n = 18$ agents, i, has a two dimensional feature vector and a private label can take on one of three values: "red", "green", or "blue".

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

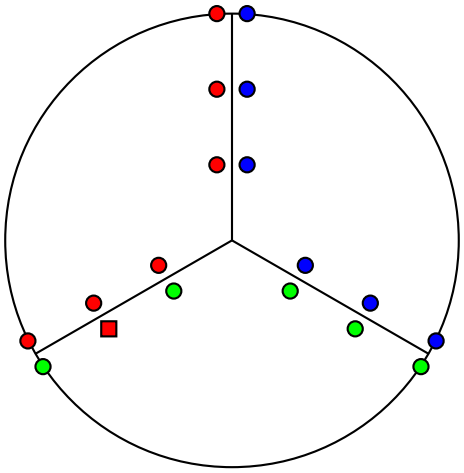

Figure 1: Incentive Incompatible Example

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 point, labeled by a square in the plot, that is on the "incorrect" side of the boundary. Suppose the point corresponds to the feature vector of an agent i with private label "red", then 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" with probability 0.4966. Therefore, 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.

However, this dataset is incentive compatible if zero-one loss is minimized when estimating the logistic parameters, and if the Bayes classifier is used with maximum likelihood estimation. In general, misreporting will always lead to a lower classification probability of the agent's true label for these classifiers.

Previous work on mechanism design for machine learning with strategic data sources focus 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 machine learner (principal).

The first group of papers focuses on principal-agent problems 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' feature vectors are public, but their labels are private. **?**, **?**, and **?** focus on strategy-proof linear regression algorithms and introduced clockwise repeated median estimators, generalized resistant hyperplane estimators, and modified generalized linear squares estimators. **?** 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 **?** assumes the labels are discrete variables (classification problems) and proposes a class of random dictator mechanisms.
2. Some models assume the agents' feature vectors are also private. **?** investigates such problems for linear regressions.
3. Other models do not distinguish between feature vectors and labels. Each agent has a private valuation. These problems are usually modeled as facility locations problems and the solution involves some variant of the Vickrey-Clarke-Groves or Meyerson auction. These include **?**, **?**, **?**, and **?**.

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. **?**

focuses on the linear regression problem in this scenario, and **?** and **?** investigates the problem for more general machine learning problems. **?** 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 **?**, **?**.

2 Logistic Regression

2.1 Model and Example

In this section, 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 feature vector, $x_i \in \mathbb{R}^m$, and private discrete label, $y_i \in \{1, 2, ..., k\}$, has the objective of maximizing the probability that her data point is labeled correctly by the principal's model, 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 $w^\star\left(y_i^\dagger\right)$, the agent's objective can be written as,

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

where,

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

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

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

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

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,

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

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

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

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

A learner (algorithm) is incentive compatible if there does not exist a dataset that is incentive incompatible.

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

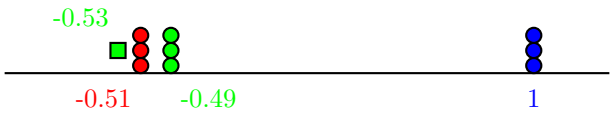

Figure 2: Second Incentive Incompatible Example

Proof. The simplest numerical example we found is the dataset with 10 points $\{(x_i, y_i)\}_{i=1}^{10} = \{(-0.53, 2), (-0.51, 1), (-0.51, 1), (-0.51, 1), (-0.49, 2), (-0.49, 2), (-0.49, 2), (1, 3), (1, 3), (1, 3)\}$. It has a similar structure as the data set shown in Figure 1 in that multiple points are close to a classification boundary, and the one point on the "incorrect" side of the boundary has the incentive to pretend to be the third class.
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 green square point correspond to agent 1 with $x_1 = (-0.53, 2)$ and $y_1 = 2$.

$$\mathbb{P}\{Y = 1 | w^\star, x_1\} = 0.2715,$$
$$\mathbb{P}\left\{Y = 1 | w^\star \left(y_1^\dagger = 3\right), x_1\right\} = 0.3709.$$

Here, $w^\star$ is given by, with class 1 weights normalized to 0,

Class	(Intercept)	x1
2	25.73034	50.40985
3	21.77434	63.65515

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

Class	(Intercept)	x1
2	6.603135	13.09844
3	8.231013	18.48809

It turns out that the incentive incompatibility heavily depends the numerical stability of the weights-finding algorithm: using BFGS to maximize the likelihood shows that this dataset is actually incentive compatible for all points.

□

The example described in Figure 1 is a dataset that is also 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 correspond to agent 1 with $x_1 = (-1.63, -1.17)$ and $y_1 = 1$.

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

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
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
2	-0.1915645	3.473426	-5.507418
3	0.8273350	4.309293	-1.200060

Using BFGS to maximize likelihood leads to the same incentive incompatibility result. Currently there is no formal proof that the result is not due to numerical instability.

2.2 Binary Classification

Proposition 2. *Binary classification using ERM with any loss function ℓ is incentive compatible.*

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

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

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

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

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

$$\ell\left(h^\star, x_i, y_i\right) > \ell\left(h^\star\left(y_i^\dagger\right), x_i, y_i\right),$$

which implies, since $y_i^\dagger = 1 - y_{i,}$

$$\ell\left(h^\star, x_i, y_i^\dagger\right) < \ell\left(h^\star\left(y_i^\dagger\right), x_i, y_i^\dagger\right).$$

Note that the above implication only works for binary classification.
Due to the optimality of $h^\star\left(y_i^\dagger\right)$,

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

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

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

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

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

□

2.3 Zero-One Loss Logistic Regression

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 **?**. Their result on incentive compatibility of empirical risk minimization in the regression setting is applicable in our model. In addition to absolute value loss, which is not a meaningful loss function for multi-class logistic regression, zero-one loss logistic regression with deterministic predictions is also incentive compatible.

Proposition 3. *Multi-class deterministic classifiers estimated by empirical risk minimization with zero-one loss is incentive compatible.*

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

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

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

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

If $y_i = h^\star(x_i)$, then the classifier is already classifying x_i correctly, misreporting will not improve the outcome for i. Now let the prediction be $h^\star(x_i) = y^\star \neq y_i$, and suppose $h^\star$ is making k mistakes, meaning,

$$k = \min_{h\in\mathcal{H}} \mathcal{L}\left(h^\star, y_i\right).$$

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,

   $$\mathcal{L}\left(h^\dagger, y^\star\right) \leqslant k-1,$$

   because $\mathcal{L}\left(h^\dagger, y^\star\right) > k-1 = \mathcal{L}\left(h^\star, y^\star\right)$ contradicts the optimality of $h^\dagger$.
   Now suppose that agent i could get her true label with $h^\dagger$, meaning $h^\dagger(x_i) = y_i$, then,

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

   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,

   $$\mathcal{L}\left(h^\dagger, y'\right) \leqslant k,$$

   because if $\mathcal{L}\left(h^\dagger, y'\right) > k = \mathcal{L}\left(h^\star, y'\right)$ contradicts the optimality of $h^\dagger$.
   Now suppose that agent i could get her true label with $h^\dagger$, then,

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

   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.

3 Bayes Classifier

The example given previously is incentive compatible with respect to the Naive Bayes classifier. None of the agents have the incentive to misreport their labels. This is always true in general for any parametric Bayesian classifier estimated using maximum likelihood.

Proposition 4. *Bayesian classifiers are incentive compatible.*

Proof. Suppose the loglikelihood function of class y given the feature vector x and the parameter w is $\ell(x; w)$, and define the optimal parameter, $w^\star$ for class y_i, of the truthful model as,

$$\begin{aligned} w^\star &= \arg\max_w \sum_{i':y_{i'}=y_i} \ell(x_{i'}; w) \\ &= \arg\max_w \sum_{i'\neq i:y_{i'}=y_i} \ell(x_{i'}; w) + \ell(x_i; w). \end{aligned}$$

Let the optimal parameter when agent i reports $y_i^\dagger \neq y_i$ be $w^\dagger$,

$$w^\dagger = \arg\max_w \sum_{i'\neq i:y_{i'}=y_i} \ell(x_{i'}; w).$$

In particular, these implies the following optimality conditions,

$$\begin{aligned} \sum_{i'\neq i:y_{i'}=y_i} \ell(x_{i'}; w^\star) + \ell(x_i; w^\star) &\geqslant \sum_{i'\neq i:y_{i'}=y_i} \ell\left(x_{i'}; w^\dagger\right) + \ell\left(x_i; w^\dagger\right), \\ \sum_{i'\neq i:y_{i'}=y_i} \ell\left(x_{i'}; w^\dagger\right) &\geqslant \sum_{i'\neq i:y_{i'}=y_i} \ell(x_{i'}, w^\star). \end{aligned}$$

