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

Full version:

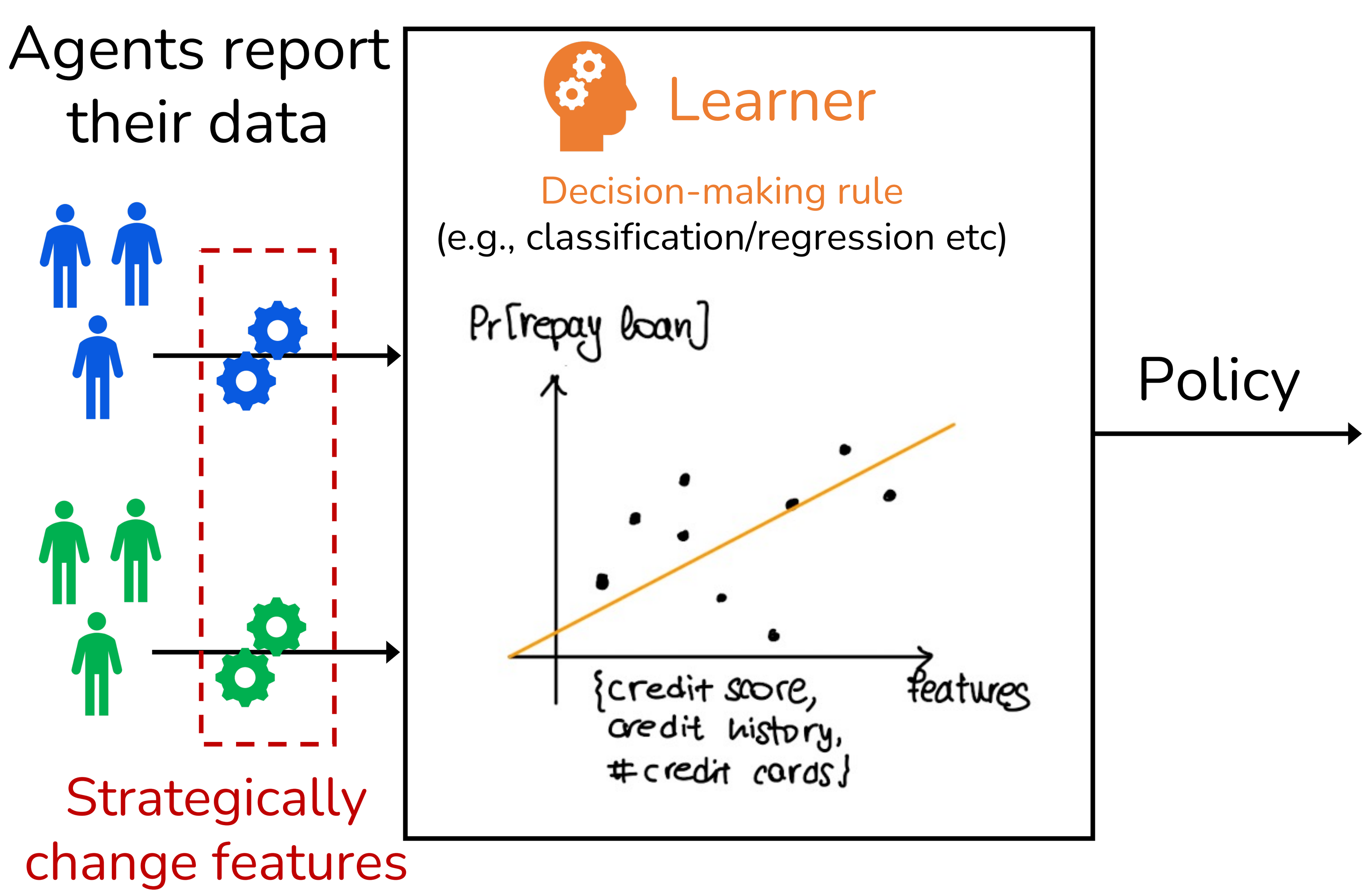
SCAN ME

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?

Setup

What is "**strategic learning**"?



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**

Model (Formally)

1. Nature decides the ground truth assessment: $\boldsymbol{w}^\star \in \mathbb{R}^d$.
2. Learner deploys score rule $\boldsymbol{w} \in \mathbb{R}^d$ but does **not** reveal it to agents.
3. Agents (per subgroup g) draw their private feature vectors from space $\mathcal{X}$: $\boldsymbol{x}_1 \sim \mathcal{D}_1$ and $\boldsymbol{x}_2 \sim \mathcal{D}_2$.
4. Given peer dataset S_g, private feature vector $\boldsymbol{x}_g$, & their utility $u(\boldsymbol{x}_g, \boldsymbol{x}'_g; g)$, the agents best-respond with feature vector: $\hat{x}_g = \arg\max_{\boldsymbol{x}'} u(\boldsymbol{x}_g, \boldsymbol{x}'; g)$.

Subgroup Feature Vector Discrepancies

- $\mathcal{S}_1, \mathcal{S}_2$: subspaces of $\mathcal{X}$ defined by supports of $\mathcal{D}_1, \mathcal{D}_2$
- $\Pi_1, \Pi_2 \in \mathbb{R}^d$: orthogonal projection matrices onto $\mathcal{S}_1, \mathcal{S}_2$

  → $\boldsymbol{x}_g = \Pi_g \boldsymbol{x}_g$ (feature discrepancy)

Why is $\boldsymbol{w}^\star \neq \boldsymbol{w}$?

- $\boldsymbol{w}^\star$ is such that $TrueScore = \langle \boldsymbol{w}^\star, \boldsymbol{x} \rangle$ for the **private** $\boldsymbol{x}$.
- $\boldsymbol{w}$ is the rule that maximizes the agents' **Social Welfare** after *best-responding*:

$$\boldsymbol{w} = \arg\max_{\boldsymbol{w}'} \left(\mathbb{E}_{x_1 \sim \mathcal{D}_1}[\langle \hat{\boldsymbol{x}}_1, \boldsymbol{w}^\star \rangle] + \mathbb{E}_{x_2 \sim \mathcal{D}_2}[\langle \hat{\boldsymbol{x}}_2, \boldsymbol{w}^\star \rangle] \right)$$

Subgroup's estimated rule using S_g

- Subgroups use **ERM** on their respective S_g.
- Each group g obtains estimate rule: $\boldsymbol{w}_{est}(g) = \Pi_g \boldsymbol{w}$.

Subgroup's Best-Response

- $utility(\boldsymbol{x}_g, \boldsymbol{x}'; g) \coloneqq Score(\boldsymbol{x}') - Cost(\boldsymbol{x}_g \to \boldsymbol{x}')$
  $= \langle x', w^\star \rangle - ||A_g(\boldsymbol{x}' - \boldsymbol{x}_g)||^2$
- Agents move in direction of w_{est}, scaled by cost matrix A_g:
  $\hat{\boldsymbol{x}}_g = \boldsymbol{x} + A_g^{-1} \Pi_g \boldsymbol{w}$

Learner's Rule

$$\boldsymbol{w} = \frac{(\Pi_1 A_1^{-1} + \Pi_2 A_2^{-1}) \boldsymbol{w}^\star}{||(\Pi_1 A_1^{-1} + \Pi_2 A_2^{-1}) \boldsymbol{w}^\star||}$$

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?"

Main Results

Thm. 1: Do-no-harm is not always guaranteed.

→ Negative externality (outcome deterioration) due to information discrepancy is possible.

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

Notable Examples:

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

Thm. 3: Characterization of conditions for improvement effort to be optimally exerted.

Experiments

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