Incentive Mechanisms for Strategic Classification and Regression Problems

KUN JIN, University of Michigan
XUERU ZHANG, Ohio State University
MOHAMMAD MAHDI KHALILI, Yahoo! Inc.
PARINAZ NAGHIZADEH, Ohio State University
MINGYAN LIU, University of Michigan

We study the design of a class of incentive mechanisms that can effectively prevent cheating in a strategic classification and regression problem. A conventional strategic classification or regression problem is modeled as a Stackelberg game, or a principal-agent problem between the designer of a classifier (the principal) and individuals subject to the classifier's decisions (the agents), potentially from different demographic groups. The former benefits from the accuracy of its decisions, whereas the latter may have an incentive to game the algorithm into making favorable but erroneous decisions. While prior works tend to focus on how to design an algorithm to be more robust to such strategic maneuvering, this study focuses on an alternative, which is to design incentive mechanisms to shape the utilities of the agents and induce effort that genuinely improves their skills, which in turn benefits both parties in the Stackelberg game. Specifically, the principal and the mechanism provider (which could also be the principal itself) move together in the first stage, publishing and committing to a classifier and an incentive mechanism. The agents are (simultaneous) second movers and best respond to the published classifier and incentive mechanism. When an agent's strategic action merely changes its observable features, it hurts the performance of the algorithm. However, if the action leads to improvement in the agent's true label, it not only helps the agent achieve better decision outcomes, but also preserves the performance of the algorithm. We study how a subsidy mechanism can induce improvement actions, positively impact a number of social well-being metrics, such as the overall skill levels of the agents (efficiency) and positive or true positive rate differences between different demographic groups (fairness).

CCS Concepts: • **Theory of computation** → **Algorithmic game theory**; **Algorithmic mechanism design**; • **Computing methodologies** → *Supervised learning by regression*; *Supervised learning by classification*.

Additional Key Words and Phrases: Subsidy mechanisms; Strategic classification; Strategic regression.

ACM Reference Format:
Kun Jin, Xueru Zhang, Mohammad Mahdi Khalili, Parinaz Naghizadeh, and Mingyan Liu. 2022. Incentive Mechanisms for Strategic Classification and Regression Problems. In *Proceedings of the 23rd ACM Conference on Economics and Computation (EC '22), July 11–15, 2022, Boulder, CO, USA.* ACM, New York, NY, USA, 31 pages. https://doi.org/10.1145/XXXXXX.XXXXXX

1 INTRODUCTION

This paper studies the impact of adding a subsidy mechanism in strategic classification and regression problems. Conventional strategic classification and regression model the interaction between a decision maker (algorithm designer) and individuals who are subject to the decision outcomes. While the former benefits from the accuracy of its decisions, the latter may have an incentive to *game* the

EC '22, July 11–15, 2022, Boulder, CO, USA.

ACM ISBN 978-1-4503-9150-4/22/07…$15.00
https://doi.org/10.1145/XXXXXX.XXXXXX

algorithm into making favorable but erroneous decisions. Recognizing the potential for such misuse, prior works tend to focus on designing an algorithm that is more robust to such strategic maneuvering, see e.g., [1–5, 7, 10, 13, 14]. Equally important, however, is the possibility for a mechanism designer to *incentivize* effort by the users who genuinely improve their true label; this would benefit the users and the decision maker by preserving the algorithm performance at the same time.

Toward this end, we present a strategic learning problem augmented by a subsidy mechanism (augmented strategic learning problem) modeled as a Stackelberg game between the decision maker, the mechanism designer (which could be the decision maker itself or a third party) and individuals from different demographic groups who are subject to the classifiers' decisions (the agents). The decision maker and the mechanism designer move in the first stage by publishing and committing to a decision rule (a binary classifier or a regression function) and an incentive mechanism. The published decision rule takes as input the agents' *observable* features and outputs decision outcomes that impact the agents' utilities. The agents are (simultaneous) second movers and best respond to the published decision rule and incentive mechanism. To capture the agent's ability to both game the decision rule and make real changes, we assume each agent has an endowed pre-response attribute (endowed private information), that is causal [13] to a set of observable features as well as its true label, also referred to as its *qualification status* in the context of the strategic learning problem.

An agent can exert effort to improve this causal state, thereby improving its features and its underlying attributes, or choose to game the classifier by employing non-causal schemes to improve only its features without changing its underlying attributes [13], or use a combination of them. Both choices of action, referred to as *improvement* (or honest effort) and *gaming* (or cheating, or dishonest effort), respectively, come at a cost to the agent. As pointed out in [14], gaming is much more frequently seen and studied due to its much lower cost compared to improvement. This difference in cost results in Goodhart's Law ("Once a measure becomes a target, it ceases to be a good measure" [17]), since gaming invariably degrades the performance of a classifier. The goal of this study is to see whether, beyond preventing gaming, the incentive mechanism can elicit sufficient *improvement* from the agents.

The decision maker derives its utility from the prediction accuracy, thus even a selfish decision maker may have an incentive to motivate the agents to choose improvements over gaming. When the decision maker is also the mechanism designer, one such incentive mechanism is for the decision maker to subsidize the agents' improvement costs, thereby making improvement more appealing compared to gaming. We characterize the Stackelberg equilibrium in this setting, where the decision maker determines the optimal decision rule as well as the incentive mechanism (a subsidy policy) in anticipation of the agents' best response. In addition, we also study the impact of the equilibrium classifier and incentive mechanism on the fairness and qualification status, when agents come from different demographic groups which differ in their pre-response attribute distribution (e.g., an advantaged group may have higher pre-response attributes that map to higher qualification rates and features) or action cost (e.g., an advantaged group may have lower action cost than a disadvantaged group). Alternatively, we also study the case where the mechanism designer is a third party (e.g., a government) with social well-being metrics as its objective. The third party designs a mechanism that incentivizes agents' improvement action and charges a price to the decision maker for this *improvement service* to ensure budget balance, while also making sure that incentive compatibility and individual rationality constraints are satisfied for both the agents and the decision maker. We compare the outcomes of the augmented strategic learning in these two settings, as well as the conventional strategic learning problem without an incentive mechanism, and investigate how the mechanism designer's objectives influence the fairness and qualification status.

Our work differs from previous works on incentive mechanisms in the presence of strategic agents [6, 10, 12, 16] in the following ways. Firstly, subsidies are also used in [10] for strategic classification;

however, all actions considered in [10] are gaming and thus all subsidies go toward gaming. Both our work and [10] show that subsidizing gaming is a strictly dominated strategy for the decision maker, but our work further shows the potential benefit of subsidizing improvement actions. Secondly, [6, 12] use the classifier decision rule as a proxy for designing incentives, while we take a combination of the decision rule and an incentive mechanism choice to provide incentives; this is noteworthy because there are cases where the decision rule alone fails to incentivize improvement, such that one can only resort to the incentive mechanism to serve this purpose (see further discussion in Section 2). Thirdly, the decision maker in our model is selfish (i.e. profit maximizing) and the third party optimizes social well-being metrics (e.g., social welfare, or fairness metrics); in contrast, the decision maker is welfare maximizing in [6], is either selfish or welfare maximizing in [16], and works toward effort profiles with desired characteristics in [12]. Fourthly, while [16] focuses on the linear regression problem and [6, 12] on binary classification problems, we study both types of problems and elaborate on the similarities and differences between these setups. Our main contributions are as follows.

(1) We formulate the problem of adding a subsidy mechanism in strategic classification and regression problems as a Stackelberg game, where the decision maker and mechanism designer commit to a classifier and an incentive mechanism, and agents follow by choosing an action to best respond (Section 2). This model substantially extends existing literature.

(2) We begin with the setting in which the decision maker is the mechanism designer, and study the incentive mechanism design and the Stackelberg equilibrium of the classification and regression models (Sections 3 and 4). We identify conditions under which the incentive mechanisms satisfy individual rationality, incentive compatibility, and budget balance.

(3) We study the social well-being of the augmented strategic learning system, focusing on both efficiency and fairness properties (Section 5). We also consider the case of a third party mechanism designer, and discuss its influence on these social well-being metrics (Section 6).

(4) We illustrate our analytical findings through numerical experiments based on the FICO dataset [8] (Section 7).

2 MODEL

We first introduce our augmented strategic learning model. In particular, we focus on a single-round, two-stage Stackelberg game, where the decision maker and the mechanism designer move first to design, publish, and commit to a decision rule f combined with an incentive mechanism G; the agents then best respond to both the incentive mechanism and the decision rule in the second stage.

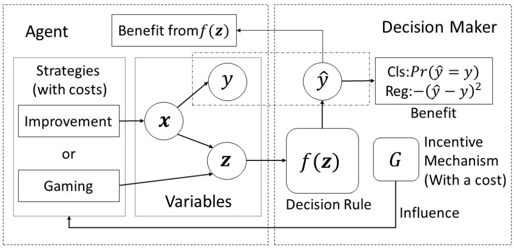


Fig. 1. The Augmented Strategic Classification/Regression Problem.

2.1 Attributes, Features, and Labels

An agent has an N-dimensional *pre-response attribute* $\boldsymbol{x} \in \mathcal{X}, \mathcal{X} \subseteq \mathbb{R}^N_{\geq 0}$, which is its private information. Its probability density function (pdf) is $p(\boldsymbol{x})$, which is public information. In the response phase, an agent takes an M-dimensional action $\boldsymbol{a} := (\boldsymbol{a}_+, \boldsymbol{a}_-)$, where $\boldsymbol{a}_+ \in \mathbb{R}^{M_+}_{\geq 0}$ denotes an *improvement action* profile while $\boldsymbol{a}_- \in \mathbb{R}^{M_-}_{\geq 0}$ is a *gaming action* profile, with $M_+ + M_- = M$ (with action indices ordered such that $\forall i \leq M_+$ is an improvement action).

The agent's action impacts its attribute as well as feature through a *projection matrix*

$P = [P_+, P_-], P \geq 0$, where $P_+ \in \mathbb{R}^{N \times M_+}$ (resp. $P_- \in \mathbb{R}^{N \times M_-}$) is the improvement (resp. gaming) projection in the following sense. The action results in the agent having a *post-response attribute* $\boldsymbol{x}' = \boldsymbol{x} + P_+ \boldsymbol{a}_+ = \boldsymbol{x} + \hat{P}\boldsymbol{a}$, where $\hat{P} = [P_+, \boldsymbol{0}] \in \mathbb{R}^{N \times M}$, and a post-response *observable feature* (simply feature for brevity) $\boldsymbol{z} = \boldsymbol{x} + P\boldsymbol{a} = \boldsymbol{x} + P_+\boldsymbol{a}_+ + P_-\boldsymbol{a}_-$. Crucially, the post-response attribute is the agent's private information, whereas the post-response feature is observable by the decision maker. An agent's action may or may not be directly observable to the decision maker, but is anticipated given the game setting and an equilibrium concept.

This model captures the fact that improvement actions can improve an agent's underlying attribute as well as observable feature, while gaming actions only affect the outward feature without changing its underlying attribute. We can think of attributes as actual skills and features as test scores; working hard can be a type of improvement action and cheating can be a type of gaming action.

In general, the projection matrix P is not full rank, which means there are multiple choices of $\boldsymbol{a}$ for the agent to obtain the same feature $\boldsymbol{z}$ and thus the same decision outcome (next subsection).

An agent with pre- (resp. post-)response attribute $\boldsymbol{x}$ (resp. $\boldsymbol{x}'$) has a pre- (resp. post-)response *true label* y (resp. y') that indicates the quality of an agent. For strategic regression, we use the same setting as in [16]:

$$y = q(\boldsymbol{x}) := \boldsymbol{\theta}^T \boldsymbol{x} + \eta, \; y' = q(\boldsymbol{x}') = \boldsymbol{\theta}^T \boldsymbol{x}' + \eta, \tag{1}$$

where $\boldsymbol{\theta} \geq 0$ is the quality coefficient vector, and η is a subgaussian noise with 0 mean and variance σ. For strategic classification, $y, y' \in \{0, 1\}$, and we use a similar setting as in [10]:

$$Pr(y = 1) = l(\boldsymbol{\theta}^T \boldsymbol{x}), \; Pr(y' = 1) = l(\boldsymbol{\theta}^T \boldsymbol{x}'), \; y' \geq y, \tag{2}$$

where we can interpret $l : \mathbb{R} \mapsto [0, 1]$ as a likelihood function that is weakly increasing (l is a step-function in [10]). We assume that $y' \geq y$ holds for every agent, with improvement actions weakly improving the agent's true label, and gaming actions leaving it unchanged.

REMARK 1. *The projection matrix P, the available action dimensions, and the quality coefficients* $\boldsymbol{\theta}$ *are assumed to be public information for the remainder of the paper. We discuss in the appendix when these parameters are initially unknown for the decision maker. Parameter acquisition requires multi-round online learning [9, 16], which is different from the model setting in this paper. However, we show that our incentive mechanisms can aid parameter learning in the multi-round online strategic learning models.*

2.2 The Decision Rule

The decision rule $f : \mathbb{R}^N \mapsto \mathbb{R}$ takes as input an agent's feature $\boldsymbol{z}$ and returns a decision outcome $f(\boldsymbol{z})$. For regression, $f(\boldsymbol{z}) = \boldsymbol{w}^T \boldsymbol{z}$; for classification, $f(\boldsymbol{z}) = \boldsymbol{1}\{\boldsymbol{w}^T \boldsymbol{z} \geq \tau\}$, for some $\boldsymbol{w} \in \mathbb{R}^N_{\geq 0}$ (since the true labels are weakly increasing in every attribute).

2.3 Three Learning/Game Problems

We will consider three different strategic learning systems/game settings:

(1) The *conventional strategic (CS)* problem where the agents and the decision maker play the standard Stackelberg game without any added incentive mechanism, both being fully strategic.

(2) The *limited strategic (LS)* problem where the agents are fully strategic and expect the decision maker to be strategic, but the latter does not anticipate the agents' strategic behavior and applies the optimal non-strategic decision rule, e.g., $f(\boldsymbol{z}) = \boldsymbol{\theta}^T \boldsymbol{z}$ in regression, as a sub-optimal option.[1]

[1] The agents in LS behave the same as in CS problems. One reason to consider LS is because the CS problem is in general NP-hard for the decision maker [12].

(3) The *augmented strategic (AS)* problem, where the agents and the decision maker play the Stackelberg game with a subsidy mechanism.

We use the CS and LS problems as benchmarks to show how subsidy mechanisms influence the equilibrium system outcome. We next detail the utility functions and the incentive mechanism.

2.4 Utilities and Optimal Strategies in Conventional & Limited Strategic Learning

In a conventional strategic learning problem, it is assumed that an agent has the following utility function $u_C(\boldsymbol{x}, \boldsymbol{a}) = f(\boldsymbol{x} + P\boldsymbol{a}) - h(\boldsymbol{a})$, where the agent benefits from the decision outcome $f(\boldsymbol{z})$ and incurs a cost of $h(\boldsymbol{a}) := \boldsymbol{c}^T\boldsymbol{a}$.

Denote by $\boldsymbol{a}^*_C(\boldsymbol{x}) := \arg\max_{\boldsymbol{a}} u_C(\boldsymbol{x}, \boldsymbol{a})$ the agent's *conventional best response* or *CS best response*, with ties broken in favor of its qualification status $\boldsymbol{\theta}^T\boldsymbol{x}'$. In the same problem, denote y'_C as the *CS post-response label*. The decision maker's utility is

$$
\begin{aligned}
U^{(cls)}_C(f) &= \int_X Pr\big(f(\boldsymbol{x} + P\boldsymbol{a}^*_C(\boldsymbol{x})) = y'_C\big)p(\boldsymbol{x})d\boldsymbol{x}; \\
U^{(reg)}_C(f) &= \int_X \mathbb{E}_\eta\big[-\big(f(\boldsymbol{x} + P\boldsymbol{a}^*_C(\boldsymbol{x})) - y'_C\big)^2\big]p(\boldsymbol{x})d\boldsymbol{x} \qquad (3)
\end{aligned}
$$

for strategic classification and regression, respectively. Here the decision maker aims to maximize the classification accuracy and minimize the mean squared error in regression, respectively. We will use $f^*_C := \arg\max_f U_C(f)$ to denote the decision maker's optimal conventional strategic decision rule, where the type of problem (*cls* vs. *reg*) will be clear from context. In the limited strategic (LS) problem, the agents' utilities and best responses are the same as the CS problem, but the decision maker instead maximizes, respectively:

$$
\begin{aligned}
U^{(cls)}_L(f) &= \int_X Pr\big(f(\boldsymbol{x}) = y\big)p(\boldsymbol{x})d\boldsymbol{x}; \\
U^{(reg)}_L(f) &= \int_X \mathbb{E}_\eta\big[-\big(f(\boldsymbol{x}) - y\big)^2\big]p(\boldsymbol{x})d\boldsymbol{x}. \qquad (4)
\end{aligned}
$$

REMARK 2. *Our findings generalize to other cost functions such as L2 cost* $h(\boldsymbol{a}) = \sqrt{\boldsymbol{a}^T C\boldsymbol{a}}$ *or quadratic cost* $h(\boldsymbol{a}) = \frac{1}{2}||\boldsymbol{a}||_2^2$. *More details are provided in the appendix.*

2.5 Incentive Mechanisms and Utilities in Augmented Strategic Learning

Different from previous works, we focus on how an incentive mechanism can influence the strategic interaction between the decision maker and the agents. We consider two types of mechanism providers. We will start with the setting where the mechanism provider is the decision maker itself. Our analysis and results are then extended in Section 6 to a second setting where the mechanism is provided or implemented by a *third-party* organization, e.g., the government.

We focus on *discount mechanisms* that are based on providing a *discount on actions*, where the mechanism provider has the ability to lower the cost of agents' actions, e.g., making the cost of getting tutoring or exam preparation cheaper during the school admission process.[2] We use G to denote the discount mechanism where the designer chooses a *rate discount* value on each action dimension $\triangle\boldsymbol{c} = (\triangle c_i)_{i=1}^M$, $\triangle c_i < c_i$, and set a *discount amount range* $[\underline{c}, \overline{c}]$. Then with G, the agent's utility function in the augmented strategic learning problem becomes

$$
u_A(\boldsymbol{x}, \boldsymbol{a}) = f(\boldsymbol{x} + P\boldsymbol{a}) - h_A(\boldsymbol{a}), \text{ where } h_A(\boldsymbol{a}) = h(\boldsymbol{a}) - \triangle\boldsymbol{c}^T\boldsymbol{a} \cdot \mathbf{1}\{\triangle\boldsymbol{c}^T\boldsymbol{a} \in [\underline{c}, \overline{c}]\}. \qquad (5)
$$

[2] In the appendix, we discuss an alternative mechanism where the designer cannot change the action cost, and show that the resulting mechanism design problem is computationally intractable.

With G, $\boldsymbol{a}_A^*(\boldsymbol{x}) := \arg\max_{\boldsymbol{a}} u_A(\boldsymbol{x}, \boldsymbol{a})$ denotes the agent's *augmented best response* or *AS best response*, with ties broken in favor of maximizing $\boldsymbol{\theta}^T\boldsymbol{x}'$ unless otherwise suggested by the mechanism designer. The designer incurs a *subsidy cost*

$$H(G) = \int_{\mathcal{X}} \triangle\boldsymbol{c}^T\boldsymbol{a}_A^*(\boldsymbol{x}) \cdot \mathbf{1}\{\triangle\boldsymbol{c}^T\boldsymbol{a}_A^*(\boldsymbol{x}) \in [\underline{c}, \overline{c}]\}p(\boldsymbol{x})d\boldsymbol{x}. \tag{6}$$

Denote by y_A' the *AS post-response label*. The augmented utility of the decision maker is then:

$$\begin{aligned} U_A^{(cls)}(f) &= \int_{\mathcal{X}} Pr\big(f(\boldsymbol{x} + P\boldsymbol{a}_A^*(\boldsymbol{x})) = y_A'\big)p(\boldsymbol{x})d\boldsymbol{x} - H(G); \\ U_A^{(reg)}(f) &= \int_{\mathcal{X}} \mathbb{E}_\eta\big[-\big(f(\boldsymbol{x} + P\boldsymbol{a}_A^*(\boldsymbol{x})) - y_A'\big)^2\big]p(\boldsymbol{x})d\boldsymbol{x} - H(G), \end{aligned} \tag{7}$$

for the classification and regression problems, respectively. In designing G, we will consider three commonly studied properties in the mechanism design literature:

(1) Individual rationality (IR): The agents are better off participating in the mechanism than not.
(2) Incentive compatibility (IC): The agents act in self-interest.
(3) Budget balance (BB): This only applies to the third party mechanism; see Section 6.

3 AUGMENTED STRATEGIC CLASSIFICATION

In this and the next section, we consider agents from a single demographic group. Throughout our analysis, we will provide pictorial interpretations of our results, using an example with 2 action dimensions: a_1 is an improvement action and a_2 is a gaming action.

We begin with some preliminaries. The next two lemmas characterize the magnitude and direction of the agents' best responses $\boldsymbol{a}_t^*(\boldsymbol{x})$ ($t \in \{C, A\}$) in the conventional and augmented strategic games.

LEMMA 3.1. *For CS and AS classification,* $\boldsymbol{w}^T(\boldsymbol{x} + P\boldsymbol{a}_t^*(\boldsymbol{x})) = \tau \Leftrightarrow \boldsymbol{a}_t^*(\boldsymbol{x}) \neq \mathbf{0}, \forall t.$

PROOF. For $\forall\boldsymbol{a}$ such that $\boldsymbol{w}^T(\boldsymbol{x} + P\boldsymbol{a}) < \tau$, $f(\boldsymbol{z}) = 0$; thus it is dominated by $\mathbf{0}$ due to $h(\boldsymbol{a}) \geq h(\mathbf{0}) = 0$ and $h_A(\boldsymbol{a}) \geq h_A(\mathbf{0}) = 0$. On the other hand, for $\forall\boldsymbol{a}$ such that $\boldsymbol{w}^T(\boldsymbol{x} + \boldsymbol{a}^*) > \tau$, there exists an $\alpha \in (0, 1)$ such that $\boldsymbol{w}^T(\boldsymbol{x} + \alpha P\boldsymbol{a}) = \tau$. Both $\boldsymbol{a}$ and $\alpha\boldsymbol{a}$ guarantee $f(\boldsymbol{z}) = 1$, and thus $\boldsymbol{a}$ is dominated by $\alpha\boldsymbol{a}$ due to $h(\boldsymbol{a}) > h(\alpha\boldsymbol{a})$ and $h_A(\boldsymbol{a}) > h_A(\alpha\boldsymbol{a})$ if $\boldsymbol{a} \neq \mathbf{0}$. □

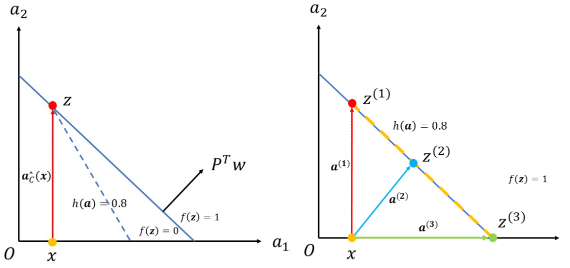
Lemma 3.1 describes the magnitude of the best response in CS and AS classification: it is such that the feature $\boldsymbol{z}$ reaches the decision boundary but not beyond, as going beyond the boundary only increases the cost without affecting the decision. This is illustrated by the red arrow in Figure 2a.

LEMMA 3.2. *For CS and AS classification,*

$$\begin{aligned} &(\boldsymbol{a}_C^*(\boldsymbol{x}))_i \geq 0, \ \textit{if}\ i \in \{\arg\max_j (P^T\boldsymbol{w})_j/c_j\};\ (\boldsymbol{a}_t^*(\boldsymbol{x}))_i = 0,\ \textit{o.w.},\ \forall\boldsymbol{x}. \\ &(\boldsymbol{a}_A^*(\boldsymbol{x}))_i \geq 0, \ \textit{if}\ i \in \{\arg\max_j (P^T\boldsymbol{w})_j/(c_j - \triangle c_j)\};\ (\boldsymbol{a}_A^*(\boldsymbol{x}))_i = 0,\ \textit{o.w.},\ \forall\boldsymbol{x}. \end{aligned} \tag{8}$$

PROOF. Assume by contradiction $a_j^* > 0, j \neq i_C = \arg\max_k \frac{(P^T\boldsymbol{w})_k}{c_k}$. By Lemma 3.1, as $\boldsymbol{a}^* \neq \mathbf{0}$ we have $\boldsymbol{w}^T(\boldsymbol{x} + P\boldsymbol{a}^*) = \tau$. Denote $\tilde{\boldsymbol{a}} = \boldsymbol{a}^* - a_j^*\boldsymbol{e}_j + \frac{a_j^*(P^T\boldsymbol{w})_j}{(P^T\boldsymbol{w})_{i_C}}\boldsymbol{e}_{i_C}$, where $\boldsymbol{e}_i$ is the i-th orthonormal base vector of $\mathbb{R}^M$. It is easy to see that $\boldsymbol{w}^T(\boldsymbol{x} + P\tilde{\boldsymbol{a}}) = \tau$ and thus $f(\boldsymbol{z}) = 1$, while $h(\tilde{\boldsymbol{a}}) < h(\boldsymbol{a}^*)$, indicating that $\tilde{\boldsymbol{a}}$ achieves a higher utility than $\boldsymbol{a}^*$, contradicting the optimality of $\boldsymbol{a}^*$. The proof for AS classification is similar. □

Lemma 3.2 describes the directional properties of the best response: the agent will invest in the action dimension(s) with the highest *return on investment* $(P^T\boldsymbol{w})_j/c_j$ (in CS) or $(P^T\boldsymbol{w})_j/(c_j -$



(a) CS Best Response

(b) Tie Breaking

Fig. 2. An illustration of a CS best response in classification, where $P = [1, 1]$, $w = 1$, $P^T w = (1, 1)$. The solid blue line is the decision boundary. In (a), the blue dashed line is an equal cost contour; $c_2 < c_1$, thus gaming is cheaper than improving leading to the best response shown in red. (b) illustrates tie breaking in best responses, where $c_1 = c_2$, with the equal cost contour shown with the yellow dashed line.

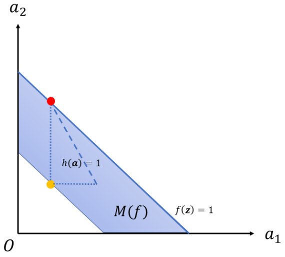


Fig. 3. An illustration of the manipulation margin in classification, given by the shaded region; every agent inside can reach the boundary with an action cost no more than 1.

$\triangle c_j$) (in AS). Without loss of generality, we assume that the optimal CS action dimension $i_C := \arg\max_j (P^T \boldsymbol{w})_j / c_j$ is unique. This property is also shown in Figure 2a, where $i_C = 2$ is the action with the highest return on investment.

We note that there may be multiple actions that are tied in their return on investment. In such cases, we assume the agent follows the algorithm designer's recommendation if any, and otherwise chooses the one that leads to the maximum improvement (i.e., the one maximizing $\boldsymbol{\theta}^T \hat{P} \boldsymbol{a}$). Figure 2b explains this tie breaking: here $c_1 = c_2$ and every point on the yellow contour has equal cost and benefit to the agent, making the agent indifferent between $\boldsymbol{a}^{(1)}, \boldsymbol{a}^{(2)}, \boldsymbol{a}^{(3)}$. We take $\boldsymbol{a}^{(3)}$, the largest improvement, to be the agent's choice.

Using Lemmas 3.1 and 3.2, we have

$$\boldsymbol{a}_C^*(\boldsymbol{x}) = \frac{\tau - \boldsymbol{w}^T \boldsymbol{x}}{(P^T \boldsymbol{w})_{i_C}} \boldsymbol{e}_{i_C}, \text{ if } \boldsymbol{x} \in \mathcal{M}(f); \; \boldsymbol{a}_C^*(\boldsymbol{x}) = \boldsymbol{0}, \text{ o.w.,} \tag{9}$$

where $\mathcal{M}(f) := \left\{ \boldsymbol{x} \middle| \frac{(\tau - \boldsymbol{w}^T \boldsymbol{x}) c_{i_C}}{(P^T \boldsymbol{w})_{i_C}} \in (0, 1] \right\}$ denotes the *manipulation margin* of f: every agent in the manipulation margin has non-zero best response to improve their decision outcome to 1. This is illustrated in Figure 3.

In classification, if $i_C \leq M_+$, we say the decision rule *incentivizes improvement*, otherwise we say the decision rule *incentivizes gaming*. The theorem below shows conditions under which it is impossible for the decision maker to have a decision rule that incentivizes improvement; the proof is given in Appendix B.2.

THEOREM 3.3. *Let κ_i denote the substitutability of action dimension i [11, 12]. Formally,*

$$\kappa_i := \min_{\boldsymbol{a} \in \mathbb{R}^M, \boldsymbol{a} \geq 0} \frac{\boldsymbol{c}^T \boldsymbol{a}}{c_i}, \; s.t. \; P\boldsymbol{a} - \boldsymbol{p}_i \geq 0, \tag{10}$$

where $\boldsymbol{p}_i$ is the i-th column of P. If $\kappa_i = 1$, then there exists a $\boldsymbol{w}$ that can incentivize action dimension i, and the $\boldsymbol{w}$ can be found in polynomial time. On the other hand, if $\kappa_i < 1, \forall i \leq M_+$, then there always exist linear combinations of gaming actions that weakly dominate every improvement action, in which case there is no f that can incentivize improvement, and the decision maker's CS optimal strategy f_C^* satisfies $\boldsymbol{w} = \boldsymbol{\theta}$.

We next consider designing an incentive mechanism, with the decision rule f treated as given.

LEMMA 3.4. *To induce an agent to take an AS best response with non-zero investment in action dimension $j \leq M_+$, i.e., $[\boldsymbol{a}_A^*(\boldsymbol{x})]_j > 0$, the discount value $\triangle c_j$ should satisfy $(P^T\boldsymbol{w})_j/(c_j - \triangle c_j) \geq (P^T\boldsymbol{w})_{i_C}/(c_{i_C})$, i.e., $\triangle c_j \geq c_j - \frac{(P^T\boldsymbol{w})_j}{(P^T\boldsymbol{w})_{i_C}} c_{i_C}$.*

Based on Lemma 3.4, we denote the *minimum effective discount value* as

$$\triangle c_j^* := c_j - \frac{(P^T\boldsymbol{w})_j}{(P^T\boldsymbol{w})_{i_C}} c_{i_C}. \tag{11}$$

Intuitively, Lemma 3.4 states that to induce a best response in action j, the discount has to make j the action with the highest (potentially tied) return on investment. Figure 4a illustrates an example of how the discount mechanism with minimum effective discount value works. By choosing $\triangle c_1 = \triangle c_1^*$, the two actions have the same return on investment; the agents choose $\boldsymbol{a}_A^*(\boldsymbol{x})$, the maximum improvement action, in this case. In contrast, the CS action $\boldsymbol{a}_C^*(\boldsymbol{x})$ is a gaming action.

Before we move on to the optimal mechanism design, we define the *subsidy surplus.*

DEFINITION 1. *In classification the subsidy surplus is*

$$S(f,G) = \int_X \left[Pr(f(\boldsymbol{x} + P\boldsymbol{a}_A^*(\boldsymbol{x})) = y_A') - Pr(f(\boldsymbol{x} + P\boldsymbol{a}_C^*(\boldsymbol{x})) = y_C')\right] p(\boldsymbol{x}) d\boldsymbol{x} - H(G), \tag{12}$$

where y_t' denotes the post-response label such that $Pr(y_t' = 1) = l(\boldsymbol{x} + \hat{P}\boldsymbol{a}_t^*(\boldsymbol{x})), \forall t \in \{C, A\}$.

The integral part in $S(f,G)$ is the *benefit gain* of the decision maker and the value in the square bracket is the *individual subsidy benefit*. The decision maker's problem is equivalent to maximizing $S(f,G)$ under IC and IR.

THEOREM 3.5. *For general $f(\boldsymbol{z}) = \mathbf{1}\{\boldsymbol{w}^T\boldsymbol{z} \geq \tau\}$, p, and l functions, finding the optimal IC and IR discount mechanism requires solving non-convex optimization problems and thus is NP-hard.*

While finding the optimal mechanism under IC and IR constraints is NP-hard, we can develop an efficient algorithm (Algorithm 1) for a special case when the likelihood function l is convex.

THEOREM 3.6. *Algorithm 1 runs in polynomial time, and if l is convex on $[0, \max_{\boldsymbol{x}:\boldsymbol{w}^T\boldsymbol{x}=\tau} l(\boldsymbol{x})]$, then any $G \neq 0$ returned by Algorithm 1 is IC, IR, and satisfies $S(f,G) \geq 0$.*

From Algorithm 1 we see that the decision maker prefers subsidizing agents that are "closer" to the boundary when l is convex on $[0, \max_{\boldsymbol{x}:\boldsymbol{w}^T\boldsymbol{x}=\tau} l(\boldsymbol{x})]$. This is because when l is convex, the subsidy benefit becomes concave while the subsidy cost is linear in the "distance to the boundary"; thus the agents close enough to the boundary can have positive individual subsidy surplus; Figure 5 provides an illustration of this.

ALGORITHM 1: Find a $G \neq 0$ that is IC, IR and $S(f,G) > 0$ for Classification

```
$\boldsymbol{x}_1 \leftarrow \arg\min_{\boldsymbol{x}:\boldsymbol{w}^T\boldsymbol{x}=\tau} \boldsymbol{\theta}^T\boldsymbol{x}$;
$i_C \leftarrow \arg\max_j (P^T\boldsymbol{w})_j / c_j$;
for $j = 1 : M_+$ do
    $\triangle\boldsymbol{c} \leftarrow \mathbf{0}$; $\bar{c} \leftarrow 0$; $l_+ \leftarrow 0$;
    $\triangle c_j \leftarrow c_j - \frac{(P^T\boldsymbol{w})_j}{(P^T\boldsymbol{w})_{i_C}} c_{i_C}$;
    Define function
    $\boldsymbol{a}(\delta) := \delta\boldsymbol{e}_j - \delta\frac{(P^T\boldsymbol{w})_j}{(P^T\boldsymbol{w})_{i_C}}\boldsymbol{e}_{i_C}$;
    $l_+(\delta) := l(\boldsymbol{\theta}^T\boldsymbol{x}_1) - l(\boldsymbol{\theta}^T\boldsymbol{x}_1 - \boldsymbol{a}(\delta))$;

    $\delta^* \leftarrow \arg\max_\delta \; s.t. \; l_+(\delta) \geq \delta\triangle c_j$;
    if $\delta^* = 0$ then
        Go back to for loop
    end
    $\bar{c} \leftarrow \min\{\delta^*, 1/(c_j - \triangle c_j)\} \cdot \triangle c_j$;
    Return $(\triangle\boldsymbol{c}, 0, \bar{c})$
end
Return $(\mathbf{0}, 0, 0)$
```

The convexity requirement of l on a low range is satisfied in real-world datasets such as the FICO credit score dataset, in which the likelihood function l frequently has an S-shape (see Section 7). We discuss the case of other likelihood function types (including concave) in the appendix. Also note that in Algorithm 1 the mechanism designer places discount on only one dimension. This is because even though it technically can set the discount $\triangle c_i > 0$ for multiple improvement actions, ultimately the agent either finds the dimension with the highest return on investment or breaks ties in favor of the largest improvement.[3]

The optimal mechanism can be found more efficiently for the special case when $\boldsymbol{w} = \boldsymbol{\theta}$ in f (this happens, e.g., in the optimal LS strategy as shown in Lemma B.1 in the appendix, or in the optimal CS strategy when $\kappa_i < 1, \forall i \leq M_+$ in Theorem 3.3). This can be done in a fixed number of steps (faster than polynomial) using Algorithm 2.

ALGORITHM 2: A G that is IC, IR and $S(f, G) \geq 0$ for Classification when $\boldsymbol{w} = \boldsymbol{\theta}$

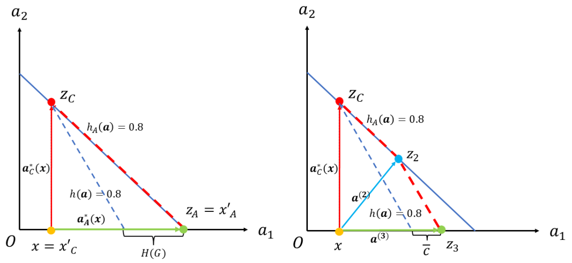
```
$i_A \leftarrow \arg\max_{j \leq M_+} (P^T\boldsymbol{\theta})_j / c_j$;
$\triangle \boldsymbol{c} \leftarrow \boldsymbol{0}$;  $\triangle c_{i_A} \leftarrow c_{i_A} - \frac{(P^T\boldsymbol{\theta})_{i_A}}{(P^T\boldsymbol{\theta})_{i_C}} c_{i_C}$;
Define functions
$s_1(r) := l(\tau) - l(r) - \frac{(\tau - r)\triangle c_{i_A}}{(P^T\boldsymbol{\theta})_{i_A}}$;
$s_2(r) := l(\tau) + l(r) - 1 - \frac{(\tau - r)\triangle c_{i_A}}{(P^T\boldsymbol{\theta})_{i_A}}$;
$r \leftarrow \arg\min_r \; s.t. \; s_1(r) \geq 0$;
**if** $l(r) < 0.5$ **then**
    $r \leftarrow \arg\min_r \; s.t. \; s_2(r) \geq 0$;
**end**
$\overline{c} = (\tau - r)\triangle c_{i_A} / (P^T\boldsymbol{\theta})_{i_A}$;
Return $(\triangle \boldsymbol{c}, 0, \overline{c})$.
```

THEOREM 3.7. *If $\boldsymbol{w} = \boldsymbol{\theta}$ in f, f incentivizes gaming, and l is convex on $[0, \tau]$, then Algorithm 2 finds a G that is IC, IR, and satisfies $S(f, G) \geq 0$. In addition, algorithm 2 finds the optimal G if $l(\tau) - l(\underline{r}_f) \leq \frac{(\tau - \underline{r}_f)\triangle c^*_{i_A}}{(P^T\boldsymbol{\theta})_{i_A}}$, where $\underline{r}_f = \min_{\boldsymbol{x} \in \mathcal{M}(f)} \boldsymbol{\theta}^T\boldsymbol{x}$.*

Intuitively, the condition $l(\tau) - l(\underline{r}_f) \leq \frac{(\tau - \underline{r}_f)\triangle c^*_{i_A}}{(P^T\boldsymbol{\theta})_{i_A}}$ indicates the subsidy cost is larger than the subsidy gain for an agent on the "far side" boundary of $\mathcal{M}(f)$ in (9). This holds when improvement costs are much larger than gaming costs, so that the discount payment is higher than the resulting benefit from the agent's improvement. Such a condition is needed to enable the efficient calculation of the optimal mechanism for the following reason. If the condition does not hold, the mechanism can further increase the cost discount rate on the actions and let agents with a pre-response attribute such that $\boldsymbol{\theta}^T\boldsymbol{x} < \underline{r}_f$ to also take improvement actions. However, this would again make the problem hard for the decision maker, since it has to jointly optimize $\triangle c_j$ and $\overline{c}$, and such optimization is non-convex.

We note that the s_1 and s_2 functions in Algorithm 2 capture the following properties of individual subsidy surplus: for agents in $\mathcal{M}(f)$, these agents' qualification status improvement equals the individual subsidy benefit $l(\boldsymbol{\theta}^T\boldsymbol{x}'_A) - l(\boldsymbol{\theta}^T\boldsymbol{x}'_C)$, but for agents not in $\mathcal{M}(f)$, the individual subsidy benefit is not the qualification status improvement, but instead $l(\boldsymbol{\theta}^T\boldsymbol{x}'_A) - [1 - l(\boldsymbol{\theta}^T\boldsymbol{x}'_C)]$ since these agents are supposed to receive 0 decision outcomes (rejections) in the CS problem. The green curve in Figure 5 also illustrates the above.

[3] When placing discounts on multiple actions, finding the optimal tie-breaking rule is a non-convex problem.


a_2
z_C
$h_A(\boldsymbol{a}) = 0.8$
$\boldsymbol{a}_C^*(\boldsymbol{x})$
$h(\boldsymbol{a}) = 0.8$
$\boldsymbol{a}_A^*(\boldsymbol{x})$
$z_A = x'_A$
a_1
O
$x = x'_C$
$H(G)$
a_2
z_C
$h_A(\boldsymbol{a}) = 0.8$
z_2
$\boldsymbol{a}_C^*(\boldsymbol{x})$
$\boldsymbol{a}^{(2)}$
$h(\boldsymbol{a}) = 0.8$
a_1
O
x
$\boldsymbol{a}^{(3)}$
$\bar{c}$
z_3

(a) Discount Mechanism (b) Designer's Suggestion

Fig. 4. An illustration of the discount mechanism in classification, $P = [1, 1], w = 1, P^T w = (1, 1), c_2 < c_1$, the red dashed line is the discounted equal cost contour with a minimum effective discount. In Figure 4b, the $\bar{c}$ is of a smaller value, and the equal cost contour has a different shape. The decision maker suggests the agents choose $\boldsymbol{a}_C^*(\boldsymbol{x})$ instead of $\boldsymbol{a}^{(3)}$ in tie breaking in Algorithm 1 and 2 when l is convex.

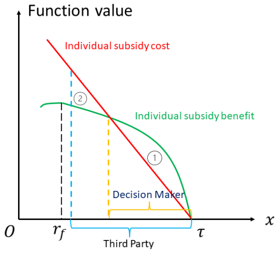


Fig. 5. A simplified illustration of the individual subsidy benefit and cost in the mechanism. Region 1 (resp. 2) corresponds to agents with subsidy surplus (resp. deficit). The third party (Section 6) incentivizes region 2 agents for social well-being objectives. r_f represents the lower boundary of $\mathcal{M}(f)$.

4 AUGMENTED STRATEGIC REGRESSION

We now turn to the strategic regression problem. For CS and AS regression, the best response directions are the same as CS and AS classification, as given in Lemma 3.2.

However, different from the strategic classification problem, the agents can have best responses with infinite magnitude. For example, if $(P^T \boldsymbol{w})_{i_C} \geq c_{i_C}$, the agent will invest an infinite amount in action i_C. To handle this issue, we assume that the agents' actions are bounded by an action budget $h(\boldsymbol{a}) \leq B$ in CS (and LS) regression, and $h_A(\boldsymbol{a}) \leq B$ in AS regression.[4]

Given these bounds on the agents' budgets, the agents' best responses can be characterized as follows: if $(P^T \boldsymbol{w})_{i_C} \geq c_{i_C}$, then $\boldsymbol{a}_C^*(\boldsymbol{x}) = \frac{B}{c_{i_C}} \boldsymbol{e}_{i_C}$; otherwise $\boldsymbol{a}_C^*(\boldsymbol{x}) = \boldsymbol{0}$. Similarly, let $i_A = \arg\max_j (P^T \boldsymbol{w})_j / (c_j - \triangle c_j)$, if $(P^T \boldsymbol{w})_{i_A} \geq c_{i_A} - \triangle c_{i_A}$. Then, the AS-discount best response is $\boldsymbol{a}_A^*(\boldsymbol{x}) = \frac{B}{c_{i_A} - \triangle c_{i_A}} \boldsymbol{e}_{i_A}$; otherwise $\boldsymbol{a}_A^*(\boldsymbol{x}) = \boldsymbol{0}$.

An interesting difference to highlight is that the agents' best responses in strategic classification depend on both the pre-response attributes of the agents and the decision rule, whereas in strategic regression, the best responses are the same for all agents and only depend on the decision rule.

In this strategic regression setting, we will say f incentivizes *0 responses* if $\boldsymbol{a}_C^*(\boldsymbol{x}) = \boldsymbol{0}$. Otherwise, if $i_C \leq M_+$ (resp. $i_C > M_+$), we say f incentivizes improvement (resp. gaming).

If f incentivizes non-zero responses (improvement or gaming), the cost discount rates will again follow Lemma 3.4, with the minimum effective discount rate still the same as in (11); otherwise, the minimum effective cost discount rate on action j will be such that $(P^T \boldsymbol{w})_j = (c_j - \triangle c_j)$, $\triangle c_j^* = \max\{c_j - c_{i_C} (P^T \boldsymbol{w})_j / (P^T \boldsymbol{w})_{i_C}, c_j - (P^T \boldsymbol{w})_j\}$.

Using this, the error incurred by the designer on an agent with pre-response attributes $\boldsymbol{x}$ will consist of two parts, an *equilibrium coefficient error* and an inevitable error due to noises,

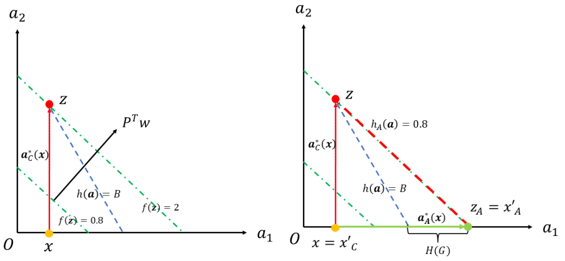
$$\mathcal{E}(f, \boldsymbol{a}, \boldsymbol{x}) = [\boldsymbol{w}^T (\boldsymbol{x} + P\boldsymbol{a}) - \boldsymbol{\theta}^T (\boldsymbol{x} + \hat{P}\boldsymbol{a})]^2 + err(\eta). \tag{13}$$

Note that although the agents' best responses are independent of $\boldsymbol{x}$, the equilibrium individual errors depend on $\boldsymbol{x}$ for any $\boldsymbol{w} \neq \boldsymbol{\theta}$.

We next consider the problem of designing an incentive (discount) mechanism.

[4] Such bound was not needed in the classification setting, as the fact that $f(\boldsymbol{z}) \leq 1$ naturally provided this.

THEOREM 4.1. *For general $f(\boldsymbol{z}) = \boldsymbol{w}^T\boldsymbol{z}$ and $p(\boldsymbol{x})$, finding the optimal IC, IR, and discount mechanism requires solving non-convex optimization problems and thus is NP-hard.*


a_2
z
$P^T w$
$\boldsymbol{a}_C^*(\boldsymbol{x})$
$h(\boldsymbol{a}) = B$
$f(\boldsymbol{z}) = 2$
$f(\boldsymbol{z}) = 0.8$
O
x
a_1
a_2
z
$h_A(\boldsymbol{a}) = 0.8$
$\boldsymbol{a}_C^*(\boldsymbol{x})$
$h(\boldsymbol{a}) = B$
$z_A = x'_A$
$\boldsymbol{a}_A^*(\boldsymbol{x})$
O
$x = x'_C$
$H(G)$
a_1

(a) CS Best Response (b) Discount Mechanism

Fig. 6. An illustration of the CS best response and the discount mechanism in regression, where the green dashed lines are equal decision outcome contours, $P = [1,1], w = 1, P^T w = (1,1), c_2 < c_1$, the red dashed line is the discounted equal cost contour with a minimum effective discount.

The difficulty of designing incentive mechanisms for strategic regression problems stems from the fact that the equilibrium individual errors depend on $\boldsymbol{x}$ and thus the overall prediction error depends largely on $p(\boldsymbol{x})$. Moreover, the individual equilibrium error is not monotone in any action dimension for a general $\boldsymbol{w} \neq \boldsymbol{\theta}$. As a result, we can not follow the same methods used in the strategic classification setting to find sufficient conditions that simplify the search for the optimal mechanism.

However, the mechanism designer can now leverage the fact that the agents have identical best responses to facilitate the search for IC and IR discount mechanisms that satisfy $S(f,G) \geq 0$, as shown in the following theorem.

THEOREM 4.2. *Suppose the computation of integration $\int_{\mathcal{X}} \mathcal{E}(f,\boldsymbol{a},\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}, \forall \boldsymbol{a}$ can be done in finite time. Then, Algorithm 3 runs in polynomial time and any $G \neq 0$ it returns is IC, IR and satisfies $S(f,G) > 0$.*

The finite computation time assumption is met, for example, when the distribution $\mathcal{X}$ is discrete or when $p(\boldsymbol{x})$ is uniform.

If f incentivizes non-zero responses, then Algorithm 3 sets $\triangle c_j$ at the minimum effective discount value, and sets no discount on other actions. Then, it chooses $\underline{c} = 0, \overline{c} = \frac{\alpha B \triangle c_j}{c_j - \triangle c_j}$ so that it incentivizes all agents to take an AS best response $\boldsymbol{a}_A^*(\boldsymbol{x}) = \alpha \frac{B}{c_j - \triangle c_j}\boldsymbol{e}_j + (1-\alpha)\frac{B}{c_{i_C}}\boldsymbol{e}_{i_C}$.[5] If f incentivizes 0 responses, then the decision maker can choose $\triangle c_j = c_j - (P^T\boldsymbol{w})_j$ and set $\overline{c} = \alpha B$ in Algorithm 3 so that $\boldsymbol{a}_A^*(\boldsymbol{x}) = \alpha \boldsymbol{e}_j$.

Below, we also discuss the cases when $\boldsymbol{w} = \boldsymbol{\theta}$, e.g., the decision maker's optimal LS strategy $f_L^*(\boldsymbol{z}) = \boldsymbol{\theta}^T\boldsymbol{z}$.[6]

LEMMA 4.3. *If $\boldsymbol{w} = \boldsymbol{\theta}$ in f and f incentivizes 0 responses or improvement, then the optimal IC and IR discount mechanism is $G = 0$.*

This is straightforward since the decision maker cannot further lower the error from $err(\eta)$ and thus does not want to pay the agents.

If f incentivizes gaming, then the equilibrium individual error becomes, $\mathcal{E}(f,\boldsymbol{a}_C,\boldsymbol{x}) = [\boldsymbol{\theta}^T(\boldsymbol{x} + P\boldsymbol{a}_C^*) - \boldsymbol{\theta}^T\boldsymbol{x}]^2 + err(\eta) = (\boldsymbol{\theta}^T P\boldsymbol{a}_C^*)^2 + err(\eta)$, which is independent of the pre-response attribute $\boldsymbol{x}$.

THEOREM 4.4. *If $\boldsymbol{w} = \boldsymbol{\theta}$ in f, and f incentivizes gaming, then the optimal IC, IR, and BB $G \neq 0$ can be found as follows:*

Choose $i_A = \arg\max_{j \leq M_+} (P^T\boldsymbol{\theta})_j / c_j$ as the target dimension, and set $\triangle c_{i_A} = \triangle c_{i_A}^*$.

[5] Similar to the classification setting, we let the algorithm put discount on one action dimension. Any $\underline{c} \leq \overline{c}$ is equivalent to both the agents and the designer here since the agent will by default use the discount amount $\overline{c}$ for the maximum improvement. The algorithm can return on condition $S > 0$ as well.

[6] The optimal CS strategy in regression does not guarantee $\boldsymbol{w} = \boldsymbol{\theta}$ when incentivizing improvement is impossible.

Then, derive the alternative form of individual subsidy surplus as $s(\alpha) = (2\alpha - \alpha^2)(\boldsymbol{\theta}^T P \boldsymbol{a}_C^*)^2 - \alpha B \triangle c_{i_A}(c_{i_A} - \triangle c_{i_A})^{-1}$ *and get* $\alpha^* = \arg\max_{\alpha \le 1} s(\alpha) = 1 - \frac{B \triangle c_{i_A}(c_{i_A} - \triangle c_{i_A})^{-1}}{2(\boldsymbol{\theta}^T P \boldsymbol{a}_C^*)^2}$. *Then find the optimal* $\bar{c}$ *by* $\bar{c} = \alpha^* B \triangle c_{i_A}(c_{i_A} - \triangle c_{i_A})^{-1}$.

An interesting observation is that the decision maker does not try to completely remove gaming with the discount mechanism. This is because when the error drops to a sufficiently low level, the marginal subsidy benefit becomes lower than the marginal subsidy cost, which is a constant.

ALGORITHM 3: Grid Search an IC, IR and $S(f, G) > 0$ Mechanism for Regression

```
Choose $\epsilon > 0$;
$\boldsymbol{a}_C \leftarrow \frac{B}{c_{i_C}} \boldsymbol{e}_{i_C}$; $S_{max} \leftarrow 0$;
 $ans \leftarrow (\boldsymbol{0}, [0, 0])$;
$E_C \leftarrow \int_X \mathcal{E}(f, \boldsymbol{a}_C, \boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x}$;
for $j = 1 : M_+$ do
    $\triangle \boldsymbol{c} \leftarrow \boldsymbol{0}$; $S \leftarrow 0$; $\alpha \leftarrow \epsilon$;
    $\triangle c_j \leftarrow c_j - \frac{(P^T \boldsymbol{w})_j}{(P^T \boldsymbol{w})_{i_C}} c_{i_C}$;
    while $S \ge 0$ do
        $\alpha \leftarrow \alpha + \epsilon$; $\bar{c} = \frac{\alpha B \triangle c_j}{c_j - \triangle c_j}$;
        $\boldsymbol{a}_A =$
          $\alpha \frac{B}{c_j - \triangle c_j} \boldsymbol{e}_j + (1 - \alpha) \frac{B}{c_{i_C}} \boldsymbol{e}_{i_C}$;
        $E_A \leftarrow \int_X \mathcal{E}(f, \boldsymbol{a}_A, \boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x}$;
        $S \leftarrow E_C - E_A - \bar{c}$;
    end
    if $S > S_{max}$ then
        $S_{max} \leftarrow S$;
         $ans \leftarrow (\triangle \boldsymbol{c}, [0, \bar{c}])$;
    end
end
Return $ans$.
```

5 DEMOGRAPHIC GROUPS AND SOCIAL WELL-BEING

Consider now the case where agents come from two demographic groups distinguished by a *sensitive attribute* $d \in \{1, 2\}$ (e.g., gender, race), which is not a part of the N skill-related attributes (not in $\boldsymbol{x}$) and is never influenced by an agent's action $\boldsymbol{a}$. Suppose the decision rule is *not allowed* to use the sensitive attribute as input but that it can be used to design *group specific* subsidies, so that different groups are subject to different incentive mechanisms provided the group identities are truthfully revealed.

We are particularly interested in how the subsidy mechanisms and their corresponding AS outcomes influence the fairness of the system. Below we introduce a number of commonly used definitions on group differences and social well-being measures related to fairness. Here, the term *well-being* is used to refer to a broader set of metrics defined below whereas *welfare* is used in the narrower sense of sum utility.

5.1 Group Differences

Without loss of generality, we will refer to group 1 as the *advantaged* group and 2 as the *disadvantaged* group.[7] We consider the following set of definitions; the first is new to the best of our knowledge and the other two were introduced in [14].

DEFINITION 2 (GROUP DISADVANTAGES). *We say group 2 is*

(1) *disadvantaged in attributes in classification if* $F^{(2)}(l) > F^{(1)}(l)$ *for* $l \in (0, 1)$, *where* $F^{(d)}$ *is the cumulative density function (cdf) of the conditional pre-response qualification status conditioned on* $d \in \{1, 2\}$; *the same in regression if* $F^{(2)}(y) > F^{(1)}(y)$ *for* $y \in (0, \max_{\boldsymbol{x}} q(\boldsymbol{x}))$.

(2) *disadvantaged in positive individuals (in classification) if* $F_+^{(2)}(l) > F_+^{(1)}(l)$, *where* $F_+^{(d)}$ *is the cdf of conditional pre-response qualification status* $(l(\boldsymbol{x}) | Y = 1, D = d), d \in \{1, 2\}$.

(3) *disadvantaged in action cost if* $h^{(2)}(\boldsymbol{a}) > h^{(1)}(\boldsymbol{a}), \forall \boldsymbol{a} \neq \boldsymbol{0}$, *where* $h^{(d)}$ *denotes the action cost functions with sensitive attribute* $d \in \{1, 2\}$. *Moreover, the minimum effective discount values satisfy* $(\triangle c^{(1)})_i^* \le (\triangle c^{(2)})_i^*, \forall i$.

[7] The group index shows up in superscripts.

5.2 Social Well-being Metrics

We will use the equilibrium qualification status $\mathbb{E}[y_t'], t \in \{C, A\}$ as an *efficiency* oriented social well-being metric. We also introduce *fairness* oriented well-being metrics.

DEFINITION 3 (QUALITY GAIN). *Quality gain measures the increase in agents' expected qualification status (positive rate in classification) in the response phase:*

$$\triangle Q_t := \mathbf{E}[Y_t'] - \mathbf{E}[Y_t];\ \triangle Q_t^d := \mathbf{E}[Y_t'|D = d] - \mathbf{E}[Y|D = d];\ \forall d \in \{1, 2\}, \forall t \in \{A, C\}. \tag{14}$$

$\gamma_t^Q(f, G) := \triangle Q_t^{(1)} - \triangle Q_t^{(2)}$ *further measures the group difference in this gain under game type t.*

Clearly, if f incentivizes improvement, then $\triangle Q_C > 0$; if $G \neq 0$ incentivizes improvement, then $\triangle Q_A > 0$. What's more interesting is to compare the quality gains across different groups and under different game types.

DEFINITION 4 (CLASSIFICATION FAIRNESS). *Considering two commonly used fairness criteria in classification, Equal Opportunity (*`EO`*) (equalized true positive rates) [8] and Demographic Parity (*`DP`*) (equalized positive decision rates), and define their respective group differences:*

$$\begin{aligned} \gamma_t^{EO}(f, G) &:= Pr(f(\boldsymbol{z}_t) = 1|Y_t' = 1, D = 1) - Pr(f(\boldsymbol{z}_t) = 1|Y_t' = 1, D = 2), t \in \{A, C\}; && (15) \\ \gamma_t^{DP}(f, G) &:= Pr(f(\boldsymbol{z}_t) = 1|D = 1) - Pr(f(\boldsymbol{z}_t) = 1|D = 2). && (16) \end{aligned}$$

5.3 Fairness Issues in the CS/LS Equilibrium

We start with a number of fairness limitations of the CS equilibria in classification and regression; the same results apply to LS.

THEOREM 5.1. *In the equilibrium CS outcome of classification where two groups have the same action cost, then (i) if group 2 is disadvantaged in attributes, then there is a* `DP` *gap no matter if f incentivizes improvement or gaming; and (ii) if group 2 is disadvantaged in positive individuals, then there is an* `EO` *gap if f incentivizes gaming but not necessarily if f incentivizes improvement.*

Part (1) is a direct result of $1 - F^{(1)}(l) > 1 - F^{(2)}(l)$, and the two groups have the same implicit threshold, which is the lower side boundary of their manipulation margins (since every agent above it will manipulate to get a positive decision outcome), and $\mathcal{M}^{(1)}(f) = \mathcal{M}^{(2)}(f)$ since the two groups have the same action cost. For part (2), whether there is a quality gain gap entirely depends on whether f incentivizes improvement and the distribution of each group in its manipulation margin $\mathcal{M}^{(d)}(f)$. For example, we can have $Pr(\boldsymbol{x} \in \mathcal{M}^{(2)}(f)|D = 2) > Pr(\boldsymbol{x} \in \mathcal{M}^{(1)}(f)|D = 1)$ and thus group 2 have more agents to improve and may have an inverse quality gain gap.

THEOREM 5.2. *In the equilibrium CS outcome of classification and regression, if group 2 is disadvantaged in action cost but has the same pre-response attribute distribution as group 1 (for positive individuals as well), then there is (i) a quality gain gap only if f incentivizes improvement; (ii) an* `EO` *gap no matter if f incentivizes improvement or gaming; and (ii) a* `DP` *gap no matter if f incentivizes improvement or gaming.*

To understand the above result, we note that if group 2 is disadvantaged in cost, we have $\mathcal{M}^{(1)}(f) \supseteq \mathcal{M}^{(2)}(f)$, so even when group 2 has the same pre-response attribute distribution, a larger portion of group 1 are accepted in the equilibrium, causing the `DP` gap. This is similar to the reason of an `EO` gap when f incentivizes gaming. If f incentivizes improvement, then a larger portion of group 1 will improve and be accepted in the equilibrium, causing a quality gain gap and an `EO` gap simultaneously.

5.4 Influence of the Discount Mechanism on Fairness

Here we analyze how the discount mechanism G alone may influence the fairness.

THEOREM 5.3. *If group 2 is disadvantaged in cost but has the same pre-response attribute distribution, then a rational decision maker will choose a G that widens the quality gain gap in both classification and regression.*

Theorem 5.3 means that a rational mechanism for the decision maker is always making the system more unfair when the quality gain gap is the metric. The rational mechanism influences the `DP` and `EO` gap but does not always make them worse.

6 THIRD PARTY MECHANISMS

We next discuss an alternative system where the discount mechanism is implemented by a third party, who subsidizes the agents' improvement actions in the same way as described in Section 2 and charges the decision maker a tax $\mathcal{T}(G)$ for improved decision performance. The decision maker's AS utility in this alternative system is

$$U_A^{(cls)}(f) = \int_X Pr\big(f(\boldsymbol{x} + P\boldsymbol{a}_A^*(\boldsymbol{x})) = y_A'\big)p(\boldsymbol{x})d\boldsymbol{x} - \mathcal{T}(G),$$

$$U_A^{(reg)}(f) = \int_X \mathbb{E}_\sigma\big[-\big(f(\boldsymbol{x} + P\boldsymbol{a}_A^*(\boldsymbol{x})) - y_A'\big)^2\big]p(\boldsymbol{x})d\boldsymbol{x} - \mathcal{T}(G)\ .$$

The IR condition for the decision maker is $U_A^{(cls)}(f) \geq U_C^{(cls)}(f)$ or $U_A^{(reg)}(f) \geq U_C^{(reg)}(f)$. In addition, we also consider the common mechanism criterion of budget balance: if the charged price is no less than the subsidy cost of the third party, then the mechanism is (weakly) budget balanced:

DEFINITION 5 (BUDGET BALANCE). *The third party is considered (weakly) budget balanced if $\mathcal{T}(G) \geq H(G)$.*

The mechanism designer can induce truthful revelation of the sensitive attribute by the agents as follows: (1) Let G consist of two group-specific mechanisms $G^{(1)}$ and $G^{(2)}$; agents who do not reveal their d participate in $G^{(1)}$; (2) Ensure that $\triangle c_i^{(1)} \leq \triangle c_i^{(2)}, \forall i$ and $(\triangle\boldsymbol{c}^{(1)})^T\boldsymbol{a} \in [\underline{c}^{(1)}, \overline{c}^{(1)}] \Rightarrow (\triangle\boldsymbol{c}^{(2)})^T\boldsymbol{a} \in [\underline{c}^{(2)}, \overline{c}^{(2)}]$. Then, group 1 agents are indifferent about revealing d while revealing d is the dominant strategy for group 2 agents. Figure 7 illustrates the three-party AS learning system.

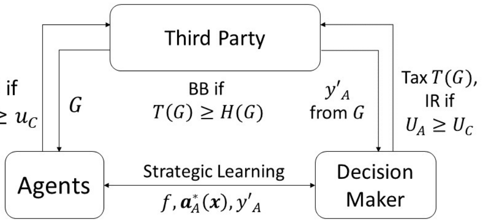


Fig. 7. An illustration of the alternative three-party augmented strategic learning system.

6.1 Objectives of the Third Party

We introduce two types of third party mechanism designers, *efficiency* oriented and *fairness* oriented. An *efficiency* oriented third party tries to maximize the equilibrium social qualification status $W_{eff}^{(cls)}(f, G) := Pr(y_A' = 1)$, $W_{eff}^{(reg)}(f, G) := \mathbb{E}[y_A']$; we call the corresponding equilibrium AS outcome the *efficient AS outcome* or *AS-eff* in short. On the other hand, a *fairness* oriented third party aims to minimize a non-negative and non-zero linear combination of the fairness gaps, or equivalently, maximizing

$$W_{fair}^{(cls)}(f, G) := -\beta^Q\gamma_A^Q(f, G) - \beta^{DP}\gamma_A^{DP}(f, G) - \beta^{EO}\gamma_A^{EO}(f, G);\ W_{fair}^{(reg)}(f, G) := -\gamma_A^Q(f, G),$$

for some $\beta^Q, \beta^{EO}, \beta^{DP} \geq 0, \beta^Q + \beta^{EO} + \beta^{DP} > 0$. We call the corresponding equilibrium AS outcome the *fair* AS outcome or *AS-fair* in short.

For conciseness, we use *AS-dm* to denote the decision maker's equilibrium AS outcome.

THEOREM 6.1. *If there is a mechanism that is IC and IR and satisfies $S(f, G) > 0$, then a mechanism that satisfies IC, IR, and BB criteria exists and weakly improves the third party's social well-being objective (either efficiency or fairness oriented) compared to the original AS equilibrium.*

6.2 Influence of Mechanism Designers' Objectives

Finally, we discuss how the objective of the mechanism designer and the corresponding incentive mechanisms influence the equilibrium efficiency and fairness oriented social well-being metrics. We compare the different AS, CS, and LS equilibrium outcomes where they have the same decision rule f and focus on how the incentive mechanisms for different objectives affect the outcome.

DEFINITION 6. *We say a mechanism $G^{(d)} \neq 0$ is an* ideal mechanism *if it is IC and IR for group d agents and achieves $S(f, G^{(d)}) > 0$ on group d, $\forall d \in \{1, 2\}$.*

THEOREM 6.2. *If group 2 is disadvantaged in action cost but has the same pre-response attribute distribution as group 1 (for positive individuals as well), then in the equilibrium,*

(1) *the* `DP` *gap in weak ascending order is: AS-fair, CS(LS), AS-dm, AS-eff;*
(2) *the* `EO` *gap (or quality gain gap) in weak ascending order is: AS-fair, CS(LS), AS-dm, AS-eff;*
(3) *The social quality improvement in weak descending order is: AS-eff, AS-dm, CS(LS).*

If there is an ideal mechanism for group 1, then AS-fair is strictly the lowest in `DP` *gap; the orders in* `EO` *gap (or quality gain gap) and quality improvement becomes strict for CS(LS), AS-dm and AS-eff. Moreover, if there is an ideal mechanism for group 2, AS-fair is strictly the lowest in* `EO` *gap (or quality gain gap).*

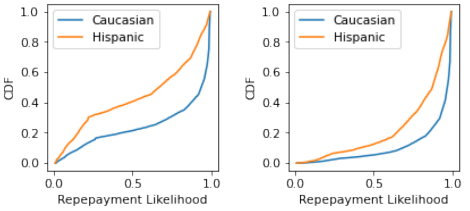
Below we provide some explanations of the statements in Theorem 6.2. For an efficiency oriented third party, the set of agents it incentivizes is a superset of the agents incentivized by the decision maker, making AS-eff the best in part (3). This is because subsidizing the agents with a positive individual subsidy surplus not only helps the third party improve the objective but also raises the budget to subsidize agents with a negative individual subsidy surplus (individual subsidy deficit). Moreover, the efficiency oriented third party tries to incentivize more agents from group 1 since they are "cheaper" to incentivize and thus exacerbates the fairness issues in parts (1) and (2).

