Information Discrepancy in Strategic Learning

Yahav Bechavod, Chara Podimata, Steven Wu, and Juba Ziani

yahav.bechavod@cs.huji.ac.il, podimata@g.harvard.edu, zstevenwu@cmu.edu, juba.ziani@isye.gatech.com

Main Question
How does information discrepancy regarding the learner’s decision rule affect the different subgroups of the population with respect to their ability to improve their outcomes?

Model (Formally)
1. Nature decides the ground truth assessment: \( w^* \in \mathbb{R}^d \).
2. Learner deploys score rule \( w \in \mathbb{R}^d \) but does not reveal it to agents.
3. Agents (per subgroup \( g \)) draw their private feature vectors from space \( X: x_1 \sim D_1 \) and \( x_2 \sim D_2 \).
4. Given peer dataset \( S_g \), private feature vector \( x_g \), & their utility \( u(x_g, x'; g) \), the agents best-respond with feature vector: \( \hat{x}_g = \arg\max_x u(x_g, x'; g) \).

Subgroup Feature Vector Discrepancies
- \( S_1, S_2 \): subspaces of \( X \) defined by supports of \( D_1, D_2 \)
- \( \Pi_1, \Pi_2 \in \mathbb{R}^d \): orthogonal projection matrices onto \( S_1, S_2 \)
- \( x_g = \Pi_g x_g \) (feature discrepancy)

Why is \( w^* \neq w \)?
- \( w^* \) is such that \( \text{TrueScore} = (w^*, x) \) for the private \( x \).
- \( w \) is the rule that maximizes the agents’ Social Welfare after best-responding:

\[
\text{Score}(x') - \text{Cost}(x_g \rightarrow x') = (x', w^*) - \| \Pi_g (x' - x_g) \|^2
\]

Subgroup’s estimated rule using \( S_g \)
- Subgroups use ERM on their respective \( S_g \).
- Each group \( g \) obtains estimate rule: \( w_{est}(g) = \Pi_g w \).

Subgroup’s Best-Response
- \( \text{utility}(x_g, x'; g) = \text{Score}(x') - \text{Cost}(x_g \rightarrow x') \)
- \( \Pi_g (x' - x_g) \) scaled by cost matrix \( A_g \):

\[
\hat{x}_g = x + A_g^{-1} \Pi_g w
\]

Learner’s Rule
\[
w = \frac{(\Pi_1 A_1^{-1} + \Pi_2 A_2^{-1})w^*}{\| (\Pi_1 A_1^{-1} + \Pi_2 A_2^{-1})w^* \|}
\]

Setup
What is “strategic learning”?

Agents report their data

Strategically change features

Standard assumption in all prior work: learner’s rule is fully known by the agents (i.e., full transparency).
- Far-fetched assumption
- In reality: banks, institutions rarely reveal their decision rules (reasons: privacy, proprietary software etc).
- Instead of full revelation: examples with explanations, examples of past decisions etc.

Our Setup at a High Level
- Agents belong in 2 subgroups (green, blue).
- Agents do not know the decision rule.
- Agents have information about past decision among their subgroup peers (peer dataset).
- Using this, they try to recover the decision rule. → information discrepancy

Improvement in Equilibrium
Three measures of interest:
1. Do-no-harm: “Are all individuals better off?”
2. Total improvement: “By how much?”
3. Per-unit improvement: “Is effort exerted optimally?”

Improve best-response using \( S_g \)
- Minimize \( \Pi_g \)
- \( w^* \) is such that \( \text{TrueScore} = (w^*, x) \) for the private \( x \).
- \( w \) is the rule that maximizes the agents’ Social Welfare after best-responding:

\[
\text{Score}(x') - \text{Cost}(x_g \rightarrow x') = (x', w^*) - \| \Pi_g (x' - x_g) \|^2
\]

Subgroup’s estimated rule using \( S_g \)
- Subgroups use ERM on their respective \( S_g \).
- Each group \( g \) obtains estimate rule: \( w_{est}(g) = \Pi_g w \).

Subgroup’s Best-Response
- \( \text{utility}(x_g, x'; g) = \text{Score}(x') - \text{Cost}(x_g \rightarrow x') \)
- \( \Pi_g (x' - x_g) \) scaled by cost matrix \( A_g \):

\[
\hat{x}_g = x + A_g^{-1} \Pi_g w
\]

Learner’s Rule
\[
w = \frac{(\Pi_1 A_1^{-1} + \Pi_2 A_2^{-1})w^*}{\| (\Pi_1 A_1^{-1} + \Pi_2 A_2^{-1})w^* \|}
\]

Main Results
Thm. 1: Do-no-harm is not always guaranteed.
- Negative externality (outcome deterioration) due to information discrepancy is possible.

Notable Examples:
- Manipulation costs that are proportional.
- Costs only differ outside of the information overlap.

Thm. 2: Characterization of (mild) conditions to guarantee individual outcomes improve.

Experiments
- Datasets: Taiwan-Credit, Adult
- Validation of theoretical results even despite not fully satisfying assumptions of Thms.

Full version: