Linear Models are Robust Optimal Under Strategic Behavior

Wei Tang [*]   Chien-Ju Ho [†]   Yang Liu [‡]

Abstract

There is an increasing use of algorithms to inform decisions in many settings, from student evaluations, college admissions, to credit scoring. These decisions are made by applying a decision rule to individual's observed features. Given the impacts of these decisions on individuals, decision makers are increasingly required to be transparent on their decision making to offer the "right to explanation." Meanwhile, being transparent also invites potential manipulations, also known as gaming, that the individuals can utilize the knowledge to strategically alter their features in order to receive a more beneficial decision.

In this work, we study the problem of *robust* decision-making under strategic behavior. Prior works in decision making under strategic behavior often assume that the decision maker has full knowledge of individuals' cost structure for manipulations. We study the robust variant that relaxes this assumption: The decision maker does not have full knowledge but knows only a subset of the individuals' available actions and associated costs. To approach this non-quantifiable uncertainty, we define robustness based on the worst-case guarantee of a decision, over all possible actions (including actions unknown to the decision maker) individuals might take. A decision rule is called *robust optimal* if its worst case performance is (weakly) better than that of all other decision rules. Our main results provide a crisp characterization of the above robust optimality: For any decision rules under mild conditions that are robust optimal, there exists a linear decision rule that is equally robust optimal. We believe this characterization promotes the use of simple linear decisions with uncertain individual manipulations.

1 Introduction

Algorithms have been increasingly engaged in making consequential decisions across a variety of sectors in our society. Examples include judges using defendant risk scores to set bail decisions and banks evaluating individuals' profiles to make loan decisions. In all these scenarios, the decision maker aims to determine a decision rule (or a model), which takes a set of individual's observed behavior or features as input, and output decisions that maximize some given utility function[1].

Given the consequential impacts to individuals, there is an increasing demand to make the decision rule transparant to offer "right to explanation" (See, for example, Goodman and Flaxman (2017)).

[*] Washington University in St. Louis; w.tang@wustl.edu

[†] Washington University in St. Louis; chienju.ho@wustl.edu

[‡] University of California, Santa Cruz; yangliu@ucsc.edu

[1] Throughout the work, we address the decision maker as "she" and the individual as "he". We also use the terms individual and agent interchangeably.

Transparency not only allows the public to audit models to mitigate potential fairness concerns but also enables the participants to understand what decisions they might receive if they have different features (See, for example, “right to recourse” (Ustun et al., 2019)). However, on the flip side, transparency simultaneously incentivizes individuals to “game” the deployed model. Specifically, if individuals understand how their observed features affect decisions, they may strategically alter their features to obtain a more favorable decision.

In response to this “gaming” behavior, there has been a recent flurry of work in studying these decison making under strategic behavior (Brückner et al., 2012; Brückner and Scheffer, 2011; Hardt et al., 2016; Kleinberg and Raghavan, 2019; Alon et al., 2020). To make the analysis tractable, almost all the works explicitly assume the decision maker has the *full knowledge* of costs that agents would incur to manipulate their behavior, as well as the action spaces that the agents might take actions from. The above knowledge enables a game theoretic analysis with characterizing agents' best responses when offered a particular decision rule.

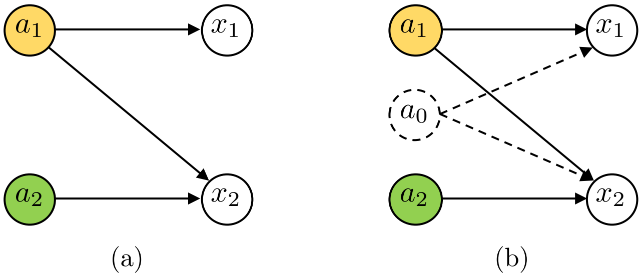


Figure 1: An instance of student evaluation problem.

However, the “full information” assumption is often not true in practice. Consider an example of student evaluation in Fig. 1 (Kleinberg and Raghavan, 2019; Alon et al., 2020). The student's observed features are their exam score (x_1) and homework score (x_2). The student can choose to take either actions, studying (a_1) or copying homework answers (a_2), to alter their features. Studying improves both exam score (x_1) and homework score (x_2), while copying homework only improves homework score. The teacher evaluates the student through a final score, which is a function of x_1 and x_2, and students are assumed to aim to maximize their final score minus the cost of the actions. If the teacher knows the actions a_1 and a_2, and they are indeed the only actions the student can take, a decision rule can be designed (outputs a final score as a function of x_1 and x_2) that maximizes some given objective by considering students' best responses. However, in practice, the teacher might not be aware of the full set of actions the student can take. For instance, the student might consider taking action a_0 unknown to the teacher (in Fig. 1b), such as hiring a tutor or working with other students. With this incomplete knowledge of the student's actions, how should the teacher design her evaluation rule?

In this work, we answer the above question by studying the design of *robust* optimal decision rules with strategic agent, where we relax the assumption of complete knowledge over agent actions. We define the robustness notion as used in robust contract design (Carroll, 2015): Evaluate the worst-case guarantee of a decision, over all possible actions (including actions unknown to the decision maker) agents might take. More formally, the decision maker only knows a subset of actions (denoted by $\mathcal{A}_d$) among all the actions available to the the agent (denoted by $\mathcal{A}_a$). Let $V_d(f|\mathcal{A}_a)$ be the utility the decision maker obtains with decision rule f when the agent's action space is $\mathcal{A}_a$. The decision maker's goal is to maximize her *worst-case* performance $V_d(f)$ over all possible actions the agent may have access to ($\mathcal{A}_a \supseteq \mathcal{A}_d$):

$$\max_f V_d(f) = \max_f \inf_{\mathcal{A}_a \supseteq \mathcal{A}_d} V_d(f|\mathcal{A}_a). \tag{1}$$

A decision rule f^* is robust optimal if it achieves the maximum of the above worst-case utility.

Our contribution We formalize the problem of robust strategic decision-making and characterize the robust optimal decision rules. We show that under mild conditions, any decision rule could be (weakly) improved by a linear decision rule, in terms of the worst-case performance. Therefore, for any robust optimal decision rule, there exists a linear rule that is equally robust optimal. Our proof hinges on a construction of two convex sets, and a key disjointness argument over these two convex sets. This enables us to apply hyperplane separation theorem to prove that for any non-linear decision rule, there exists a linear one which gives a weakly greater guarantee to the decision maker.

We also explore the computational problem of searching for the robust optimal f^*. Given the characterization, it suffices to search over the space of linear decision rules. However, we show that finding f^* is generally NP-hard and offer conditions when efficient solvers might exist.

1.1 Related Work

Our problem closely connects to recent literature in studying classification algorithms in the presence of strategic manipulation (Hardt et al., 2016; Brückner et al., 2012; Brückner and Scheffer, 2011). Specifically, Hardt et al. (2016) study the design of optimal classification when the agents can incur costs to manipulate their features. Motivated by fairness concerns, Hu et al. (2019) and Milli et al. (2019) consider the scenario in which the costs for manipulation are different for different groups and explore the societal impacts. There are also works directly utilizing the decision rule as an incentive device (Kleinberg and Raghavan, 2019; Alon et al., 2020; Haghtalab et al., 2020; Ball, 2020; Dong et al., 2018; Tabibian et al., 2019; Miller et al., 2019). Among theses works, Kleinberg and Raghavan (2019) is closest to our work: they introduce a graphic model to capture the known agent's available actions and show that simple linear mechanisms suffice for a single known agent. Alon et al. (2020) then extends the discussion to multiple agents. Our work departs from the above works in the sense that the decision maker only has incomplete knowledge of the agent's cost structure or his available actions.

Our formulation resembles the principal-agent problem in contract theory (Grossman and Hart, 1992; Shavell, 1979; Holmstrom and Milgrom, 1987), which also studies the strategic interplay between two interest-misaligned parties. Our characterization of robust decision rule follows the works on robust contract design (Carroll, 2015; Dai and Toikka, 2017; Miao and Rivera, 2016; Carroll and Segal, 2019; Carroll, 2017; Diamond, 1998; Hansen and Sargent, 2012; Chassang, 2013) in which robustness is defined as the worst-case optimal mechanisms in various settings. Our model differs from this line of research in that the decision maker determines a decision rule (instead of a "contract" in contract theory) that is multi-dimensional and could take arbitrary forms. Moreover, we do not restrict the decision maker's utility to be in additive form (reward minus the payment). We generalize the utility to be arbitrary function that satisfies some mild conditions. Other computational approaches to contract design in computer science community can be found in the work of (Dütting et al., 2019; Babaioff et al., 2006; Ho et al., 2016; Babaioff et al., 2010). In addition, our work shares similar flavor for max-min analysis in worst-case algorithmic analysis (Azar et al., 2013; Bandi and Bertsimas, 2014). In all of these works, the setting and the formulation are different from the ones we consider in the present work.

Our work complements a recent literature on discussing the effects of linear models in social stratification. For example, Wang et al. (2018) extend the notion of interpretability to credibility and discuss the credibility in a linear setting. Fawzi et al. (2018) analyze the robustness of linear classifiers to adversarial perturbations. Ustun and Rudin (2014) and Ustun et al. (2019) discuss the interpretability and right to recourse in linear classification.

2 A Model of Robust Strategic Decision-making

In this section, we establish notations and then formalize our model. Agent features are represented by a vector $\mathbf{x} = (x_1, \ldots, x_n)$, which takes value in a compact set (it may be finite or infinite) $\mathcal{X} \subseteq \mathbb{R}^n$. An action of the agent can be represented by the outcome and the cost of the action. We use a pair $(F, c) \in \Delta(\mathcal{X}) \times \mathbb{R}_+$ to denote an action, where F is the outcome, i.e., the distribution of the agent features after taking a particular action, and c is the associated cost. Using (F, c) simplifies our presentation, and note that the physical meaning of an action does not carry much weights. All that matter to the agent are the outcome features and the cost. The cost c can be interpreted as effort, monetary cost, or as simply describing the agent's preferences over the available actions. The decision maker cannot observe the agent's action but can only observe the features, the realized outcome of the action.

Action set We define two important action sets $\mathcal{A}_a$ and $\mathcal{A}_d$. In particular, $\mathcal{A}_a \subseteq \Delta(\mathcal{X}) \times \mathbb{R}_+$ is the set of all possible actions that the agent can take, and $\mathcal{A}_d$ is the set of action that the decision maker is aware of. While the decision maker only knows $\mathcal{A}_d$ and not $\mathcal{A}_a$, she knows that $\mathcal{A}_d \subseteq \mathcal{A}_a$. The decision maker's *unquantifiable* uncertainty of $\mathcal{A}_a$ is the key conceptual element of this work. Using the student evaluation example, the available actions to the student $\mathcal{A}_a$ could be (studying, cheating, hiring tutors). The teacher only knows $\mathcal{A}_d$, (studying, cheating), a subset of $\mathcal{A}_a$ but aims to design a decision rule that is robust to this uncertainty.

Decision rule A decision rule $f : \mathcal{X} \to \mathbb{R}_+$ is a mapping from the agent's features to a decision, where the decision domain of f is normalized to be positive. For example, a FICO credit score is a positive numeric value, which predicts a consumer's creditworthiness from multiple features including his length of credit history and default rate. The decision rule f is contingent only on the observable features, but not on the actions that are not observable to the decision maker.

The decision maker aims to maximize her utility function $h : \mathcal{X} \to \mathbb{R}_+$. This function should be sufficient to characterize the utility that the agents would bring to the designer. For example, it could be a qualification function, assuming that agent's effort investment in changing their features may actually lead sometimes to self-improvement, thus in their true qualifications. We assume that there's an upper bound $\bar{C} > 0$ of $f(\mathbf{x})$ for any $\mathbf{x} \in \mathcal{X}$. In addition, we also define the following simple class of decision rules:

Definition 1 (Linear decision rule)**.** A decision rule f is *linear* if f is a linear function of the feature[2], i.e., $f(\mathbf{x}) = \alpha^\top \mathbf{x} + \beta$ for $\alpha \in \mathbb{R}^n$ and $\beta \in \mathbb{R}$. Let $\mathcal{G}^{\text{lin}} = \left\{(\alpha, \beta) \in \mathbb{R}^n \times \mathbb{R} : f(\mathbf{x}) = \alpha^\top \mathbf{x} + \beta \in [0, \bar{C}], \forall \mathbf{x} \in \mathcal{X}\right\}$ be the space of parameter pair (α, β).

The interaction between the decision maker and the agent goes as follows: (1) the decision maker publishes a decision rule f based on the knowledge of $\mathcal{A}_d$; (2) the agent, knowing $\mathcal{A}_a$, chooses action $(F, c) \in \mathcal{A}_a$ to respond to f; (3) the agent features are then moved to $\mathbf{x} \sim F$; (4) the decision maker derives utility of $h(\mathbf{x})$ and the agent derives utility of $f(\mathbf{x}) - c$.

Robustness of decision making under strategic behavior We first characterize the agent's behavior. Given the decision rule f and his available action set $\mathcal{A}_a$, the agent obtains expected utility $\mathbb{E}_F[f(\mathbf{x})] - c$ for taking action (F, c). Let $\mathcal{A}_a^*(f|\mathcal{A}_a)$ be the set of actions that maximize the agent's utility, and $V_a(f|\mathcal{A}_a)$ be the corresponding utility:

$$\mathcal{A}_a^*(f|\mathcal{A}_a) = \arg\max_{(F,c)\in\mathcal{A}_a} \left(\mathbb{E}_F[f(\mathbf{x})] - c\right), \quad V_a(f|\mathcal{A}_a) = \max_{(F,c)\in\mathcal{A}_a} \left(\mathbb{E}_F[f(\mathbf{x})] - c\right).$$

When there are multiple maximizers of the agent's objective, the agent may choose the action that is the most beneficial to the decision maker. The expected utility of the decision maker, given decision rule f and the action set available to the agent $\mathcal{A}_a$ is

$$V_d(f|\mathcal{A}_a) = \max_{(F,c)\in\mathcal{A}_a^*(f|\mathcal{A}_a)} \mathbb{E}_F[h(\mathbf{x})].$$

Note that the decision maker only knows $\mathcal{A}_d$ but not $\mathcal{A}_a$. Therefore, she cannot optimize $V_d(f|\mathcal{A}_a)$ directly. To address this nonquantifiable uncertainty, we define $V_d(f)$ as the worst case utility the decision maker obtains over all possible actions sets $\mathcal{A}_a$ that are supersets of $\mathcal{A}_d$:

$$V_d(f) = \inf_{\mathcal{A}_a \supseteq \mathcal{A}_d} V_d(f|\mathcal{A}_a). \tag{2}$$

We define the robust optimal decision rule f^* as the one that maximizes $V_d(f)$, since it is robust to any action (even unknown to the decision maker) the agent might take:

$$f^* \in \arg\max_f V_d(f) = \arg\max_f \inf_{\mathcal{A}_a \supseteq \mathcal{A}_d} V_d(f|\mathcal{A}_a). \tag{3}$$

Moreover, in this work, we will focus on the following decision rules:

Definition 2. A decision rule f is *eligible* if $V_d(f) > V_d(0)$ and $V_d(0) > 0$, where 0 denotes the zero decision rule.

Eligibility implies that we only care about scenarios that the decision maker would get benefits from posting the decision rule. While there might exist scenarios that the decision maker gets worse utility by taking interventions (posting decision rules), we focus on the positive scenario in this work.

[2] More precisely, it is an affine decision rule with the form of $\alpha^\top \mathbf{x} + \beta$.

3 Main results: Linear Model is Robust Optimal Under Strategic Behavior

In this section we establish our main result that there exists a linear decision rule that is robust optimal. The main result in this section is summarized as follows:
Theorem 1. *There exists a decision rule f that maximizes $V_d(f)$ and is linear, namely: $f \in \arg\max V_d(f)$, where $f(\mathbf{x}) = \beta + \alpha^\top \mathbf{x}$, for some $\alpha \in \mathbb{R}^n, \beta \in \mathbb{R}$.*

The above theorem characterizes the robust optimal decision rule defined in (3). The key implication of the theorem is that, when aiming to find the robust optimal model against strategic responses, it suffices to only consider linear models.

In the reminder of this section, we provide the proof sketch of the theorem and then use an example to demonstrate our results. The proof consists of three main steps. We first characterize the properties of the worst case utility $V_d(f)$ for a given decision rule f; we then show that any nonlinear decision rule can be (weakly) improved by a linear decision rule in terms of the worst case utility. Finally, we wrap up by showing the existence of an optimal linear decision rule in the linear decision space.

3.1 Characterizing the worst-case utility $V_d(f)$

We first characterize the worst-case utility guarantee for any given eligible decision rule.
Lemma 1. *Let f be any eligible decision rule. Define a set $\Gamma = \{F \in \Delta(\mathcal{X}) : \mathbb{E}_F[f(\mathbf{x})] \geq V_a(f|\mathcal{A}_d)\}$. Then one of the following two cases occurs:*

$$(i) \quad V_d(f) = \min_{F\in\Gamma} \mathbb{E}_F[h(\mathbf{x})]; \tag{4}$$

$$or \quad (ii) \max_{F\in\Delta(\mathcal{X})} \mathbb{E}_F[f(\mathbf{x})] = V_a(f|\mathcal{A}_d). \tag{5}$$

Moreover, for F attaining the minimum in (4), then the inequality in Γ will reduce to equality at F.

The key message of this lemma is that, we can replace the definition of $V_d(f)$ in (2), that depends on unknown $\mathcal{A}_a$, with an expression that depends only on variables known to the decision maker. In particular, in case (i), this is given by identifying F which is constrained by $V_a(f|\mathcal{A}_d)$ using the designer's knowledge $\mathcal{A}_d$. In case (ii), we know that the best response from the agent is indeed in $\mathcal{A}_d$, so again the designer can focus on the action space she is aware of.

We now sketch the proof as follows, the details of which we relegate to the Appendix A.

Proof Sketch. Note that for any action set $\mathcal{A}_a \supseteq \mathcal{A}_d$ the agent has, and any optimal action (F, c) he chooses under $\mathcal{A}_a$ and the eligible decision rule f, his expected utility he gets from f must satisfy:

$$\mathbb{E}_F[f(\mathbf{x})] \geq \mathbb{E}_F[f(\mathbf{x})] - c = V_a(f|\mathcal{A}_a) \geq V_a(f|\mathcal{A}_d). \tag{6}$$

Here the second inequality holds because $\mathcal{A}_a$ contains $\mathcal{A}_d$, and having more actions available can only make the agent better off. Thus, for any decision rule f, the agent will only take the actions that guarantee himself a utility that is at least $V_a(f|\mathcal{A}_d)$, these action actually formulates the set Γ.

Furthermore, the decision maker's utility $V_d(f|\mathcal{A}_a) = \mathbb{E}_F[h(\mathbf{x})]$ is at the least the minimum given by Eqn. (4). Thus, we have $V_d(f) \geq \min_{F\in\Gamma} \mathbb{E}_F[h(\mathbf{x})]$. To show this is actually tight, we then prove the other direction. To achieve that, we construct some worst case action set $\mathcal{A}_a$ to guarantee that $V_d(f)$ cannot exceed $\min_{F\in\Gamma} \mathbb{E}_F[h(\mathbf{x})]$. Case (ii) is simply the boundary case in which the agent's best action under any possible actions sets is already included in $\mathcal{A}_d$. □

3.2 Improving nonlinear decision rule to a linear rule

Having characterized the worst-case utility guarantee of decision maker, we can now show that any nonlinear decision rule can be (weakly) improved by a linear decision rule in terms of its $V_d(f)$.

Lemma 2. *Fix any h and any (nonlinear) eligible decision rule f, there exists a linear one f' such that: $V_d(f') \geq V_d(f)$.*

We give a proof sketch of Lemma 2 below, the detailed proof of this lemma is deferred to Appendix B.

Proof Sketch. We illustrate our proof sketch graphically using Figure 2. At a very high-level, we show that for every decision rule f, we can construct two convex sets, with one containing information about the agent and one about the decision maker. We then show that the two convex sets are disjoint, and therefore there exists a hyperplane that separates the two convex sets. Then it turns out that separating hyperplane is the linear decision rule that weakly improves on f.

Given a decision rule f, consider a point $(\mathbb{E}_F[\mathbf{x}], \mathbb{E}_F[f(\mathbf{x})])$ generated by any possible action (F, c). This point will be in the convex hull of $(\mathbf{x}, f(\mathbf{x}))$. We define $\mathcal{S}$ to be the convex hull of all pairs $(\mathbf{x}, f(\mathbf{x}))$, for $\mathbf{x} \in \mathcal{X}$. To construct another convex set, we separately consider the two cases in Lemma 1. For case (i), we define $t(\mathbf{x}) = \max\{V_a(f|\mathcal{A}_d), h(\mathbf{x}) + f(\mathbf{x}) - V_d(f)\}$. Intuitively, $t(\mathbf{x})$ is constructed to accommodate the constraint in the set Γ for Eqn. (4). We define $\mathcal{T}$ as the convex hull of all pairs $(\mathbf{x}, z)$ that $\mathbf{x}$ lies in the convex hull of $\mathcal{X}$, and $z > t(\mathbf{x})$. By utilizing the results in Lemma 1, we can show that the two convex sets are disjoint (details in Appendix). By hyperplane separation theorem, we can find a hyperplane f' separating $\mathcal{S}$ and $\mathcal{T}$. f' has two advantages: First it gives the agent the same incentive as f. Second, it gives a weakly greater guarantee to the decision maker. For case (ii), we change the set $\mathcal{T}$ to be the set of all $(\mathbf{x}, z)$ with $\mathbf{x}$ in the convex hull of $\mathcal{X}$ and $z > V_a(f|\mathcal{A}_d)$. Similar arguments in case (i) still apply here.

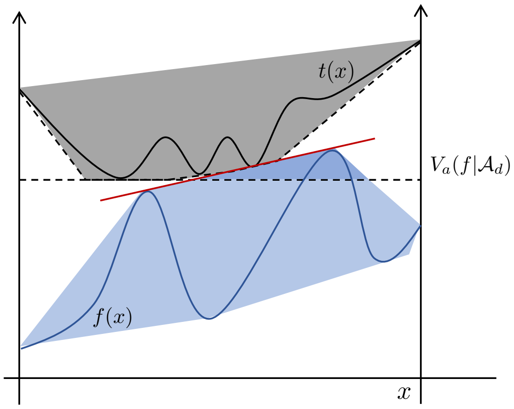


Figure 2: Illustrate $\mathcal{S}$ and $\mathcal{T}$ when $n = 1$. The blue line is $f(\mathbf{x})$ and its associated convex hull in blue shaded region (the top blue triangle is the set Γ). Black line is $t(\mathbf{x})$. The black shaded region is the convex hull for all points $(\mathbf{x}, z)$ where $z > t(\mathbf{x})$. The red line is the hyperplane to separate $\mathcal{S}$ and $\mathcal{T}$.

□

3.3 Wrapping up

We have shown that any eligible decision rule f can be (weakly) improved to a linear one. We now wrap up our analysis by showing the existence of an optimum within the class of linear decision rules.