For a fairness oriented third party, it can also incentivize a superset of agents incentivized by the decision maker, but that means incentivizing some group 1 agents, which results in two conflicting effects: it helps the third party gather more "funding" to subsidize group 2 agents, but at the same time makes the fairness issue worse. As a result, the social quality improvement in AS-fair is better than CS (LS) and worse than AS-eff, but how it compares to AS-dm depends on the specific game parameters and thus is not discussed in part (3). When there is an ideal mechanism for group 2, the third party can ignore the dilemma of subsidizing group 1 agents and focus on subsidizing only group 2 agents to improve fairness in parts (1) and (2).

The ideal mechanisms in Theorem 6.2 makes the comparison strict. The existence of an ideal $G^{(2)}$ is a sufficient condition to the existence of an ideal $G^{(1)}$ when group 2 is disadvantaged in cost but has the same distribution. This is because $G^{(2)}$ itself is ideal for group 1.

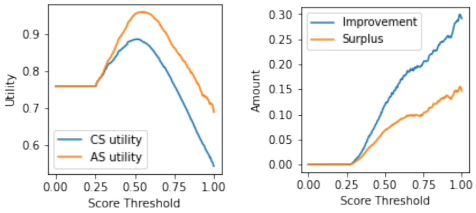
THEOREM 6.3. *In both classification and regression problems, if group 2 is disadvantaged in attributes (resp. positive individuals) but has the same action cost as group 1 then*

(1) *the* `DP` *(resp.* `EO`*) gap in AS-fair outcome is weakly the lowest, and is strictly the lowest if there is an ideal mechanism for group 2;*



(a) Entire Group (b) Positive Individuals

Fig. 9. The Likelihood CDF



(a) CS/AS Utilities (b) $\triangle Q$ and $S(f, G)$

Fig. 10. Single Group (Caucasian) Results

(2) *the social quality improvement in AS-eff outcome is weakly the highest, and is strictly the highest if there is an ideal mechanism for either group.*

When group 2 has the same cost, then an ideal $G^{(2)}$ is no longer sufficient or necessary for an ideal $G^{(1)}$ to exist for general classification problems, and that's why the condition in part (2) looks different from Theorem 6.3. But the existence of an ideal $G^{(2)}$ is sufficient and necessary for the existence of an ideal $G^{(1)}$ in regression, as well as in a special class of classification problems where $\boldsymbol{w} = \boldsymbol{\theta}$ in f and l is convex on $[0, \tau]$.[8] From Theorem 5.1, we know that DP and EO gap always exist in the CS (LS) problem, but if there is an ideal $G^{(2)}$, the fairness oriented third party can further incentivize group 2 agents to reduce the gap in part (1) (those not in $\mathcal{M}^{(2)}(f)$ to reduce the DP gap).

THEOREM 6.4. *Suppose group 2 is disadvantaged in cost but has the same pre-response distribution (for positive individuals as well). Denote $p^{(d)} := Pr(D = d)$, then an IC, IR, and BB mechanism $G \neq 0$ that satisfies $\gamma_A^Q = \gamma_A^{EO} = \gamma_A^{DP} = 0$ exists if $S(f, G^{(1)}) + (1 - p^{(1)})H(G^{(1)}) \geq p^{(2)}H(G^{(2)})$, s.t. $h_A^{(1)}(\boldsymbol{a}) = h_A^{(2)}(\boldsymbol{a}), \forall \boldsymbol{a}$.*

In general, this condition can hold if $p^{(1)}$ is much larger than $p^{(2)}$, i.e., the disadvantaged group is also the minority group in the population or $S(f, G^{(1)})$ is very high.

REMARK 3. *Our results generalize to multiple groups when the definitions of group disadvantages and fairness metrics are consistent.*

7 NUMERICAL RESULTS

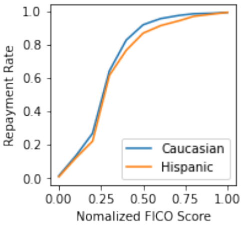


Fig. 8. Repay Rate $l(x)$

This section presents numerical results obtained using the FICO score [15] dataset preprocessed in [8]. The credit card holders are considered as agents and they have repayment rates that can map to the likelihood function l in our model. The decision maker uses binary classification to predict whether the agents will default. We assume that $\theta = 1$, $P = [1, 1]$, and the agent can either choose a_1 to improve or a_2 to game the classifier $f(z) = \mathbf{1}(z \geq \tau)$, i.e., x is the pre-response normalized FICO score as well as the attribute, $x' = x + a_1$ is the post-response attribute, and $z = x + a_1 + a_2$ is the post-response normalized FICO score. Figure 8 shows how the repayment rate $l(x)$ changes with x; it has an S-shape, with $l(x) = 0.5$ approximately corresponding to $x = 0.3$ and $l(x)$ (nearly) convex on $[0, 0.3]$. We assume that the decision maker chooses $w = 1$, which aligns with the LS and CS optimal solution from Section 3 when $c_2 < c_1$.

[8] We are excluding extreme distributions in the "iff" claim, e.g., $Pr(\boldsymbol{x} \in \mathcal{M}(f)) = 0$.

We start with the properties of the discount mechanism and show how the decision maker's CS and AS utility changes with different choices of threshold τ. We then show the impact the incentive mechanisms have on social well-being metrics.

Throughout this section, we use a quadratic outcome likelihood cost function and assume that $c_1^{(1)} = c_1 = 8$ and $c_2^{(1)} = c_2 = 4$ (for the advantaged group if there are action cost differences). For the multiple group case, we make the following two sets of comparisons. (1) Groups with different distributions: the Hispanic group is disadvantaged in features and in positive individuals compared to the Caucasian group (see Figure 9). (2) Groups with different costs: we will assume there are two subgroups (A and B) in the Caucasian group, and group 2 has higher action costs $c_1^{(2)} = 10$ and $c_2^{(2)} = 5$. We set $p^{(1)} = 0.8, p^{(2)} = 0.2$ as the population proportions.

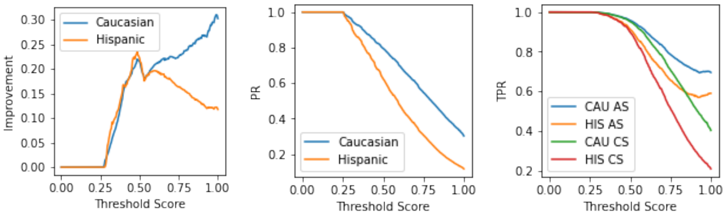
As a result, we show the AS-fair equilibrium outcome is the best well-rounded system design for the augmented strategic learning problems.

The decision maker's AS and CS utility. Using only the Caucasian data, the set of results in Figure 10 show how the AS/CS decision maker utilities, subsidy surplus and qualification status improvement change with the threshold τ.

We can see that the AS utility is always higher than the CS utility (Fig. 10). This is because their difference is the subsidy surplus, which is non-negative for a rational decision maker. We note that the CS utility should always be single-peaked but the AS utility may have multiple local maxima since the value of subsidy surplus is not monotone in τ and depends on $p(x)$. For other choices of c_1, c_2 values, we find that the larger the difference $c_1 - c_2$, the smaller the utility difference and the closer the optimal thresholds are ($|\tau_{AS}^* - \tau_{CS}^*|$ lower). Both the subsidy surplus in (12) and the qualification status improvement in (14) are positive, indicating the decision maker's selfish strategy is also benefiting the efficiency oriented social well-being. The improvement and subsidy surplus are also highly positively correlated with a correlation coefficient of 0.92.

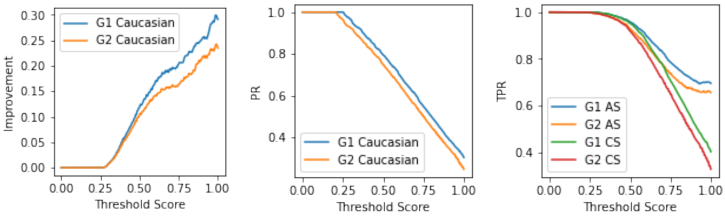
Social well-being of the strategic incentive mechanism. Figure 11 (resp. Figure 12) shows the quality improvement, PR and TPR, (and thus we can see the DP, and EO gap from the curve differences) when the Hispanic group (resp. Caucasian subgroup 2) is disadvantaged in features and positive individuals (resp. costs) compared to the Caucasian group (resp. Caucasian subgroup 1) in the CS(LS) and AS-dm equilibrium. The decision maker does not incentivize agents outside of the manipulation margin and thus the CS and AS PR curves are the same.

We can see from Figure 11a that when τ is in the lower score ranges, the Hispanic group has a slightly higher qualification status improvement compared to the Caucasian group, whereas if τ is in the higher score ranges, the Caucasian group has a much higher improvement. Intuitively, this is because the Hispanic (resp. Caucasian) group has a higher probability mass in the lower (resp. higher) score ranges and a low (resp. high) τ incentivizes a higher proportion of agents to improve in the Hispanic (resp. Caucasian) group. Figure 11b shows that the PR is 1 when $\tau < 0.25$; this is because all agents can manipulate to get $f(z) = 1$. When $\tau > 0.25$, the PR is strictly decreasing in τ for both groups and the Caucasian group always has a higher PR, i.e., the Hispanic group will suffer from a DP gap in both CS and AS-dm equilibrium. This is because the lower side boundary of the manipulation margin becomes an implicit threshold, where all agents above the implicit threshold can manipulate (no matter improvement or gaming) to get accepted. The implicit threshold is the same for both groups since they have the same action cost, and the DP gap is caused by the disadvantage in pre-response attribute distribution (Theorem 5.1 part (1)). For similar reasons, Figure 11c shows that the CS and AS TPR is 1 when $\tau < 0.3$. Therefore, we can see that the AS TPR is always higher than the CS TPR for either group, because now some agents improved their qualification status and get accepted at the same time, making the numerator and denominator of the TPR formula increase by the same amount and thus increase the TPR. On the other hand, the Hispanic group suffers from an EO gap in both the CS and the AS-dm equilibrium, as previously discussed in Theorem 5.1 part (2).



(a) Improvement (b) PR (c) AS/CS TPR

Fig. 11. Disadvantaged in features



(a) Improvement (b) PR (c) TPR

Fig. 12. Disadvantaged in costs

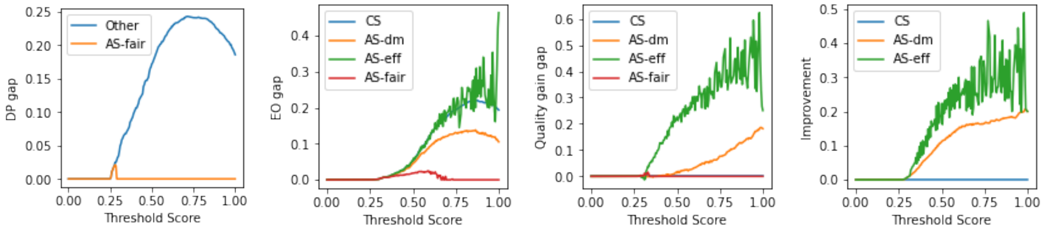
Figure 12a and 12c support our claims in Theorem 5.3 part (3), where the incentive mechanism widens the quality gain gap and the EO gap. Figure 12b shows PR curves and the DP gap between the two subgroups, which is determined by the pre-response attribute probability mass within $[\tau - 1/c_2^{(1)}, \tau - 1/c_2^{(2)}]$ (the difference between the manipulation margins in the two groups). Figure 12c shows the CS and AS TPR curves and the EO gaps; the implicit threshold creates the CS EO gap, and the fact that group 1 agents are cheaper to incentivize jointly creates the AS EO gap.

Social Well-being metrics with the third party incentive. Social well-being results under the third party model are shown in Figure 13 where groups have attribute distribution differences (Caucasian and Hispanic group), and in Figure 14 where groups have cost differences (Caucasian subgroups).

We can see in both sets of results that the AS-fair equilibrium outcome significantly reduces and even removes the fairness issues in the system, whereas the AS-eff equilibrium outcome has the worst fairness metrics. On the other hand, the AS-eff equilibrium achieves the highest social qualification status improvement. We note that the chosen AS-fair outcomes used mechanisms that incentivized a superset of agents compared to those that are incentivized by the decision maker, and thus it achieves a higher social qualification status improvement than AS-dm as well.

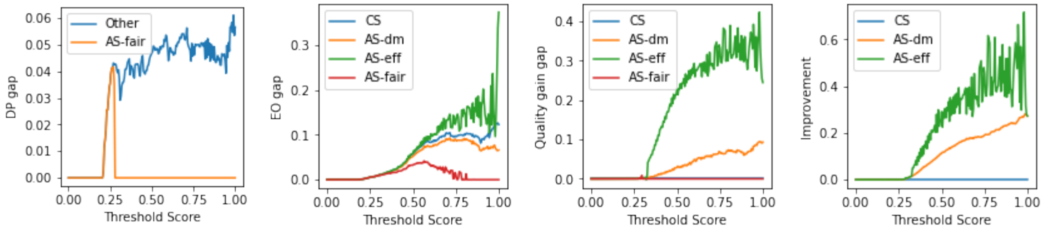
8 CONCLUSION

We formulated Stackelberg game models to study the strategic classification and regression problem, where the decision maker's strategy combines a decision rule and an incentive mechanism. Our model provides an extension of the previously studied strategic learning problems. We showed how the decision maker can design discount-based incentives mechanisms to use in conjunction with its decision rule, by providing conditions on when this problem is computationally intractable, discussing when and how approximate algorithms can find reasonable mechanisms in polynomial time, and when the optimal mechanism can be found in closed-form. We then discussed the efficiency and fairness oriented social well-being properties of the augmented strategic learning system when multiple demographic groups co-exist. We also examined an alternative model where the incentive mechanism is provided by a third party, whose objective is optimizing some of the social well-being



(a) DP gap (b) EO gap (c) Quality gain gap (d) Improvement

Fig. 13. Third Party Outcomes with Attribute Distribution Differences



(a) DP gap (b) EO gap (c) Quality gain gap (d) Improvement

Fig. 14. Third Party Outcomes with Cost Differences

metrics with an IC, IR, and BB mechanism, and showed how an efficiency-oriented and fairness-oriented third party can influence the equilibrium social well-being metrics. We conducted numerical experiments on the FICO dataset to demonstrate the impact of the incentive mechanism on the system. Our findings established that a fairness-oriented third party can provide the best well-rounded equilibrium outcomes compared to a selfish decision maker, an efficiency-oriented third party, or a system without an incentive mechanism.

ACKNOWLEDGMENTS

This material is based upon work supported by the NSF program on Fairness in AI in collaboration with Amazon under Award No. IIS-2040800. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of Amazon or the National Science Foundation.

REFERENCES

[1] Mark Braverman and Sumegha Garg. 2020. The Role of Randomness and Noise in Strategic Classification. In *1st Symposium on Foundations of Responsible Computing*.

[2] Michael Brückner, Christian Kanzow, and Tobias Scheffer. 2012. Static Prediction Games for Adversarial Learning Problems. *The Journal of Machine Learning Research* 13 (09 2012), 2617–2654.

[3] Michael Brückner and Tobias Scheffer. 2011. Stackelberg games for adversarial prediction problems. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 547–555. https://doi.org/10.1145/2020408.2020495

[4] Yatong Chen, Jialu Wang, and Yang Liu. 2020. Strategic Recourse in Linear Classification. *arXiv preprint arXiv:2011.00355* (2020).

[5] Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. 2018. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*. 55–70.

[6] Nika Haghtalab, Nicole Immorlica, Brendan Lucier, and Jack Wang. 2020. Maximizing Welfare with Incentive-Aware Evaluation Mechanisms. 160–166. https://doi.org/10.24963/ijcai.2020/23

[7] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. 2016. Strategic Classification. 111–122. https://doi.org/10.1145/2840728.2840730

[8] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*. 3315–3323.