Taking the difference between the two inequalities, we have,

$$\ell(x_i; w^\star) \geqslant \ell\left(x_i; w^\dagger\right).$$

Note that the empirical prior probability for class y_i is decreased if the number of data with label y_i is decreased by 1. Therefore, the posterior probabilities satisfy,

$$\mathbb{P}\{y_i | x_i, w^\star\} \geqslant \mathbb{P}\left\{y_i | x_i, w^\dagger\right\}.$$

Therefore, no agent can improve the outcome and the dataset is incentive compatible.

□

One special case of this is the Bayes classifier for the Gaussian Mixture Model.

4 Kernel Density Estimators

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 **?**).

The first approach suggests that,

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

Define $w^\star$ to be the weight function when all agents report truthfully, then the classification probabilities for agent i from dividing up the sum based on the class is,

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

meaning reporting truthfully results in a larger probability compared to reporting $y_i^\dagger$ instead.

Alternative if the classification probabilities are computed based on **?**,

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

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

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

K-Nearest Neighbor is a special case of this with a uniform kernel.

5 Artificial Examples

One intuition behind why the above classifiers are always incentive-compatible is that one-vs-one classification decisions are made independently for classifiers. The classifiers like logistic regression have 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.

Consider the three-class $1D$ threshold classifiers in the form,

$$\hat{y}\left(x; t\right) = \begin{cases} 1 & \text{if } x < -t \\ 2 & \text{if } -t \leqslant x \leqslant t \\ 3 & \text{if } x > t \end{cases}$$

This is effectively a one-vs-one threshold classifier with the 1-vs-2 threshold of $-t$, with the 2-vs-3 threshold of t and the 1-vs-3 threshold of any value between $-t$ and t. Now define the loss function as the error margin,

$$\ell\left(x_i, y_i; t\right) = \begin{cases} \left(x_i - (-t)\right)^2 \mathbb{1}_{\hat{y}(x_i;t)\neq y_i} & \text{if } y_i = 1 \\ \min\left\{\left(x_i - (-t)\right)^2, \left(x_i - t\right)^2\right\} \mathbb{1}_{\hat{y}(x_i;t)\neq y_i} & \text{if } y_i = 2 \\ \left(x_i - t\right)^2 \mathbb{1}_{\hat{y}(x_i;t)\neq y_i} & \text{if } y_i = 3 \end{cases}$$

Then, the learner's maximization problem is,

$$\min_{t\in\mathbb{R}} \ell\left(x_i, y_i; t\right).$$

The agents' problem is to minimize the loss for their point with their true label. The loss can be either the margin loss or zero-one loss, here, for simplicity, assume the agents also want to minimize the margin,

$$\max_{y_i^\dagger \in \{1,2,3\}} \ell\left(x_i, y_i; t^\star\left(y_i^\dagger\right)\right),$$

where $t^\star\left(y_i^\dagger\right)$ is the optimal threshold t for the learner when agent i misreports the label as $y_i^\dagger$.

For the dataset $\left\{\left(x_i, y_i\right)\right\}_{i=1}^5 = \left\{\left(-4, 1\right), \left(-3, 2\right), \left(-2, 1\right), \left(3, 2\right), \left(4, 3\right)\right\}$, the minimum loss is 1 and it occurs when $t = 3$ when every agent reports truthfully. In this case, agent $(-2, 1)$ is misclassified as $(-2, 2)$. However, if $(-2, 1)$ misreports as $(-2, 3)$, the loss from $t = 3$ is 25 and the minimum loss is 22 and it occurs when $t = 0$. In this case, agent $(-2, 1)$ is classified correctly. In

particular, the agent (-2, 1) improves the loss from 1 to 0.