Lemma 3. *There exists a robust optimal linear decision rule.*

Recall our definition of $\mathcal{G}^{\text{lin}}$ in Definition 1. The proof reduces to show that $V_d(f)$ is upper semi-continuous w.r.t. $(\alpha, \beta) \in \mathcal{G}^{\text{lin}}$, this guarantees that $V_d(f)$ has a maximum over the compact set $\mathcal{G}^{\text{lin}}$. We defer the proof to Appendix C.

3.4 Illustrating example: Student evaluation

We now use the example of student evaluation to demonstrate our results obtained in this section. We first illustrate the application of Lemma 2: For a particular nonlinear decision rule, we show how to find an improved linear decision rule. Then, we compute the worst case utility for both decision rules according to Lemma 1. Finally, we return to the environment with student being able to take actions unknown to the teacher, as depicted in Fig. 1b to discuss how these two decision rules perform.

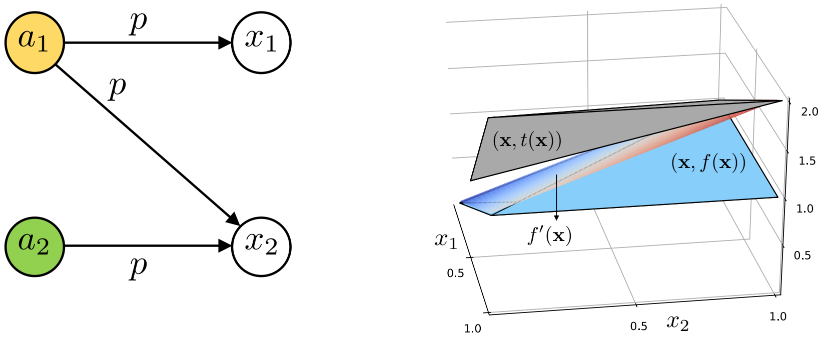
We first specify the environment details of our example. Suppose each feature is a binary variable in $\{0, 1\}$ (e.g., x_1: pass or fail the exam, x_2 : whether the homework is qualified or not). With an effort budget of size 1, the student needs to decide how to allocate his effort to each action (we denote effort for action a_j by e_j) [3]. The effort-feature conversion obeys the following rule: $\mathbb{P}(x_i = 1) = \sum_j w_{j,i} \cdot e_j$, where $w_{j,i} \in [0, 1]$ controls a weight on how the student's effort $e_j \in [0, 1]$ on action a_j contributes to the value of feature x_i. For example, a student may study for the exam and still fail with some (small) probability. The effort-feature conversion weights are detailed in Fig. 3a.

Suppose for a moment the student's available actions are $\{a_1, a_2\}$. The teacher wants to incentivize the student to invest all their efforts on studying (namely, the action a_1), thus, the teacher can set her utility function as $h(\mathbf{x}) = \alpha_h^\top \mathbf{x} + \beta_h$, where $\alpha_h = (1, 0)$ and β_h is a small positive value[4]. One (nonlinear) decision rule that maximizes $h(\mathbf{x})$ is $f(\mathbf{x}) = \max\{x_1, x_2\}$. It is easy to verify that this decision rule results in the student to invest all his effort to action a_1 (i.e., $e_1 = 1$), and leads to the teacher's utility of p. Now suppose the student can take actions unknown to the teacher. Our Lemma 2 tells us we can find a linear one to improve this nonlinear decision rule. Specifically, upon defining the convex sets $\mathcal{S}$ and $\mathcal{T}$ for f, we can find one hyperplane $f'(\mathbf{x}) = x_1 + x_2$ that separates these two sets. We illustrate this in Fig. 3b. Furthermore, we compute the worst-case utility for f' and f to see if f' is indeed better than f in terms of worst case performance. For f', by Eqn. (4) in Lemma 1, $V_d(f') = \min_{F \in \Delta(\mathcal{X})} \mathbb{E}_F[h(\mathbf{x})] = \min_{F \in \Delta(\mathcal{X})} \mathbb{E}_F[x_1] + \beta$, where F satisfies $\mathbb{E}_F[f'(\mathbf{x})] = \mathbb{E}_F[x_1 + x_2] \geq V_a(f'|\mathcal{A}_d)$. Observe that, when the student's available action set $\mathcal{A}_d$ is depicted as in Fig. 3a, $V_a(f'|\mathcal{A}_d) = 2p$. Since when F attains the minimum of $V_d(f')$,

[3] We assume the cost per unit effort is small enough so that the student is incentivized to exhaust his efforts. Another modeling choice is that the student will incur a fixed cost per unit effort with no budget. In fact, the two models are equivalent generally (Kleinberg and Raghavan, 2019).

[4] β_h can be used to guarantee the decision rules are eligible.

the inequality must bind. As a result, we will have $\min_{F\Delta(\mathcal{X})} \mathbb{E}_F[x_1] = 2p-1$, which gives us $V_d(f') = 2p - 1 + \beta_h$. However, follow the same analysis, one can compute that $V_d(f) = \beta_h$, which is smaller than $2p - 1 + \beta_h$.



(a)

(b)

Figure 3: (a): $p \in (0.5, 1)$ is the weight parameter. (b): Construct $\mathcal{S}$ and $\mathcal{T}$ when $f(\mathbf{x}) = \max\{x_1, x_2\}$. The gray shaded region is the set $\mathcal{T}$ (we actually plot the convex hull of all points $(\mathbf{x}, t(\mathbf{x}))$), while the light sky blue is the set $\mathcal{S}$. The color hyperplane is exactly $f'(\mathbf{x}) = x_1 + x_2$.

Moreover, f' does outperform f for our example with the student being able to take one action unknown to the teacher, as introduced in Fig. 1b. Suppose the student has one more action a_0 available to accomplish his course responsibilities (depicted in dashed lines in Fig. 4). The teacher is not informed by this change and may only be aware of the original student's available actions (which is $\{a_1, a_2\}$) and has to design her decision rule based on this restricted knowledge (see $\mathcal{A}_d$ and $\mathcal{A}_a$ in Table 1) [5]. Facing this uncertainty, it is easy to see that the linear one f' can guarantee teacher's maximal utility p, while f can only ensure a utility of $p - \epsilon$ to the teacher (since in this case, the student will deviate to invest all effort to action a_0), which is smaller than p.

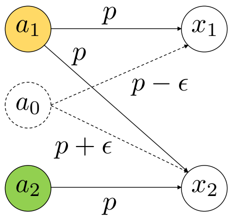


Figure 4: $\epsilon \in (0, 1-p)$ is the weight parameter.

4 The Complexity for Computing Robust Optimal Decision Rule

Having shown that a robust optimal decision rule f^* is linear, one may wonder whether it is possible to efficiently compute such f^*. Note that our analysis for robust optimality is constructive, and it establishes an algorithmic procedure to compute the optimal f^*. However, we show that computing f^* is generally hard.

Theorem 2. *We state the computation complexity for computing f^*:*

1. *Computing the linear f^* is at least as hard as solving the corresponding strategic decision making problem without robustness concern (under the linear decision space $\mathcal{G}^{lin}$).*
2. *In general, computing f^* is NP-hard since its corresponding strategic decision making problem without robustness concern (under the linear decision space $\mathcal{G}^{lin}$) is generally NP-hard.*

[5] See our Appendix D.

3. *When $\mathcal{X}$ is finite, if there is a polynomial-time algorithm for solving the corresponding strategic decision making problem without robustness concern (under the linear decision space $\mathcal{G}^{lin}$), then there is a polynomial-time algorithm for computing f^*.*

The proof and the description of a procedure for computing f^* are included in Appendix E. The key idea is to first show that we can rephrase the problem of computing f^* as an optimization problem. We then show that it can be further decomposed into two optimization problems, with one to be the same as solving (non-robust) optimal decision rule with strategic behavior (under the linear decision space $\mathcal{G}^{\text{lin}}$), and the other being a linear program with equality constraint.