[9] Keegan Harris, Hoda Heidari, and Zhiwei Steven Wu. 2021. Stateful Strategic Regression. arXiv:cs.LG/2106.03827

[10] Lily Hu, Nicole Immorlica, and Jennifer Vaughan. 2019. The Disparate Effects of Strategic Manipulation. 259–268. https://doi.org/10.1145/3287560.3287597

[11] Kun Jin, Tongxin Yin, Charles A. Kamhoua, and Mingyan Liu. 2021. Network Games with Strategic Machine Learning. In *Decision and Game Theory for Security*, Branislav Bošanský, Cleotilde Gonzalez, Stefan Rass, and Arunesh Sinha (Eds.). Springer International Publishing, Cham, 118–137.

[12] Jon Kleinberg and Manish Raghavan. 2020. How Do Classifiers Induce Agents to Invest Effort Strategically? *ACM Transactions on Economics and Computation* 8 (11 2020), 1–23. https://doi.org/10.1145/3417742

[13] John Miller, Smitha Milli, and Moritz Hardt. 2020. Strategic Classification is Causal Modeling in Disguise. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Hal Daumé III and Aarti Singh (Eds.), Vol. 119. PMLR, 6917–6926.

[14] Smitha Milli, John Miller, Anca Dragan, and Moritz Hardt. 2019. The Social Cost of Strategic Classification. 230–239. https://doi.org/10.1145/3287560.3287576

[15] US Federal Reserve. 2007. Report to the congress on credit scoring and its effects on the availability and affordability of credit. *Board of Governors of the Federal Reserve System* (2007).

[16] Yonadav Shavit, Benjamin Edelman, and Brian Axelrod. 2020. Causal Strategic Linear Regression. arXiv:cs.LG/2002.10066

[17] Marilyn Strathern. 1997. ‘Improving ratings’: audit in the British University system. *European review* 5, 3 (1997), 305–321.

A SUPPLEMENTARY MATERIAL FOR SECTION 2

A.1 Discussion on Remark 1

In this part, we discuss the case where game parameters like $\boldsymbol{\theta}$ and P are unknown and the decision maker need to be learn them. Unlike the single round, two stage game in the main article, the learning process requires online learning with multiple rounds, each containing two stages.

We note that the quality coefficients $\boldsymbol{\theta}$ can be learned in one round by setting $f = 0$ and then we have $(\boldsymbol{z}, y') = (\boldsymbol{x}, y)$, and running any suitable learning algorithm can get an estimate of $\boldsymbol{\theta}$.

However, P can not always be learned in the conventional learning problem. We can use an example can from the impossibility conditions in Theorem 3.3, given those conditions, only the columns whose index has substitutability 1 can be learned, the other columns are always unknown. Below we show how the discount mechanism help with learning the the projection matrix P.

In the regression problem with L1 cost, we can use the following procedures to learn the projection matrices,

- Choose f such that $\boldsymbol{w} > 0$ (without loss of generality, assume that $\boldsymbol{w} > 0 \Rightarrow P^T\boldsymbol{w} > 0$, otherwise some action dimensions are meaningless)
- For each time step $t = 1, \ldots, M$, get a sufficiently large sample of agents with their observable features $\boldsymbol{z}$
- At $t = 0$, $G_d = 0$, let $\bar{\boldsymbol{z}}_0 = \mathbb{E}[\boldsymbol{z}]$
- At $t = 1, \ldots, M$, let G_d induce the best response along action dimension t by lowering the cost to $\tilde{c}_t$, and let $\bar{\boldsymbol{z}}_t = \mathbb{E}[\boldsymbol{z}]$
- Compute $\boldsymbol{v}_t = (\bar{\boldsymbol{z}}_t - \bar{\boldsymbol{z}}_0)\tilde{c}_t/B$, which is an estimate of $P\boldsymbol{e}_t = \boldsymbol{p}_t$, i.e., the t-th column of P.

Discount mechanisms can enable best responses to in action dimensions that are impossible to be incentivized with the decision rule itself, and this is true for both classification and regression, both L1 cost and other types of costs like L2 or squared.

A.2 Discussion on Remark 2

We will use the L2 cost $h(\boldsymbol{a}) = ||\boldsymbol{a}||_2$ for demonstration purpose, and we note that higher orders of cost functions $h(\boldsymbol{a}) = \frac{1}{2}||\boldsymbol{a}||_2^2$ are very similar in classification but different in regression. In regression, higher order costs are convex and the marginal cost grows, and thus there is no need to be a budget constraint $B \geq h(\boldsymbol{a})$, other than that, $h(\boldsymbol{a}) = ||\boldsymbol{a}||_2$ is very representative.

For all other cost functions, we can equivalently have a set of "equal cost contour" i.e., $\{\boldsymbol{a}|h(\boldsymbol{a}) = C\}$ for some constant C is a contour. Most cost functions used in economic and computer science literature have contours with different sizes but a constant "shape" (the surface of norm balls, since the cost functions are norm based), like the L1 cost, L2 cost, tilted L2 cost $h(\boldsymbol{a}) = \sqrt{\boldsymbol{a}^T C\boldsymbol{a}}$ and squared cost $h(\boldsymbol{a}) = \frac{1}{2}||\boldsymbol{a}||_2^2$. The constant shape of contours made it possible to have a concise (closed-form in most cases) representation of the best responses' directional and magnitude properties.

For example, when $h(\boldsymbol{a}) = ||\boldsymbol{a}||_2$, the best responses satisfy $\rho(\boldsymbol{a}_t^*, P^T\boldsymbol{w}) = 1$ where $\rho(\boldsymbol{v}_1, \boldsymbol{v}_2) = \frac{\boldsymbol{v}_1^T\boldsymbol{v}_2}{||\boldsymbol{v}_1||_2||\boldsymbol{v}_2||_2}$ is the cosine similarity. We still have properties in Lemma 3.1 and in classification and regression, the best

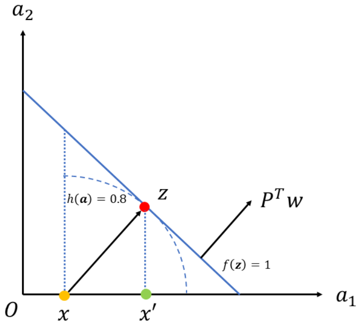


Fig. 15. An illustration of a CS best response in classification with L2 cost, where the blue dashed curve (quarter circle) is an equal cost contour, $P = [1, 1]$, $\boldsymbol{w} = (1, 1)$, a_1 is improvement and a_2 is gaming.

responses are

$$\boldsymbol{a}_C^*(\boldsymbol{x}) = \frac{\tau - \boldsymbol{w}^T\boldsymbol{x}}{||P^T\boldsymbol{w}||_2^2} P^T\boldsymbol{w},\ \boldsymbol{a}_C^*(\boldsymbol{x}) = \frac{B}{||P^T\boldsymbol{w}||_2} P^T\boldsymbol{w},$$

and we can similarly write out the expressions of the AS best responses for other cost functions.

For L2 cost $h(\boldsymbol{a}) = ||\boldsymbol{a}||_2$, we can think of discounts with minimum effective discount value as giving certain action directions a fixed discount rate or incentivizing agents to play a different action and pay the cost differences.

Therefore, the implementer will try to incentivize some of the agents to take an AS best response that also reach the boundary, this can be done by making the discount amount equal the cost difference between the AS and the CS/LS best response. The implementer wants to maximize the subsidy surplus on a given agent, which is the quality gain $l(\boldsymbol{a}) - l(\boldsymbol{a}_C^*(\boldsymbol{x}))$ minus the subsidy cost $||\boldsymbol{a}||_2 - ||\boldsymbol{a}_C^*(\boldsymbol{x})||_2$ and thus $\boldsymbol{a}_A^*(\boldsymbol{x})$ is the solution to the optimization problem

$$\begin{aligned} \text{minimize}_{\boldsymbol{a}} \ \ & ||\boldsymbol{a}||_2 - l(\boldsymbol{\theta}^T(\boldsymbol{x} + \hat{P}\boldsymbol{a})) \\ \text{subject to} \ \ & \boldsymbol{w}^T P\boldsymbol{a} = \tau - \boldsymbol{w}^T\boldsymbol{x} \end{aligned} \tag{17}$$

However, the above problem is in general not convex and can be NP hard to find the optimal solution. But the below assumption guarantees a solution.

ASSUMPTION 1. $\boldsymbol{w} = \boldsymbol{\theta}$*, and the implementer limit the AS best response to be* gaming free*, i.e.,* $[\boldsymbol{a}_A^*(\boldsymbol{x})]_j = \mathbf{1}\{j \leq M_i\} \Leftrightarrow P\boldsymbol{a}_A^*(\boldsymbol{x}) = \hat{P}\boldsymbol{a}_A^*(\boldsymbol{x})$*,*

Under Assumption 1, the problem becomes convex since $l(\boldsymbol{\theta}^T(\boldsymbol{x} + \hat{P}\boldsymbol{a})) = l(\tau)$ is constant

$$\begin{aligned} \text{minimize}_{\boldsymbol{a}} \ \ & ||\boldsymbol{a}||_2 \\ \text{subject to} \ \ & \boldsymbol{\theta}^T\hat{P}\boldsymbol{a} = \tau - \boldsymbol{\theta}^T\boldsymbol{x} \end{aligned} \tag{18}$$

and the solution (AS best response to incentivize) is

$$\boldsymbol{a}_A^*(\boldsymbol{x}) = (\tau - \boldsymbol{\theta}^T\boldsymbol{x}) \frac{\hat{P}^T\boldsymbol{\theta}}{||\hat{P}^T\boldsymbol{\theta}||_2^2} \tag{19}$$

We can then similarly define the individual subsidy surplus in the L2 case and find sufficient conditions that guarantees an IC, IR and BB mechanism $G \neq 0$ or even find the optimal solutions with the same assumptions made in the Theorems 3.6 and 3.7.

One interesting difference in the L2 cost case is that the decision rule can incentivize *partial improvement*, which can also be called *partial gaming*, which means $\boldsymbol{\theta}^T P\boldsymbol{a} > \boldsymbol{\theta}^T\hat{P}\boldsymbol{a} > 0$, and the corresponding theorems in L1 case still applies when f incentivizes pure gaming $\boldsymbol{\theta}^T P\boldsymbol{a} > \boldsymbol{\theta}^T\hat{P}\boldsymbol{a} = 0$. An example of pure gaming happens when for every improvement action j, there is a corresponding gaming action k with the an exaggerated effect $\boldsymbol{p}_k = \alpha_j\boldsymbol{p}_j, \alpha_j > 1$, which can model problems like multi-subject exams where an agent has an improvement and gaming action for each of the subject and cheating is always more cost efficient than working hard without an incentivize mechanism.

A.3 An Alternative Incentive Mechanism

An alternative mechanism to consider, the *transfer mechanisms* is based on *monetary transfer*, where the mechanism designer provides reimbursement or bonus payment when the agent meets certain feature criteria, e.g., rewards for high scores. We use G_t to denote the transfer mechanism, where the designer chooses a bonus amount $b(\boldsymbol{z}), b : \mathbb{R}^N \mapsto \mathbb{R}$, effectively revising the agent's utility to

$$u_A(\boldsymbol{x}, \boldsymbol{a}) = f(\boldsymbol{x} + P\boldsymbol{a}) - h(\boldsymbol{a}) + b(\boldsymbol{x} + P\boldsymbol{a}). \tag{20}$$

In transfer mechanisms, knowing the actual $\boldsymbol{x}$ seems to help the designer reduce the subsidy cost on agents with high endowment and low improvement, but we will show below that this extended version with bonus amount $\tilde{b}(\tilde{\boldsymbol{x}}, \boldsymbol{z})$ is equivalent as the bonus $b(\boldsymbol{z})$ that only uses features as input, where $\tilde{x}$ is the reported pre-response attribute. This is because $\tilde{b}(\tilde{\boldsymbol{x}}, \boldsymbol{z})$ either can not incentivize agents to truthfully report $\tilde{\boldsymbol{x}} = \boldsymbol{x}$, or it can not generate more benefit for the mechanism designer.

With the alternative version of the monetary transfer mechanism, the agent's utility now becomes

$$\tilde{u}_A(\boldsymbol{x}, \boldsymbol{a}, G_t) = f(\boldsymbol{x} + P\boldsymbol{a}) - h(\boldsymbol{a}) + \max_{\tilde{\boldsymbol{x}}} \tilde{b}(\tilde{\boldsymbol{x}}, \boldsymbol{x} + P\boldsymbol{a}),$$

and we can find the corresponding $\boldsymbol{a}_A^*(\boldsymbol{x})$, and only if

$$\boldsymbol{x} \in \arg\max_{\tilde{\boldsymbol{x}}} \tilde{b}(\tilde{\boldsymbol{x}}, \boldsymbol{x} + P\boldsymbol{a}_A^*(\boldsymbol{x})),$$

truth-reporting is incentivized. If truth reporting is not incentivized, $\tilde{b}(\tilde{\boldsymbol{x}}, \boldsymbol{z})$ and $b(\boldsymbol{z}) = \max_{\tilde{\boldsymbol{x}}} \tilde{b}(\tilde{\boldsymbol{x}}, \boldsymbol{z})$ are equivalent for both the agents and the mechanism designer. Meanwhile, for $\forall \boldsymbol{x}_1 \neq \boldsymbol{x}_2$, truth telling requires either

$$\boldsymbol{x}_1 + P\boldsymbol{a}_A^*(\boldsymbol{x}_1) \neq \boldsymbol{x}_2 + P\boldsymbol{a}_A^*(\boldsymbol{x}_2),$$

indicating that backward induction from $\boldsymbol{x} + P\boldsymbol{a}_A^*(\boldsymbol{x})$ to $\boldsymbol{x}$ is achievable, or

$$\boldsymbol{x}_1 + P\boldsymbol{a}_A^*(\boldsymbol{x}_1) = \boldsymbol{x}_2 + P\boldsymbol{a}_A^*(\boldsymbol{x}_2), \text{ and } \boldsymbol{x}_1, \boldsymbol{x}_2 \in \arg\max_{\tilde{\boldsymbol{x}}} \tilde{b}(\tilde{\boldsymbol{x}}, \boldsymbol{x} + P\boldsymbol{a}_A^*(\boldsymbol{x})).$$

In either case, $b(\boldsymbol{z})$ is sufficient.

However, the computational complexity is very high in the backward induction step for a general $b(\boldsymbol{z})$ bonus function. Recall that the AS utility of an agent is

$$u_A(\boldsymbol{x}, \boldsymbol{a}) = f(\boldsymbol{x} + P\boldsymbol{a}) - h(\boldsymbol{a}) + b(\boldsymbol{x} + P\boldsymbol{a}),$$

and thus computing $\boldsymbol{a}_A^*(\boldsymbol{x}) = \arg\max_{\boldsymbol{a}} u_A(\boldsymbol{x}, \boldsymbol{a})$ is non-convex for a non-concave $b(r)$ bonus function.

On one hand, we can't guarantee concave $b(r)$ is the optimal solution. On the other hand, for a concave $b(\boldsymbol{z})$, the computation of $\boldsymbol{a}_A^*(\boldsymbol{x}) = \arg\max_{\boldsymbol{a}} u_A(\boldsymbol{x}, \boldsymbol{a})$ is convex and but the individual subsidy surplus

$$s(\boldsymbol{x}, f, G_t) = l(\boldsymbol{\theta}^T(\boldsymbol{x} + \hat{P}\boldsymbol{a}_A^*(\boldsymbol{x}))) - l(\boldsymbol{\theta}^T(\boldsymbol{x} + \hat{P}\boldsymbol{a}_C^*(\boldsymbol{x}))) - b(\boldsymbol{x} + P\boldsymbol{a}_A^*(\boldsymbol{x}))$$

on the agents are not concave unless l is convex (we are supposing $\boldsymbol{x} \in \mathcal{M}(f)$ here, otherwise more non-convexity is introduced). Moreover, the overall objective depends on the integration on a subset of $\hat{\mathcal{X}} \subseteq \mathcal{X}$

$$S(f, G_t) = \int_{\hat{\mathcal{X}}} s(\boldsymbol{x}, f, G_t) p(\boldsymbol{x}) d\boldsymbol{x},$$

and a general probability density function p, and the convexity of set $\hat{\mathcal{X}}$ can make the mechanism designer's objective non-convex even if l is convex.

We also note that when changing the value $b(\boldsymbol{z})$ for a certain $\boldsymbol{z}$, the AS best response for all agents with pre-response attribute $\boldsymbol{x}$ in the cone $\boldsymbol{x} - \boldsymbol{z} \leq 0$ (element wise non-positive) might change, and this also makes the analysis hard.

B SUPPLEMENTARY MATERIAL FOR SECTION 3

B.1 Characterization of the optimal LS decision rule

LEMMA B.1. *The LS optimal decision rule is* $f_L^*(\boldsymbol{z}) = \mathbf{1}\{\boldsymbol{\theta}^T\boldsymbol{z} \geq \tau_L\}$, $\tau_L = \arg\min_\tau l(\tau) \geq 0.5$.

