



SCAN ME

Information Discrepancy in Strategic Learning

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

ICML 2022

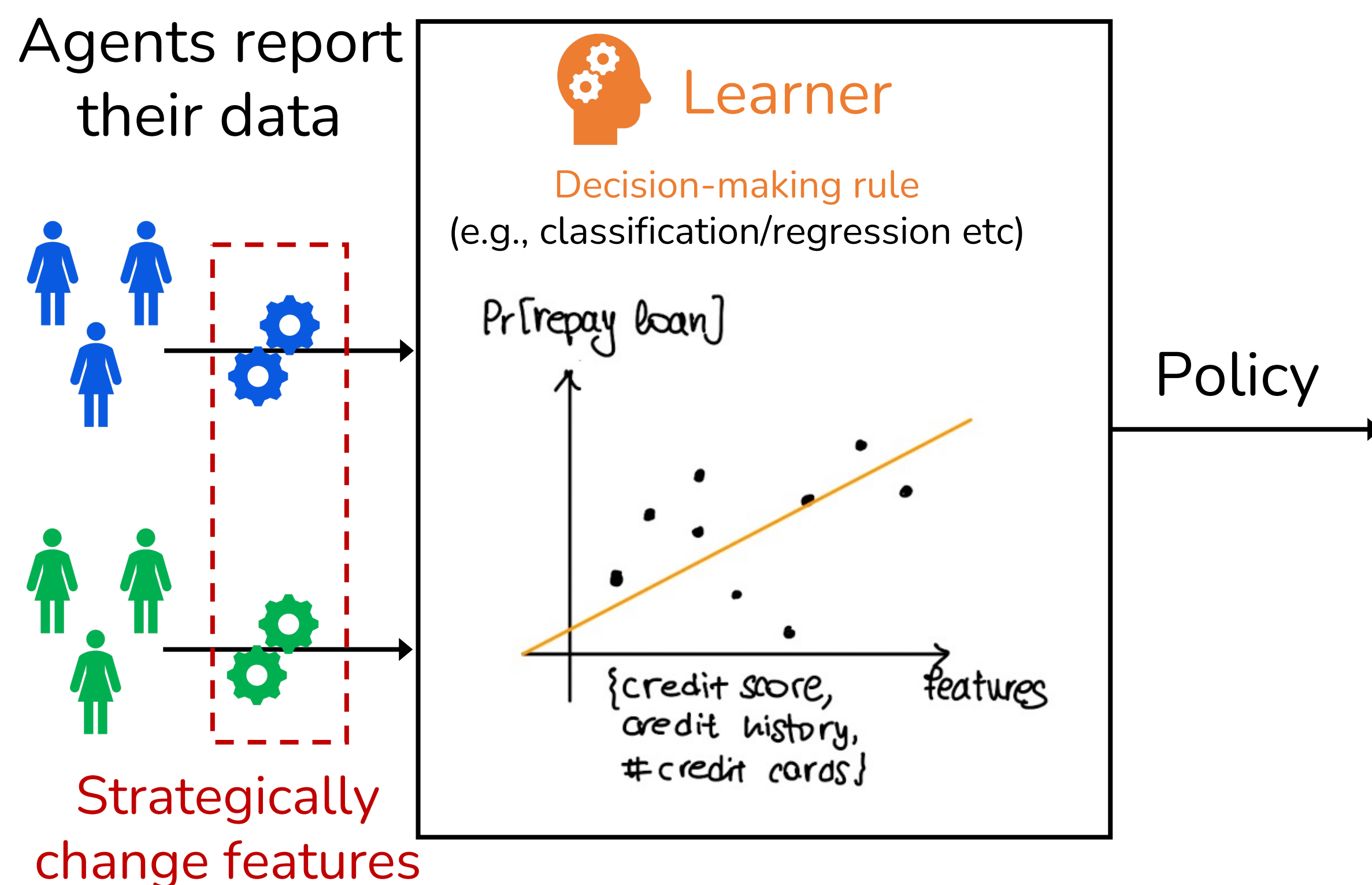
yahav.bechavod@cs.huji.ac.il, podimata@berkeley.edu, zstevenwu@cmu.edu, juba.ziani@isye.gatech.com

Main Question

How does **information discrepancies** 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.

- Far-fetched assumption
- In reality: **institutions** rarely reveal their **decision rules** (reasons: privacy, proprietary software etc).
- Instead of full revelation: examples with explanations, examples of past decisions etc.

Setup at a High Level

- Agents belong in 2 subgroups (**green, blue**).
- Agents **do not know** the **decision rule**.
- Agents have information about past decisions 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**: $w^* \in \mathbb{R}^d$.
2. Learner uses **score rule** $w \in \mathbb{R}^d$ (hidden from agents).
3. Agents (per subgroup g) draw their private feature vectors: $x_1 \sim \mathcal{D}_1$ and $x_2 \sim \mathcal{D}_2$.
4. Given peer dataset S_g , private feature vector x_g , & their utility $u(x_g, x'_g; g)$, the agents best-respond with feature vector: $\hat{x}_g = \arg \max_{x'} u(x_g, x'; g)$.

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 $\mathcal{S}_1, \mathcal{S}_2$
- $x_g = \Pi_g x_g$ (feature discrepancy)

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

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

Why is $w^* \neq w$?

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

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

Learner's Rule

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

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

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

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

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.

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

Experiments

- **Datasets:** Taiwan-Credit, Adult