More formally, let a linear decision rule be in the form of $f_{(\alpha,\beta)} = \alpha^\top \mathbf{x} + \beta$, where $(\alpha, \beta) \in \mathcal{G}^{\text{lin}}$ (see Definition 1), we use `Strategic-opt` to denote the corresponding strategic decision making problem (under linear decision space $\mathcal{G}^{\text{lin}}$) where the agent's available action set is exactly $\mathcal{A}_d$ (matching the knowledge of the decision maker):

$$\arg\max_{(\alpha,\beta)\in\mathcal{G}^{\text{lin}}} \mathbb{E}_F[h(\mathbf{x})], \quad \text{s.t. } (F, c) \in \arg\max_{(F,c)\in\mathcal{A}_d} \mathbb{E}_F[f_{(\alpha,\beta)}(\mathbf{x})] - c. \tag{\texttt{Strategic-opt}}$$

Let $(F_0, c_0) \in \mathcal{A}_d$ be the solution to the constraint in `Strategic-opt`. Then according to Lemma 1, we can compute f^* by solving:

$$\arg\max_{(\alpha,\beta)\in\mathcal{G}^{\text{lin}}} \min_{F\in\mathcal{E}} \mathbb{E}_F[h(\mathbf{x})], \tag{\texttt{Robust-strategic-opt}}$$

$$\text{s.t. } \mathcal{E} = \left\{F' : \mathbb{E}_{F'}[f_{(\alpha,\beta)}(\mathbf{x})] = \mathbb{E}_{F_0}[f_{(\alpha,\beta)}(\mathbf{x})] - c_0, F' \in \Delta(\mathcal{X})\right\}. \tag{7}$$

Note that different from the problem in `Strategic-opt`, after identifying the agent's best response $(F_0, c_0) \in \mathcal{A}_d$ under $f_{(\alpha,\beta)}$, our problem in `Robust-strategic-opt` will have an inside layer of optimization over the set $\mathcal{E}$. It is easy to see that this is a linear programming with equality constraint, where the decision variables are a probability simplex over $\mathcal{X}$.

$$\min_{F\in\mathcal{E}} \mathbb{E}_F[h(\mathbf{x})], \quad \text{s.t. } \mathcal{E} = \Big\{F' : \alpha^\top \mathbb{E}_{F'}[\mathbf{x}] = \alpha^\top \mathbb{E}_{F_0}[\mathbf{x}] - c_0, F' \in \Delta(\mathcal{X})\Big\}. \tag{$\mathcal{E}$-\texttt{LP}}$$

Therefore, the computation of `Robust-strategic-opt` can be decomposed into the computation of `Strategic-opt` and a linear program. This decomposition enables us to complete the proof.

5 Discussions and Future Work

Linear models, one of the "white-box" predictive models (contrary to the black-box models such as neural networks), have several desired properties such as nice generalizability, interpretability, transparency, and right to recourse. In this work, we further show that it is *robust* to unknown strategic manipulations when being used for making decisions. This is another dimension that is worth taking into account when deciding on which predictive models to deploy. While we also demonstrate that finding the robust optimal decision rule is generally hard, our analysis in decomposing the problem could provide some directions in figuring out efficient solvers in special cases.

There are still a number of open questions. In particular, our robustness notion could be overly pessimistic, considering the worst-case scenario over all possible unknown actions. One natural future direction is to explore Bayesian approaches, i.e., incorporating prior beliefs over all possible agent's action sets, to model and quantify these uncertainties. Secondly, our work has focused on dealing with a single agent (or more broadly, a set of homogeneous agents: The decision-maker knows the *common* subset of all agents' available actions). It would be interesting to extend the discussion to heterogeneous agents or a distribution of agents.

References

Tal Alon, Magdalen Dobson, Ariel D Procaccia, Inbal Talgam-Cohen, and Jamie Tucker-Foltz. Multiagent evaluation mechanisms. 2020.

Pablo Azar, Silvio Micali, Constantinos Daskalakis, and S Matthew Weinberg. Optimal and efficient parametric auctions. In *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*, pages 596–604. SIAM, 2013.

Moshe Babaioff, Michal Feldman, and Noam Nisan. Combinatorial agency. In *Proceedings of the 7th ACM conference on Electronic commerce*, pages 18–28, 2006.

Moshe Babaioff, Michal Feldman, and Noam Nisan. Mixed strategies in combinatorial agency. *Journal of Artificial Intelligence Research*, 38:339–369, 2010.

Ian Ball. Scoring strategic agents. 2020.

Chaithanya Bandi and Dimitris Bertsimas. Optimal design for multi-item auctions: a robust optimization approach. *Mathematics of Operations Research*, 39(4):1012–1038, 2014.

Michael Brückner and Tobias Scheffer. Stackelberg games for adversarial prediction problems. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 547–555, 2011.

Michael Brückner, Christian Kanzow, and Tobias Scheffer. Static prediction games for adversarial learning problems. *Journal of Machine Learning Research*, 13(Sep):2617–2654, 2012.

Gabriel Carroll. Robustness and linear contracts. *American Economic Review*, 105(2):536–63, 2015.

Gabriel Carroll. Robustness and separation in multidimensional screening. *Econometrica*, 85(2): 453–488, 2017.

Gabriel Carroll and Ilya Segal. Robustly optimal auctions with unknown resale opportunities. *The Review of Economic Studies*, 86(4):1527–1555, 2019.

Sylvain Chassang. Calibrated incentive contracts. *Econometrica*, 81(5):1935–1971, 2013.

Tianjiao Dai and Juuso Toikka. Robust incentives for teams. *Unpublished manuscript, Mass. Inst. of Technology, Cambridge, MA*, 2017.

Peter Diamond. Managerial incentives: on the near linearity of optimal compensation. *Journal of Political Economy*, 106(5):931–957, 1998.

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

Paul Dütting, Tim Roughgarden, and Inbal Talgam-Cohen. Simple versus optimal contracts. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 369–387, 2019.

Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Analysis of classifiers' robustness to adversarial perturbations. *Machine Learning*, 107(3):481–508, 2018.

Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57, 2017.

Sanford J Grossman and Oliver D Hart. An analysis of the principal-agent problem. In *Foundations of Insurance Economics*, pages 302–340. Springer, 1992.

Nika Haghtalab, Nicole Immorlica, Brendan Lucier, and Jack Wang. Maximizing welfare with incentive-aware evaluation mechanisms. Technical report, working paper, 2020.

Lars Peter Hansen and Thomas J Sargent. Three types of ambiguity. *Journal of Monetary Economics*, 59(5):422–445, 2012.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pages 111–122, 2016.

Chien-Ju Ho, Aleksandrs Slivkins, and Jennifer Wortman Vaughan. Adaptive contract design for crowdsourcing markets: Bandit algorithms for repeated principal-agent problems. *Journal of Artificial Intelligence Research*, 55:317–359, 2016.

Bengt Holmstrom and Paul Milgrom. Aggregation and linearity in the provision of intertemporal incentives. *Econometrica: Journal of the Econometric Society*, pages 303–328, 1987.

Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. The disparate effects of strategic manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 259–268, 2019.

Jon Kleinberg and Manish Raghavan. How do classifiers induce agents to invest effort strategically? In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 825–844, 2019.

Jianjun Miao and Alejandro Rivera. Robust contracts in continuous time. *Econometrica*, 84(4): 1405–1440, 2016.

John Miller, Smitha Milli, and Moritz Hardt. Strategic adaptation to classifiers: A causal perspective. *arXiv preprint arXiv:1910.10362*, 2019.

Smitha Milli, John Miller, Anca D Dragan, and Moritz Hardt. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 230–239, 2019.

Steven Shavell. Risk sharing and incentives in the principal and agent relationship. *The Bell Journal of Economics*, pages 55–73, 1979.