PROOF. This is because it is optimal for the decision maker to accept every agent with $l(\boldsymbol{\theta}^T\boldsymbol{x}) \geq 0.5$, since rejecting this agent results in a decrease in the expected individual prediction outcome $1 - l(\boldsymbol{\theta}^T\boldsymbol{x}) \leq l(\boldsymbol{\theta}^T\boldsymbol{x})$. Similarly, the decision maker wants to reject every agent with $l(\boldsymbol{\theta}^T\boldsymbol{x}) < 0.5$. □

B.2 Proof of Theorem 3.3

PROOF. The proofs of the claims

(1) If $\kappa_j = 1$, then there exists a $\boldsymbol{w}$ in f that can incentivize action dimension j, and the $\boldsymbol{w}$ can be found in polynomial time;

(2) if $\kappa_j < 1$, meaning there always are linear combinations of gaming actions weakly dominate every action j, then there is no f that can incentivize best response on action j.

are covered in [11, 12]. Intuitively, if $\kappa_j < 1, \forall j \leq M_+$, the corresponding $\boldsymbol{a}$ is the combination that strictly dominates $\boldsymbol{e}_j$ for any f and thus there is no f that can incentivize improvement.

We will proceed to show the decision maker's CS optimal strategy satisfy $\boldsymbol{w} = \boldsymbol{\theta}$. The main idea is that when f always incentivizes gaming, then the CS decision outcomes with $f_C(\boldsymbol{z}) = \mathbf{1}\{\boldsymbol{w}_C^T \boldsymbol{z} \geq \tau_C\}$ always have an equivalent LS decision outcomes with $f_L(\boldsymbol{z}) = \mathbf{1}\{\boldsymbol{w}_L^T \boldsymbol{z} \geq \tau_L\}$, where the $\boldsymbol{w}_C = \boldsymbol{w}_L$, and τ_C, τ_L satisfy

$$\tau_L = \min \left\{ 0, \tau_C - \frac{(P^T \boldsymbol{w})_k}{c_k} \right\}.$$

In other words, we can show that $\forall \boldsymbol{x}, f_L(\boldsymbol{x}) = f_C(\boldsymbol{x} + P\boldsymbol{a}^*)$, and thus is equivalent for the decision maker to find an optimal f_L which guarantees $\boldsymbol{w}_L = \boldsymbol{\theta}$ as the Lemma B.1 suggests. □

B.3 Proof of Theorem 3.5

PROOF. We will first show the problem is non-convex when discount is placed on multiple actions, then show even the discount is only on one action, the problem is still non-convex.

When the discount is on multiple actions, providing the optimal tie breaking strategy for an agent with $\boldsymbol{x}$ requires solving

$$\text{maximize}_{\boldsymbol{a}} \; l(\boldsymbol{\theta}^T (\boldsymbol{x} + \hat{P}\boldsymbol{a})) - \triangle \boldsymbol{c}^T \boldsymbol{a},$$

which is non-convex for a general l function. This is for individual subsidy surplus for a fixed $\triangle \boldsymbol{c}$, and it has to be integrated over $\mathcal{X}$ to compute the overall subsidy surplus $S(f, G)$. So finding the optimal mechanism will only have higher computational complexity when the decision maker has to optimize over $\triangle \boldsymbol{c}, \underline{c}, \overline{c}$, and take into account the influence of $p(\boldsymbol{x})$.

When the discount is only on one action, from Lemma 3.4, the mechanism designer need to choose $\triangle \boldsymbol{c}$ such that

$$\triangle \boldsymbol{c}_j \geq \triangle c_j^* = c_j - \frac{(P^T \boldsymbol{w})_j}{(P^T \boldsymbol{w})_{i_C}} c_{i_C},$$

for some improvement action dimension $j \leq M_+$ that it wants to incentivize the agents.

Then for the decision maker, maximizing its AS utility is equivalent as maximizing the subsidy surplus, so the decision maker solves

$$\begin{aligned} \text{maximize}_{j, \triangle \boldsymbol{c}_j, \underline{c}, \overline{c}} \quad & \int_{\mathcal{X}} [Pr(f(\boldsymbol{x} + P\boldsymbol{a}_A^*(\boldsymbol{x})) = y_A') - \mathbf{1}\{\triangle \boldsymbol{c}^T \boldsymbol{a}_A^*(\boldsymbol{x}) \in [\underline{c}, \overline{c}]\}] p(\boldsymbol{x}) d\boldsymbol{x} \\ \text{subject to} \quad & \triangle \boldsymbol{c} \in [\triangle \boldsymbol{c}^*, c_j), j \leq M_+ \end{aligned}$$

where the problem can be non-convex and not monotone for general p and l. Specifically, when j has the highest return of investment after the discount, the backward induction that anticipates the agent's AS best response is,

$$\boldsymbol{a}_A^*(\boldsymbol{x}) = \begin{cases} \frac{\tau - \boldsymbol{w}^T \boldsymbol{x}}{(P^T \boldsymbol{w})_j} \boldsymbol{e}_j, \text{ if } \frac{\triangle c_j (\tau - \boldsymbol{w}^T \boldsymbol{x})}{(P^T \boldsymbol{w})_j} \in [\underline{c}, \overline{c}], \\ \frac{\tau - \boldsymbol{w}^T \boldsymbol{x}}{(P^T \boldsymbol{w})_{i_C}} \boldsymbol{e}_{i_C}, \text{ o.w.} \end{cases}$$

This indicates that agents with $\boldsymbol{x}$ in a belt shape subset of $\mathcal{X}$ will be incentivized to improve, but the overall subsidy surplus is in general not convex, not concave and not monotone in either the upper bound (determined by f and $\bar{c}$) or the lower bound (determined by f and $\underline{c}$) of the belt even when the other is fixed. Moreover, the minimum effective discount value $\triangle c_j^*$ is not always the optimal solution, adding more complexity to the problem. This is because sometimes the decision maker wants to put more discount on the action dimension and incentivize some agents outside of the manipulation margin to improve and accept them rather than reject them. For example, if 80 percent agent has attribute that makes their likelihood 0.49, the minimum effective discount value still makes them rejected and take 0 AS best response, but a slightly higher discount can incentivize them all to improve to the threshold value, say 0.7, the $0.7 - (1 - 0.49) \cdot 0.8 = 0.152$ amount of improvement may largely outweigh the extra subsidy cost.

Overall speaking, the difference between $\boldsymbol{w}$ and $\boldsymbol{\theta}$ in f, the global properties of p, l and their local properties influenced by τ all makes the problem hard to solve. □

B.4 Proof of Theorem 3.6

PROOF. We will show that any $G \neq 0$ returned by Algorithm 1 is IC, IR and satisfies $S(f, G) \geq 0$.

The IC part follows that the participants act in self-interest. Also, as previously discussed, the minimum effective discounted value $\triangle c_j = \triangle c_j^* = c_j - \frac{(P^T \boldsymbol{w})_j}{(P^T \boldsymbol{w})_{i_C}} c_{i_C}$ makes sure the agents are weakly better off in the AS game than the CS game (given the same f).

We note that for all f that incentivizes gaming, the decision maker would prefer $\boldsymbol{w} = \boldsymbol{\theta}$ and we can use Theorem 3.7 to find G, so below we have $i_C \leq M_+$.

The basic logic of ensuring $S(f, G) \geq 0$ is that the algorithm finds a specific agent that is incentivized, and if this specific agent has a non-negative individual subsidy surplus, it is sufficient that all the other incentivized agents also have non-negative individual subsidy surplus and thus $S(f, G) \geq 0$.

In Algorithm 1, the designer finds (a convex problem and easy to solve)

$$\underline{\boldsymbol{x}} = \underset{\boldsymbol{x}: \boldsymbol{w}^T \boldsymbol{x} = \tau - \delta_j (P^T \boldsymbol{w})_j}{\arg\min} \boldsymbol{\theta}^T \boldsymbol{x},$$

which is the attribute of the specific agent. From the upper bound set on δ_j in the algorithm, we assume the specific agent is in $\mathcal{M}(f)$, and then uses

$$\underline{s} = l_+ - \delta_j \triangle c_j = l(\boldsymbol{\theta}^T (\underline{\boldsymbol{x}} + \delta_j \hat{P} \boldsymbol{e}_j)) - l(\boldsymbol{\theta}^T (\underline{\boldsymbol{x}} + \delta_{i_C} \hat{P} \boldsymbol{e}_{i_C})) - \delta_j \triangle c_j,$$

as a benchmark, where δ_j is the δ in the algorithm and $\delta_{i_C} = \frac{(P^T \boldsymbol{w})_j}{(P^T \boldsymbol{w})_{i_C}} \delta_j$. $\delta_j \boldsymbol{e}_j$ and $\delta_{i_C} \boldsymbol{e}_{i_c}$ help the agent achieve the same $\boldsymbol{w}^T \boldsymbol{z}$, $\underline{c} = 0, \bar{c} = \delta_j \triangle c_j$ here.

Then $\underline{s}$ is the specific agent's individual subsidy surplus, i.e.,

$$s(\underline{\boldsymbol{x}}, f, G) = l(\boldsymbol{\theta}^T (\underline{\boldsymbol{x}} + \hat{P} \boldsymbol{a}_A^*(\underline{\boldsymbol{x}}))) - l(\boldsymbol{\theta}^T (\underline{\boldsymbol{x}} + \hat{P} \boldsymbol{a}_C^*(\underline{\boldsymbol{x}}))) - \mathbf{1}\{\triangle \boldsymbol{c}^T \boldsymbol{a}_A^*(\boldsymbol{x}) \in [\underline{c}, \bar{c}]\} = \underline{s}.$$

We start with agents with CS best response $\boldsymbol{a}_C^*(\boldsymbol{x}) = \delta_{i_C} \boldsymbol{e}_{i_C}$, i.e., $\boldsymbol{w}^T \boldsymbol{x} = \boldsymbol{w}^T \underline{\boldsymbol{x}}$. For them, the AS best response is $\boldsymbol{a}_A^*(\boldsymbol{x}) = \delta_j \boldsymbol{e}_j$, the individual subsidy surplus is then

$$s(\boldsymbol{x}, f, G) = l(\boldsymbol{\theta}^T (\boldsymbol{x} + \delta_j \hat{P} \boldsymbol{e}_j)) - l(\boldsymbol{\theta}^T (\boldsymbol{x} + \delta_{i_C} \hat{P} \boldsymbol{e}_{i_C})) - \delta_j \triangle c_j,$$

since (1) $\boldsymbol{\theta}^T \hat{P} (\delta_j \boldsymbol{e}_j - \delta_{i_C} \boldsymbol{e}_{i_C})$ is constant, (2) $\boldsymbol{\theta}^T \boldsymbol{x} \geq \boldsymbol{\theta}^T \underline{\boldsymbol{x}}$ and (3) l is convex on this range, we have $s(\boldsymbol{x}, f, G) \geq \underline{s} \geq 0$.

For agents with "higher endowment" $\boldsymbol{w}^T \boldsymbol{x} > \boldsymbol{w}^T \underline{\boldsymbol{x}}$, i.e., with CS best response $\boldsymbol{a}_C^*(\boldsymbol{x}) = \alpha_{i_C} \boldsymbol{e}_{i_C}$, $\alpha_{i_C} < \delta_{i_C}$, we denote $\alpha_j = \alpha_{i_C} (P^T \boldsymbol{w})_{i_C} / (P^T \boldsymbol{w})_j$, then the (sub-optimal) AS best response is $\boldsymbol{a}_A^*(\boldsymbol{x}) =$

$\alpha_j \boldsymbol{e}_j$, and the individual subsidy surplus is

$$
\begin{aligned}
s(\boldsymbol{x}, f, G) &= l(\boldsymbol{\theta}^T(\boldsymbol{x} + \alpha_j \hat{P}\boldsymbol{e}_j)) - l(\boldsymbol{\theta}^T(\boldsymbol{x} + \alpha_{i_C}\hat{P}\boldsymbol{e}_{i_C})) - \alpha_{i_C}\bar{c}/\delta_{i_C} \\
&\geq \frac{\alpha_{i_C}}{\delta_{i_C}}[l(\boldsymbol{\theta}^T(\boldsymbol{x} + \delta_j \hat{P}\boldsymbol{e}_j)) - l(\boldsymbol{\theta}^T(\boldsymbol{x} + \delta_{i_C}\hat{P}\boldsymbol{e}_{i_C})) - \bar{c}] \\
&\geq \frac{\alpha_{i_C}}{\delta_{i_C}}\underline{s} \geq 0,
\end{aligned}
$$

where the second inequality comes from the convexity of l.

For agents with "lower endowment" i.e., with CS best response $\boldsymbol{a}_C^*(\boldsymbol{x}) = \beta_{i_C}\boldsymbol{e}_{i_C}$, $\beta_{i_C} < \delta i_C$, the mechanism designer suggest that they break tie choosing $\boldsymbol{a}_A^*(\boldsymbol{x}) = \beta_{i_C}\boldsymbol{e}_{i_C}$ as the AS best response and thus the individual subsidy surplus is 0. For $\beta_j = \beta_{i_C}(P^T\boldsymbol{w})_{i_C}/(P^T\boldsymbol{w})_j$, we note that $\boldsymbol{a}_A^*(\boldsymbol{x}) = \beta_j \boldsymbol{e}_j$ is a dominated strategy since $\triangle \boldsymbol{c}^T \boldsymbol{a}_A^*(\boldsymbol{x}) > \bar{c}$.

□

We see from the proof that when l is convex, agents with "high endowment" will have "high return on investment" for the mechanism designer when utilizing the discount. On the other hand, if l is concave on $[0, \max_{\boldsymbol{x}:\boldsymbol{w}^T\boldsymbol{x}=\tau} l(\boldsymbol{x})]$, we can infer that the agents with "low endowment" will have "high return on investment" when utilizing the discount. So then finding a suitable $\underline{c}$ becomes important, finding the approximate optimal mechanism can follow similar steps in Algorithm 4.

In general, real world data like FICO shows that the likelihood function has an S-shape and is concave on higher score range, and choosing a threshold too high hurts the decision maker.

We also note that the minimum effective discount value is used because it is also "sufficient". For convex l, if an agent cannot guarantee a non-negative individual subsidy surplus under the minimum effective discount value, it can not have a non-negative individual subsidy surplus for any other effective discount value. Not only because the individual subsidy cost goes up, but also because the marginal quality improvement is lower for agents farther away from the boundary while the marginal cost is constant.

ALGORITHM 4: Extended Grid Search an IC, IR and BB Discount Mechanism for Classification

```
Choose $\epsilon > 0$, set $\bar{c}_{max} \leftarrow 0$, $ans \leftarrow (\mathbf{0}, 0)$;
Define $a(r, j) = (\tau - r)\boldsymbol{e}_j/(\hat{P}^T\boldsymbol{w})_j$;
Define $r(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x}$;
Define $s(\boldsymbol{x}, j, \triangle\boldsymbol{c}) = l(\boldsymbol{\theta}^T(\boldsymbol{x} + \hat{P}a(r, j))) - l(\boldsymbol{\theta}^T(\boldsymbol{x} + \hat{P}a(r, i_C))) - (\tau - r)\triangle c_j/(\hat{P}^T\boldsymbol{w})_j$;
for $j = 1 : M_+$ do
    $\triangle\boldsymbol{c} \leftarrow \mathbf{0}$; $\bar{c} \leftarrow 0$;
    $S \leftarrow 0$;
    $\triangle c_j \leftarrow c_j - \frac{(P^T\boldsymbol{w})_j}{(P^T\boldsymbol{w})_{i_C}} c_{i_C}$;
    while $S \geq 0$ and $\delta_j \leq \underline{\boldsymbol{x}}_j$ do
        $\bar{c} \leftarrow \bar{c} + \epsilon$;
        $\underline{r} \leftarrow \tau - \delta_j(\hat{P}^T\boldsymbol{w})_j$;
        $S \leftarrow \int_{\{\boldsymbol{x}: r(\boldsymbol{x}) \in [\underline{r}, \tau]\}} s(\boldsymbol{x}, j, \triangle\boldsymbol{c})p(\boldsymbol{x})d\boldsymbol{x}$;
        if $S > S_{max}$ then
            $S_{max} \leftarrow S$; $ans \leftarrow (\triangle\boldsymbol{c}, \bar{c})$;
        end
    end
end
Return $ans$.
```

B.5 Proof of Theorem 3.7

PROOF. When $\boldsymbol{w} = \boldsymbol{\theta}$, the mechanism designer is indifferent about AS best responses along any improvement action dimension.