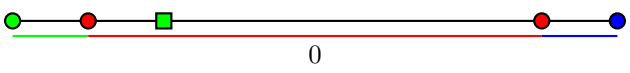

Figure 3: $1D$ Artificial Incentive Incompatible Example (Truthful)

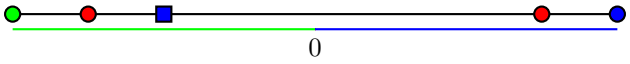

Figure 4: $1D$ Artificial Incentive Incompatible Example (Misreport)

Intuitively, when the misclassified class-1 point pretends to be a class-3 point with a large loss, the classifier modifies the 2-vs-3 decision boundary to minimize that loss and the point benefits from the interdependence between the 2-vs-3 decision and the 1-vs-2 decision.

Consider another three-class $2D$ threshold classifier in the form,

$$\hat{y}\left(x_1, x_2; a, b, c, d\right) = \begin{cases} 1 & \text{if } ax_1 + bx_2 + c \leqslant 0 \text{ and } bx_1 - ax_2 + d > 0 \\ 2 & \text{if } ax_1 + bx_2 + c \leqslant 0 \text{ and } bx_1 - ax_2 + d \leqslant 0 \\ 3 & \text{if } ax_1 + bx_2 + c > 0 \end{cases}$$

$$\ell\left(x_i, y_i; t\right) = d\left(x_i, t\right) \mathbb{1}_{\hat{y}(x_i;t) \neq y_i}, t = (a, b, c, d) \in \mathbb{R}^4$$

where $d\left(x_i, t\right)$ is the minimum distance from x_i to one of the lines $ax_1 + bx_2 + c = 0$ and $bx_1 - ax_2 + d = 0$, or,

$$d\left(x_i, t\right) = \frac{\min\left(|ax_1 + bx_2 + c|, |bx_1 - ax_2 + d|\right)}{\sqrt{a^2 + b^2}}.$$

For the dataset $\{(x_{i1}, x_{i2}, y_i)\}_{i=1}^5 = \{(0, 0, 1), (0, 0, 3), (0, -1, 1), (0, -1, 2), (1, 0, 2)\}$, the minimum loss is 0 and it occurs when $a = b = c = d = 0$ when every agent reports truthfully. In this case, agent $(0, -1, 1)$ is misclassified as $(0, -1, 2)$. However, if $(0, -1, 1)$ misreports as $(0, -1, 3)$, the loss from $a = b = c = d = 0$ is 1 and the minimum loss is $\frac{\sqrt{2}}{2}$ and it occurs when $a = -1, b = c = d = 1$. In this case, agent $(0, -1, 1)$ is classified correctly.

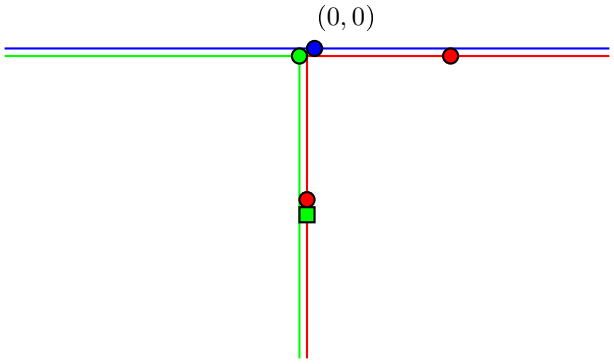

Figure 5: $2D$ Artificial Incentive Incompatible Example (Truthful)

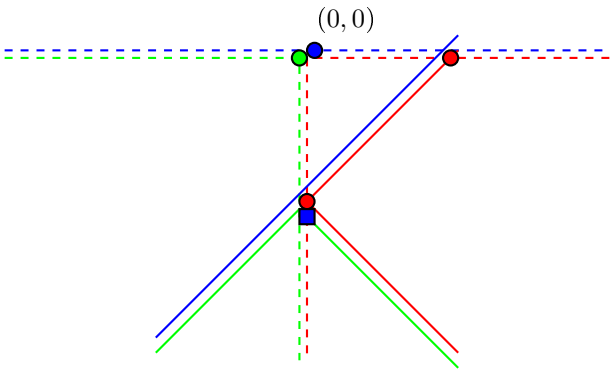

Figure 6: $2D$ Artificial Incentive Incompatible Example (Misreport)

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.

Proposition 5. *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 $\{(x_i,y_i)\}_{i=1}^n$, 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,

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

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),

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

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 $h^\dagger$,

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

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

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

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

□

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 1 so it behaves the same way as a probabilistic classifier, or we could use stronger Monotonic and IIA Conditions.