Behzad Tabibian, Stratis Tsirtsis, Moein Khajehnejad, Adish Singla, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. Optimal decision making under strategic behavior. *arXiv preprint arXiv:1905.09239*, 2019.

Berk Ustun and Cynthia Rudin. Methods and models for interpretable linear classification. *arXiv preprint arXiv:1405.4047*, 2014.

Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 10–19, 2019.

Jiaxuan Wang, Jeeheh Oh, Haozhu Wang, and Jenna Wiens. Learning credible models. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2417–2426, 2018.

A Proof for Lemma 1

Proof. Let us first fix an arbitrary action set $\mathcal{A}_a \supseteq \mathcal{A}_d$, and an eligible decision rule f, then by (6), we have that the agent's utility is at least $V_a(f|\mathcal{A}_d)$, that is, any action (F, c) the agent would chose under the decision rule f must satisfy:

$$\mathbb{E}_F[f(\mathbf{x})] \geq \mathbb{E}_F[f(\mathbf{x})] - c = V_a(f|\mathcal{A}_a) \geq V_a(f|\mathcal{A}_d).$$

Thus, the decision maker's utility $V_d(f|\mathcal{A}_a) = \mathbb{E}_F[h(\mathbf{x})]$ is at the least the minimum given by the (4). This implies the following guarantee of worst-case utility $V_d(f)$:

$$V_d(f) \geq \min_{F \in \Delta(\mathcal{X})} \mathbb{E}_F[h(\mathbf{x})] \quad \text{s.t.} \quad \mathbb{E}_F[f(\mathbf{x})] \geq V_a(f|\mathcal{A}_d). \tag{8}$$

We now show that (8) is tight. Let $\mathtt{supp}(F)$ denote the support of distribution F. Let F_0 be a distribution attaining the minimum in (4) and also satisfying the constraint. We consider following two cases:

Case 1: $\mathtt{supp}(F_0) \not\subset \arg\max_{\mathbf{x}} f(\mathbf{x})$. Then let F_1 be a distribution which achieves a higher value of $\mathbb{E}_F[f(\mathbf{x})]$. Let F' be a mixture distribution $F' = (1-\epsilon)F_0 + \epsilon F_1$, with a small positive ϵ. Then we have $\mathbb{E}_{F'}[f(\mathbf{x})] = (1-\epsilon)\mathbb{E}_{F_0}[f(\mathbf{x})] + \epsilon\mathbb{E}_{F_1}[f(\mathbf{x})] > \mathbb{E}_{F_0}[f(\mathbf{x})]$. Now take $\mathcal{A}'_a = \mathcal{A}_d \cup \{(F', 0)\}$, then the agent's unique optimal action under $\mathcal{A}'_a$ is $(F', 0)$. This brings the decision maker with utility of $V_d(f|\mathcal{A}'_a) = (1-\epsilon)\mathbb{E}_{F_0}[h(\mathbf{x})] + \epsilon\mathbb{E}_{F_1}[h(\mathbf{x})]$. Since $V_d(f|\mathcal{A}'_a) \geq V_d(f)$, we further have

$$V_d(f) \leq V_d(f|\mathcal{A}'_a) = (1-\epsilon)\mathbb{E}_{F_0}[h(\mathbf{x})] + \epsilon\mathbb{E}_{F_1}[h(\mathbf{x})]. \tag{9}$$

When $\epsilon \to 0$, the RHS in (9) will converge to $\mathbb{E}_{F_0}[h(\mathbf{x})]$ from above. This implies $V_d(f) \leq \mathbb{E}_{F_0}[h(\mathbf{x})]$ when $\epsilon \to 0$. Recall our definition of F_0, and together with the lower bound we have shown for $V_d(f)$ in (8), we can conclude our results in (4) for this case.

Case 2: $\mathtt{supp}(F_0) \subset \arg\max_{\mathbf{x}} f(\mathbf{x})$. For this case, we discuss following two situations.
(i): $\mathbb{E}_{F_0}[f(\mathbf{x})] > V_a(f|\mathcal{A}_d)$, we now consider action set $\mathcal{A}'_a = \mathcal{A}_d \cup \{(F_0, 0)\}$. Since $\mathbb{E}_{F_0}[f(\mathbf{x})] > V_a(f|\mathcal{A}_d)$, then the agent will uniquely chose action $(F_0, 0)$ for f under the action set $\mathcal{A}'_a$. This brings the decision maker with the utility of $V_d(f|\mathcal{A}'_a) = \mathbb{E}_{F_0}[h(\mathbf{x})]$. Again, with the fact that $V_d(f|\mathcal{A}'_a) \geq V_d(f)$ and the definition of F_0, we have now proved (4).
(ii): $\mathbb{E}_{F_0}[f(\mathbf{x})] = V_a(f|\mathcal{A}_d) = \max f(\mathbf{x})$, this situation can only be satisfied when $\mathcal{A}_d$ contains some action of the form $(F', 0)$ with $\mathtt{supp}(F') \subset \arg\max f(\mathbf{x})$. Thus, we define

$$\mathcal{G} := \left\{(F', 0) \in \mathcal{A}_d : \mathtt{supp}(F') \subset \arg\max f(\mathbf{x})\right\} \neq \emptyset.$$

Then, under action set $\mathcal{A}_d$, the agent will choose an action in $\mathcal{G}$ which would benefit decision maker (since we have assumed that when there are multiple optimal actions for agent, agent will choose the one which maximizes decision maker's utility.), leading the decision maker's utility $V_d(f|\mathcal{A}_d) = \max_{(F,0)\in\mathcal{G}} \mathbb{E}_F[h(\mathbf{x})] \geq V_d(f)$. But then the agent would have been willing to choose the same action under the zero decision rule (and any $\mathcal{A} \supseteq \mathcal{A}_d$), leading the decision maker's utility $V_d(0|\mathcal{A}) = \max_{(F,0)\in\mathcal{G}} \mathbb{E}_F[h(\mathbf{x})] = V_d(0)$. This means $V_d(0) \geq V_d(f)$, which contradicts our eligibility assumption.

Without loss of generality, we may assume the agent has a costless action $(\delta_{\underline{\mathbf{x}}}, 0)$ in $\mathcal{A}_d$.[6] Now let $F_0 \in \Delta(\mathcal{X})$ attain the minimum in (4). We have $\mathbb{E}_{F_0}[h(\mathbf{x})] = V_d(f) > 0$ by eligibility. On the other hand, we have $h(\underline{\mathbf{x}}) = 0$. If we have $\mathbb{E}_{F_0}[f(\mathbf{x})] > V_a(f|\mathcal{A}_d)$ strictly, then replace F_0 by a mixture distribution $F' = (1-\epsilon)F_0 + \epsilon\delta_{\underline{\mathbf{x}}}$ for small ϵ. Consider $\mathcal{A}'_a = \mathcal{A}_d \cup \{(F', 0)\}$, then the agent's utility by taking the action $(F', 0)$ is given by $V_a(f|\mathcal{A}'_a) = (1-\epsilon)\mathbb{E}_{F_0}[f(\mathbf{x})] + \epsilon f(\underline{\mathbf{x}})$, then one can find a small ϵ such that $V_a(f|\mathcal{A}'_a)$ is strictly larger than $V_a(f|\mathcal{A}_d)$. As a result, this brings the policymaker with a utility of $V_d(f|\mathcal{A}'_a) = (1-\epsilon)\mathbb{E}_{F_0}[h(\mathbf{x})] + \epsilon h(\underline{\mathbf{x}}) = (1-\epsilon)\mathbb{E}_{F_0}[h(\mathbf{x})]$. Since $V_d(f|\mathcal{A}'_a) \geq V_d(f)$, given any positive ϵ, this implies that $V_d(f) \leq (1-\epsilon)\mathbb{E}_{F_0}[h(\mathbf{x})] < \mathbb{E}_{F_0}[h(\mathbf{x})]$, which contradicts the minimality of F_0. Hence we have equality, $\mathbb{E}_{F_0}[f(\mathbf{x})] = V_a(f|\mathcal{A}_d)$, as claimed.