The mechanism designer find the "cheapest to incentivize" target action dimension

$$
i_A = \arg\max_{j \leq M_+} (P^T\boldsymbol{\theta})_j / c_j \;\Leftrightarrow\; i_A = \arg\min \frac{\triangle c_j^*(\tau - \boldsymbol{\theta}^T\boldsymbol{x})}{(P^T\boldsymbol{\theta})_j}
$$

and set $\triangle \boldsymbol{c}$ so that $\triangle c_{i_A} \geq \triangle c^*_{i_A} = c_{i_A} - \frac{(P^T\boldsymbol{\theta})_{i_A}}{(P^T\boldsymbol{\theta})_{i_C}} c_{i_C}$.

The choice of $\bar{c}$ depends on the individual subsidy surplus, which is the quality improvement of an incentivized agent minus the subsidy cost, denote $r_{\boldsymbol{x}} = \boldsymbol{\theta}^T\boldsymbol{x}$, then [9]

$$s(r_{\boldsymbol{x}}, f, G_d) := l(\tau) - l(r_{\boldsymbol{x}})\mathbf{1}\{\boldsymbol{x} \in \mathcal{M}(f)\} - [1 - l(r_{\boldsymbol{x}})]\mathbf{1}\{\boldsymbol{x} \notin \mathcal{M}(f)\} - \frac{(\tau - r_{\boldsymbol{x}})\triangle c^*_{i_A}}{(P^T\boldsymbol{\theta})_{i_A}}, \qquad (21)$$

which is because when agents break tie choosing the action with the largest improvement, we have

$$\boldsymbol{\theta}(\boldsymbol{x} + \hat{P}\boldsymbol{a}^*_A(\boldsymbol{x})) = \tau.$$

When the minimum effective discount value is chosen, and the condition

$$l(\tau) - l(\underline{r}_f) \leq \frac{(\tau - \underline{r}_f)\triangle c^*_{i_A}}{(P^T\boldsymbol{\theta})_{i_A}} = (\tau - \underline{r}_f)\left[\frac{c_{i_A}}{(P^T\boldsymbol{w})_{i_A}} - \frac{c_{i_C}}{(P^T\boldsymbol{w})_{i_C}}\right] \qquad (22)$$

holds, all incentivized agents satisfy $\boldsymbol{x} \in \mathcal{M}(f)$ and $s(r_{\boldsymbol{x}}, f, G_d) = l(\tau) - l(r_{\boldsymbol{x}}) - \frac{(\tau - r_{\boldsymbol{x}})\triangle c^*_{i_A}}{(P^T\boldsymbol{\theta})_{i_A}}$, which is concave in $r, \forall r \leq \tau$ since l is convex on $[0, \tau]$. A rational decision maker will make sure that an agent with $r_{\boldsymbol{x}} \geq 0.5$ is in $\mathcal{M}(f)$, and $r_{\boldsymbol{x}} < 0.5$ is not. And similar to the case in Theorem 3.6, agents that fully spends $\bar{c}$ but still need $(\boldsymbol{a}_A)_{i_C} > 0$ are suggested to stick with their CS best responses.

The decision maker chooses $\bar{c}$ by

$$\bar{c} = (\tau - \underline{r})\triangle c^*_{i_A}/(P^T\boldsymbol{\theta})_{i_A}, \text{ where } \underline{r} = \arg\min_r \; s.t. \; s(r, f, G) \geq 0,$$

intuitively, it incentivizes every agent with non-negative individual subsidy surplus.

Here we highlight some of the key reasons why the mechanism is still IC, IR and satisfies $S(f, G)$ if the condition in (22) does not hold.

In fact, when $\boldsymbol{w} = \boldsymbol{\theta}$ in f, we can assume that a rational decision maker makes sure if $\boldsymbol{x} \notin \mathcal{M}(f)$, then $l(r_{\boldsymbol{x}}) < 0.5 \Leftrightarrow 1 - l(r_{\boldsymbol{x}}) > l(r_{\boldsymbol{x}})$. As a result, we know that

$$\bar{s}(r, f, G) = l(\tau) - l(r) - \frac{(\tau - r)\triangle c_{i_A}}{(P^T\boldsymbol{\theta})_{i_A}} \geq s(r, f, G),$$

is concave in r and

$$\underline{s}(r, f, G) = l(\tau) + l(r) - 1 - \frac{(\tau - r)\triangle c_{i_A}}{(P^T\boldsymbol{\theta})_{i_A}} \leq s(r, f, G),$$

is increasing in $r, \forall r \; s.t. \; l(r) < 0.5$. Therefore, if the condition in (22) does not hold, we have $l(\underline{r}) < 0.5$, where $\underline{r} = \arg\min_r \; s.t. \; s(r, f, G) \geq 0$ we can also conclude that $r_{\boldsymbol{x}} in[\underline{r}, \tau]$ satisfies $s(r_{\boldsymbol{x}}, f, G) \geq 0$, i.e., every agent incentivized has non-negative individual subsidy surplus. □

When (22) does not hold or f incentivizes improvement, the mechanism designer can approximate the optimal $\triangle c_{i_A}$ by doing a grid search on the value of $1/(c_{i_A} - \triangle c_{i_A})$ with step size ϵ to find the corresponding $\triangle c_{i_A}$ and the related optimal $\bar{c}$ like in Theorem 3.7, which guarantees $\max_G S(f, G) - \max S_{grid}(f, G_{grid})$ is $O(\epsilon)$ if $\max_v \int_v^{v+\epsilon} p_R(r)dr$ is $O(\epsilon)$.

This is because if the optimal discounted value is $\triangle\tilde{c}_{i_A}$, there is one scanned value $1/(c_{i_A} - \triangle c_{i_A})$ that is at most $\epsilon/2$ away from $1/(c_{i_A} - \triangle\tilde{c}_{i_A})$, meaning that

(1) the optimal scanned value at worst failed to incentivize $O(\epsilon)$ of agents to improve with a highest subsidy benefit of $O(\epsilon)$ and possibly an infinitesimal extra amount of subsidy cost;

[9] If $\boldsymbol{x} \in \mathcal{M}(f)$, incentivizing this agent will result in the same decision outcome and an improvement equilibrium qualification status and thus the subsidy benefit is $l(\tau) - l(r_{\boldsymbol{x}})$; if $\boldsymbol{x} \notin \mathcal{M}(f)$, subsidizing this agent will change the decision outcome from 0 to 1, and the subsidy benefit is $l(\tau) - [1 - l(r_{\boldsymbol{x}})]$. When applying the minimum effective discount value, the agent's equilibrium action cost is the same in AS and CS outcomes, and thus $\boldsymbol{x} \in \mathcal{M}(f)$ are incentivized to improve.

(2) the optimal scanned value at worst paid an extra subsidy cost of $O(\epsilon)$ ($O(\epsilon)$ probability mass of agents with individual payment no more than 1) and has no improved a highest subsidy benefit.

C SUPPLEMENTARY MATERIAL FOR SECTION 4

C.1 Proof of Theorem 4.1

Proof. Recall that the AS utility of the decision maker is

$$U_A^{(reg)}(f) = \int_X \mathbb{E}_\sigma\big[- \big(f(\boldsymbol{x} + P\boldsymbol{a}_A^*(\boldsymbol{x})) - y_A'\big)^2\big] p(\boldsymbol{x}) d\boldsymbol{x} - H(G),$$

which if we rewrite the equilibrium individual error as

$$\mathcal{E}(f, \boldsymbol{a}, \boldsymbol{x}) = [\boldsymbol{w}^T(\boldsymbol{x} + P\boldsymbol{a}) - \boldsymbol{\theta}^T(\boldsymbol{x} + \hat{P}\boldsymbol{a})]^2 + err(\sigma),$$

the objective becomes

$$U_A^{(reg)}(f) = \int_X -\mathcal{E}(f, \boldsymbol{a}, \boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} - H(G).$$

The integral part is non-concave for general $p(\boldsymbol{x})$.

On the other hand, for a target AS best response where $\alpha < 1$, $\boldsymbol{a}_A = \alpha \frac{B}{c_j - \triangle c_j^*} \boldsymbol{e}_j + (1-\alpha) \frac{B}{c_{i_C}} \boldsymbol{e}_{i_C}$, we have $H(G) = \bar{c} = \frac{\alpha B \triangle c_j^*}{c_j - \triangle c_j^*}$, and the where $H(G)$ is linear in $\boldsymbol{a}$. for a target AS best response where $\alpha > 1$, $\boldsymbol{a}_A = \alpha \frac{B}{c_j - \triangle c_j^*} \boldsymbol{e}_j$, we have

$$\frac{\alpha}{c_j - \triangle c_j^*} = \frac{1}{c_j - \triangle c_j} \quad \Leftrightarrow \quad \triangle c_j = \frac{(\alpha - 1)c_j + \triangle c_j^*}{\alpha},$$

and

$$H(G) = \frac{B \triangle c_j}{c_j - \triangle c_j} = \frac{B((\alpha - 1)c_j + \triangle c_j^*)}{c_j - \triangle c_j^*}.$$

We can similarly show that $H(G)$ is piece-wise affine in $\boldsymbol{a}$ and thus the entire objective is non-concave and the problem is non-convex. □

C.2 Proof of Theorem 4.2

Proof. This algorithm has two loops, making it finish in polynomial time.

The outer loop enumerates through all improvement action dimensions and chooses the minimum effective discount amount to incentivize the agents to take an AS best response $\boldsymbol{a}_A = \alpha \frac{B}{c_j - \triangle c_j^*} \boldsymbol{e}_j + (1-\alpha) \frac{B}{c_{i_C}} \boldsymbol{e}_{i_C}$, where $\alpha < 1$. The inner loop grid searches the α values for each j to see if an IC and IR and $S(f, G) > 0$, computes the corresponding $\bar{c}$ and keeps track of the G that generates the largest $S(f, G)$. □

C.3 Proof of Theorem 4.4

Proof. In the special case, if improvement is incentivized by the mechanism, it is the dominant strategy to use the minimum effective discount amount, since a higher discount achieves the same error reduction but a higher subsidy cost.

For an AS best response $\boldsymbol{a}_A = \alpha \frac{B}{c_j - \triangle c_j^*} \boldsymbol{e}_j + (1-\alpha) \frac{B}{c_{i_C}} \boldsymbol{e}_{i_C}$, where $\alpha < 1$, the alternative form of individual subsidy benefit is the reduction in the expected prediction error

$$(\boldsymbol{\theta}^T P \boldsymbol{a}_C^*)^2 - (1-\alpha)^2 (\boldsymbol{\theta}^T P \boldsymbol{a}_C^*)^2,$$

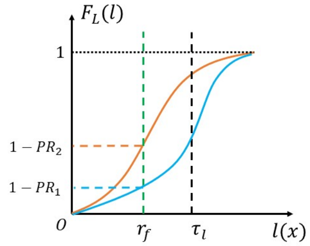


Fig. 16. An illustration of the CS DP gap when group 2 is disadvantaged in attributes.

the subsidy cost is $H(G) = \bar{c} = \frac{\alpha B \triangle c_j^*}{c_j - \triangle c_j^*}$, and thus we have the alternative individual subsidy urplus

$$s(\alpha) = (2\alpha - \alpha^2)(\boldsymbol{\theta}^T P \boldsymbol{a}_C^*)^2 - \alpha B \triangle c_{i_A} (c_{i_A} - \triangle c_{i_A})^{-1}.$$

□

D SUPPLEMENTARY MATERIAL FOR SECTION 5

D.1 Proof of Theorem 5.1

Proof. The DP gap is only related to $f(\boldsymbol{z})$ but not y or y', when the two groups have the same action cost but group 2 is disadvantaged in attribute, the implicit threshold (the lower side boundary of $\mathcal{M}^d(f)$, $\hat{\tau}_L$) is the same for both groups and from the definition of attribute disadvantage, $PR^{(1)} = 1 - F^{(1)}(\hat{\tau}_L) > 1 - F^{(2)}(\hat{\tau}_L) = PR^{(2)}$, and we know that the DP gap exists.

When f incentivizes gaming, the reason of an EO gap is similar as above $TPR^{(1)} = 1 - F_+^{(1)}(\hat{\tau}_L) > 1 - F_+^{(2)}(\hat{\tau}_L) = TPR^{(2)}$. If f incentivizes improvement, then the EO gap depends on both the CS TPR in both groups, the CS PR in both groups, and the AS quality improvement in both groups. For example, if G only not incentivize agents in the manipulation margins, then

$$TPR_A = 1 - FNR_A = 1 - FNR_C \cdot \frac{PR_C}{PR_A} = 1 - \frac{(1 - TPR_C) \cdot PR_C}{\triangle Q_A + PR_C},$$

and we know that the AS EO gap depends on $\triangle Q_A^{(1)}, \triangle Q_A^{(2)}$ which is based on $p^{(1)}(\boldsymbol{x})$ and $p^{(2)}(\boldsymbol{x})$ and we can not easily conclude the EO gap changes. □

D.2 Proof of Theorem 5.2

Proof. Part (1) is obvious since gaming results in no quality gain, and $\mathcal{M}^{(1)}(f) \supseteq \mathcal{M}^{(2)}(f)$ results in the quality gain gap if f incentivizes improvement and the DP gap no matter f incentivizes gaming or improvement.

If f incentivizes gaming, the reason of the EO gap is similar to that of the DP gap.

If f incentivizes improvement, again we can look at the formula for AS TPR

$$TPR_A = 1 - FNR_A = 1 - FNR_C \cdot \frac{PR_C}{PR_A} = 1 - \frac{(1 - TPR_C) \cdot PR_C}{\triangle Q_A + PR_C},$$

group 1 has a higher TPR_C, and a higher $\triangle Q_A / PR_C$, and thus a higher TPR_A and the EO gap always exists. □

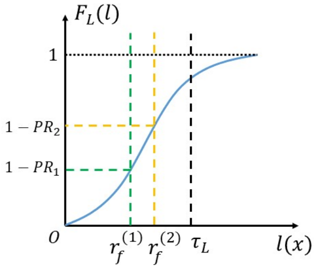


Fig. 17. An illustration of the CS DP gap when group 2 is disadvantaged in cost.

D.3 Proof and Discussion on Theorem 5.3

Proof. If group 2 is disadvantaged in cost, then it is cheaper to incentivize a group 1 agent than a group 2 agent to get the same qualification status improvement, and thus the decision maker subsidizes more group 1 agents and creates a quality gain gap. □

For DP gap, if G only incentivizes agents in the manipulation margins, the agents' CS or AS equilibrium decision remains the same.

For EO gap, we note that with G, the AS true positive increases in both groups and how the EO gap in classification changes depends on both the CS positive decision rate and the qualification status improvement and we do not have certain conclusions. For example, if G only not incentivize agents in the manipulation margins, then

$$TPR_A = 1 - FNR_A = 1 - FNR_C \cdot \frac{PR_C}{PR_A} = 1 - \frac{(1 - TPR_C) \cdot PR_C}{\triangle Q_A + PR_C},$$

we have $PR_C^{(1)} > PR_C^{(2)}$, $\triangle Q_A^{(1)} > \triangle Q_A^{(2)}$ and we can not easily conclude the EO gap changes. $FNR_A = 1 - FNR_C \cdot \frac{PR_C}{PR_A}$ because all false negative agents in CS remain to be false negatives in AS (the positive individuals with lower attribute than the lower side boundary of manipulation margins).

E SUPPLEMENTARY MATERIAL FOR SECTION 6

E.1 Proof of Theorem 6.1

Proof. We still need G to be IR for the decision maker, where the maximum tax a rational decision maker accepts is the subsidy benefit $\mathcal{T}(G) \leq S(f, G) + H(G)$, and the BB condition requires $S(f, G) + H(G) \geq \mathcal{T}(G) \geq H(G)$. So, as long as $S(f, G) \geq 0$, there is an IC, IR, and BB third party mechanism. Therefore, finding the optimal IC, IR, and BB third party mechanism is the same as

$$\text{maximize}_G\ W(f, G), \text{ subject to } S(f, G) \geq 0,$$

and if $S(f, G) > 0$ the mechanism can further improve its objective by setting the surplus at 0. □

E.2 Proof of Theorem 6.3

Proof. For Part (1), the fairness oriented third party can implement the ideal mechanism on group 2 and even further subsidize other group 2 agents to reduce the gap while avoiding subsidizing more group 1 agents to enlarge the fairness gaps.

For Part (2), any ideal mechanism makes sure the efficiency oriented third party has "remaining budget" to incentivize more agents to improve compared to AS-dm outcome and thus has the strictly highest equilibrium social quality improvement. □

E.3 Proof of Theorem 6.4

PROOF. If $h_A^{(1)}(\boldsymbol{a}) = h_A^{(2)}(\boldsymbol{a}), \forall \boldsymbol{a}$, then the equilibrium feature and attribute distribution are the same for both groups, and thus there is no fairness gap. Meanwhile, the subsidy benefit are the same in both groups, so the overall benefit is $S(f, G^{(1)}) + H(G^{(1)})$, and the overall subsidy cost is $p^{(1)}H(G^{(1)}) + p^{(2)}H(G^{(2)})$. □