Finally, if $F_0 \in \arg\max_{F\in\Delta(\mathcal{X})} \mathbb{E}_F[f(\mathbf{x})]$, and $\mathbb{E}_{F_0}[f(\mathbf{x})] = V_a(f|\mathcal{A}_d)$, then we have (5). □

After finishing the proof, we would like to give following explanation on our construction of worst-case action set in the proof.

Remark 1. The above proof relies on a construction of agent's worst case action set by adding an arbitrary action of the form $(F, 0)$. It may seem unrealistic to allow the agent to arbitrarily manipulate himself at zero cost. However, we note that the zero cost is not a substantive assumption; the logic can be carried over to more detailed models that explicitly restrict the effort costs as a function of expected manipulated feature. Then the equivalent step consists of adding an action to the action set that produces F at the lowest allowable cost.

B Proof for Lemma 2

Proof. We may assume that the convex hull of $\mathcal{X}$ is a full-dimensional set in $\mathbb{R}^n$. Now fix any nonlinear decision rule f, our proof will hinge on the discussion of two cases we have shown in Lemma 1.

Case 1. We first define

$$t(\mathbf{x}) = \max\{V_a(f|\mathcal{A}_d), h(\mathbf{x}) + f(\mathbf{x}) - V_d(f)\}.$$

Now we define two sets in $\mathbb{R}^{n+1} = \mathbb{R}^n \times \mathbb{R}$: Let $\mathcal{S}$ be the convex hull of all pairs $(\mathbf{x}, f(\mathbf{x}))$, for $\mathbf{x} \in \mathcal{X}$, let $\mathcal{T}$ be the convex hull of all pairs $(\mathbf{x}, z)$ that $\mathbf{x}$ lies in the convex hull of $\mathcal{X}$, and $z > t(\mathbf{x})$. We note that $\mathcal{T}$ is then a convex set.

We now claim that $\mathcal{S}$ and $\mathcal{T}$ are disjoint. To see this, let's suppose $\mathcal{S}$ and $\mathcal{T}$ are not disjoint, then there exists a distribution $F \in \Delta(\mathcal{X})$ such that $\mathbb{E}_F[f(\mathbf{x})] > \mathbb{E}_F[t(\mathbf{x})]$. In particular, we have

$$\mathbb{E}_F[f(\mathbf{x})] > V_a(f|\mathcal{A}_d),$$

and also

$$\begin{aligned}&\mathbb{E}_F[f(\mathbf{x})] > \mathbb{E}_F[h(\mathbf{x})] + \mathbb{E}_F[f(\mathbf{x})] - V_d(f)\\ &\Rightarrow V_d(f) > \mathbb{E}_F[h(\mathbf{x})].\end{aligned}$$

[6] This assumption is a purely additive normalization of the decision maker's utility and it can be relaxed (see our discussion at the end of the section B). Our results will not require this assumption. Other works also make similar assumption (Carroll, 2015; Dütting et al., 2019): The agent can always exert no effort, namely, the zero-cost action, to produce a minimum output (denote by 0); this corresponds to assuming $(\delta_0, 0) \in \mathcal{A}_d$.

This is a direct contradiction to our statement of (4) in Lemma 1.

The disjointness and convexity of $\mathcal{S}$ and $\mathcal{T}$ enable us to apply the separating hyperplane theorem: There exists a vector $\lambda = (\lambda_1, \dots, \lambda_n)$ and constants μ, v such that

$$\sum_i \lambda_i x_i + \mu z \le v, \quad \forall (\mathbf{x}, z) \in \mathcal{S} \tag{10}$$

$$\sum_i \lambda_i x_i + \mu z \ge v, \quad \forall (\mathbf{x}, z) \in \mathcal{T} \tag{11}$$

and λ is a non-zero vector. Note that (10) and (11) implies $\mu \ge 0$. To see this, fix a point $\mathbf{x} \in \mathcal{X}$, then for $(\mathbf{x}, z) \in \mathcal{S}$ and $(\mathbf{x}, z') \in \mathcal{T}$ we have

$$\sum_i \lambda_i x_i + \mu z' \ge \sum_i \lambda_i x_i + \mu z \Rightarrow \mu z' \ge \mu z,$$

by earlier argument on the disjointness of $\mathcal{S}$ and $\mathcal{T}$, we can conclude that $\mu \ge 0$. We now also show that μ is a positive constant. Suppose $\mu = 0$, then (10) gives $\sum_i \lambda_i x_i \le v$ and (11) gives $\sum_i \lambda_i x_i \ge v$, which leads to $\sum_i \lambda_i x_i = v$. Since not all λ_i are zero, this contradicts the full-dimensionality of $\mathcal{X}$.

Now we can rewrite (10) as following

$$f(\mathbf{x}) \le \frac{v - \sum_i \lambda_i x_i}{\mu}, \quad \forall \mathbf{x} \in \mathcal{X}$$

This motivates us to define following linear decision rule

$$f'(\mathbf{x}) = \frac{v - \sum_i \lambda_i x_i}{\mu}, \quad \forall \mathbf{x} \in \mathcal{X}. \tag{12}$$

Note that we have $f'(\mathbf{x}) \ge f(\mathbf{x})$ pointwise.

Now we are ready to check that $V_d(f') \ge V_d(f)$. Let (F_0, c_0) be the action that the agent would like to choose under f and action set $\mathcal{A}_d$. Consider any action set $\mathcal{A}_a \supseteq \mathcal{A}_d$, as we have shown before, we must have

$$V_a(f'|\mathcal{A}_a) \ge V_a(f'|\mathcal{A}_d) \ge V_a(f|\mathcal{A}_d). \tag{13}$$

Let (F, c) be the action that the agent chooses under f' and action set $\mathcal{A}_a$. Then (11) implies

$$\begin{aligned} \mathbb{E}_F[t(\mathbf{x})] &\ge \frac{v - \sum_i \lambda_i \mathbb{E}_F[x_i]}{\mu} \\ &= \mathbb{E}_F[f'(\mathbf{x})] && (14) \\ &= V_a(f'|\mathcal{A}_a) + c \\ &\ge V_a(f'|\mathcal{A}_a) && (c \in \mathbb{R}_+) \\ &\ge V_a(f|\mathcal{A}_d). && \text{(by (13))} \end{aligned}$$

It is worthy noting that if above inequality is strict, then according to our definition of $t(\mathbf{x})$, we must have

$$\mathbb{E}_F[t(\mathbf{x})] = \mathbb{E}_F[h(\mathbf{x})] + \mathbb{E}_F[f(\mathbf{x})] - V_d(f). \tag{15}$$

So we have

$$\begin{aligned} V_d(f'|\mathcal{A}_a) = \mathbb{E}_F[h(\mathbf{x})] &= \mathbb{E}_F[t(\mathbf{x})] - \mathbb{E}_F[f(\mathbf{x})] + V_d(f) \\ &\geq \mathbb{E}_F[t(\mathbf{x})] - \mathbb{E}_F[f'(\mathbf{x})] + V_d(f) && \text{(by definition of } f'\text{)} \\ &\geq V_d(f). && \text{(by 14)} \end{aligned}$$

On the other hand, if $\mathbb{E}_F[t(\mathbf{x})] = V_a(f|\mathcal{A}_d)$. This implies all the inequalities in the stacked chain above are equalities. In particular, we will have

$$V_a(f'|\mathcal{A}_a) = V_a(f'|\mathcal{A}_d) = V_a(f|\mathcal{A}_d).$$

Since the agent now does at least as well as $V_a(f|\mathcal{A}_d)$ by taking action (F_0, c_0), this action is in his choice set under f' and $\mathcal{A}_a$, as a result, the decision maker gets at least the corresponding utility: $V_d(f'|\mathcal{A}_a) \geq \mathbb{E}_{F_0}[h(\mathbf{x})] = V_d(f|\mathcal{A}_d) \geq V_d(f)$, where the first inequality is due to the tie-breaking assumption of the agent (when there are multiple maximizers, the agent will chose the most beneficial one for the decision maker).

Thus, in either case, we have $V_d(f'|\mathcal{A}_a) \geq V_d(f)$, this holds for any $\mathcal{A}_a \supseteq \mathcal{A}_d$, thus we have $V_d(f') \geq V_d(f)$.

Case 2. In this case, we define $\mathcal{S}$ to be the convex hull of all pairs $(\mathbf{x}, f(\mathbf{x}))$, and $\mathcal{T}$ to be the set of all $(\mathbf{x}, z)$ with $\mathbf{x}$ in the convex hull of $\mathcal{X}$ and $z > V_a(f|\mathcal{A}_d)$. We still claim both of $\mathcal{S}$ and $\mathcal{T}$ are convex, and disjoint: otherwise, there exists F such that

$$\mathbb{E}_F[f(\mathbf{x})] > V_a(f|\mathcal{A}_d).$$

This contradicts our statement (5) in Lemma 1. Using the same arguments as in case 1, we find a vector $\lambda = (\lambda_1, \ldots, \lambda_n)$ and constants μ, v such that (10) and (11) hold, and we can still guarantee that $\mu > 0$. Again, we define a linear decision rule f' by (12); from (10) we know that $f' \geq f$ pointwise. Consider the agent's behavior under decision rule f', for any action (F, c) chosen by the agent under any possible action set, we have

$$\mathbb{E}_F[f'(\mathbf{x})] - c = f'(\mathbb{E}_F[\mathbf{x}]) - c \leq V_a(f|\mathcal{A}_d). \qquad \text{(by (11))}$$

This means that the agent cannot earn a higher expected utility than $V_a(f|\mathcal{A}_d)$. On the other hand, the agent can always earn at least this much, since $V_a(f'|\mathcal{A}_a) \geq V_a(f'|\mathcal{A}_d) \geq V_a(f|\mathcal{A}_d)$. This means we have equality $V_a(f'|\mathcal{A}_a) = V_a(f'|\mathcal{A}_d) = V_a(f|\mathcal{A}_d)$. From here, the argument finishes just as at the end of case 1, and we have $V_d(f') \geq V_d(f)$. □

Extensions: General cost lower bounds As mentioned in Remark 1, our analysis relies on the construction of worst case action sets, using actions, that produce an undesirable distribution F, at costs of zero. This zero-cost action assumption (together with the assumption in Footnote 6) is not substantial and one natural relaxation is that the decision maker knows a lower bound on the cost of any available actions, or of producing any given level of expected output. Our analysis and results will go through for this scenario. Specifically, suppose the known lower bound cost is denoted by $\underline{c} > 0$, then our Lemma 1 can be accordingly changed to: $V_d(f) = \min_{F \in \Delta(\mathcal{X})} \mathbb{E}_F[h(\mathbf{x})]$, s.t. $\mathbb{E}_F[f(\mathbf{x})] - \underline{c} \geq V_a(f|\mathcal{A}_d)$ or $\max_{F \in \Delta(\mathcal{X})} \mathbb{E}_F[f(\mathbf{x})] - \underline{c} = V_a(f|\mathcal{A}_d)$. To get the analogous result in Lemma 2, one can change the function $t(\mathbf{x})$ as $t(\mathbf{x}) = \max\{V_a(f|\mathcal{A}_d) + \underline{c}, h(\mathbf{x}) + f(\mathbf{x}) - V_d(f)\}$, then all the analysis can be carried over here.

C Proof for Lemma 3

Proof. We prove Theorem 1 via showing the existence of an optimum within the class of linear decision rules, and this decision rule will then be optimal among all decision rules. Note that for any eligible decision rule $f(\mathbf{x})$, the value of $f(\mathbf{x})$ that it assigns to $\mathbf{x}$ is bounded within $[0, \bar{C}]$. Let a linear decision rule be the form of $f_{(\alpha,\beta)}(\mathbf{x}) = \alpha^\top \mathbf{x} + \beta$. Then it suffices to show that the guaranteed worst-case utility $V_d(f)$ is an upper semi-continuous function of $(\alpha, \beta) \in \mathcal{G}^{\text{lin}}$. Now fix a sequence $(\alpha^1, \beta^1), (\alpha^2, \beta^2), \ldots$ in $\mathcal{G}^{\text{lin}}$ converging to some $(\alpha^\infty, \beta^\infty)$ in $\mathcal{G}^{\text{lin}}$. Then it suffices to show that $V_d(f_{(\alpha^\infty,\beta^\infty)}) \geq \limsup_k V_d(f_{(\alpha^k,\beta^k)})$. To prove this, first note that by replacing the sequence $((\alpha^k, \beta^k))$ with a subsequence along which $V_d(f((\alpha^k, \beta^k)))$ converges to its lim sup on the original sequence, thus, we can assume that $V_d(f_{(\alpha^k,\beta^k)})$ converges to $\limsup_k V_d(f_{(\alpha^k,\beta^k)})$. Now for any action set $\mathcal{A}_a$, and let (F^k, c^k) be the agent's chosen action under $\mathcal{A}_a$ and the decision rule $f_{(\alpha^k,\beta^k)}$. Then if necessary, by extracting a further subsequence, we can assume that the sequence (F^k, c^k) converges to some $(F^\infty, c^\infty) \in \mathcal{A}_a$. Since the agents' utility are continuous in (α, β), then (F^∞, c^∞) is an optimal action for the agent under $f_{(\alpha^\infty,\beta^\infty)}$, and its utility to the decision maker is the limit of the corresponding utility of (F^k, c^k) under $f_{(\alpha^k,\beta^k)}$. We thus have

$$V_d(f_{(\alpha^\infty,\beta^\infty)}|\mathcal{A}_a) \geq \mathbb{E}_{F^\infty}[h(\mathbf{x})] = \lim_k \mathbb{E}_{F^k}[h(\mathbf{x})] = \lim_k V_d(f_{(\alpha^k,\beta^k)}|\mathcal{A}_a) \geq \lim_k V_d(f_{(\alpha^k,\beta^k)}).$$

Since $\mathcal{A}_a \supseteq \mathcal{A}_d$ is arbitrary, then we have $V_d(f_{(\alpha^\infty,\beta^\infty)}) \geq \lim_k V_d(f_{(\alpha^k,\beta^k)})$. □

D Missing Table in Section 3.4

Given the student's efforts $\mathbf{e}$ invested to each action, we can enumerate all possible induced distributions over $\mathcal{X}$ in $\mathcal{A}_d$ and $\mathcal{A}_a$ (see Table 1). Note that since the student can now also invest efforts to action a_0, $\mathcal{A}_a$ contains more availabilities compared to $\mathcal{A}_d$.

$\mathbf{x} = (x_1, x_2)$	F in $\mathcal{A}_d$	F in $\mathcal{A}_a$
$\mathbb{P}(\mathbf{x} = (1,1))$	$e_1 p^2$	$(e_1 p + (p - \epsilon)e_0)(p + \epsilon e_0)$
$\mathbb{P}(\mathbf{x} = (1,0))$	$e_1 p(1-p)$	$(e_1 p + (p - \epsilon)e_0)(1 - p - \epsilon e_0)$
$\mathbb{P}(\mathbf{x} = (0,1))$	$(1 - e_1 p)p$	$(1 - e_1 p - (p - \epsilon)e_0)(p + \epsilon e_0)$
$\mathbb{P}(\mathbf{x} = (0,0))$	$(1 - e_1 p)(1 - p)$	$(1 - e_1 p - (p - \epsilon)e_0)(1 - p - \epsilon e_0)$

Table 1: All possible distributions F $\mathcal{A}_d$ and $\mathcal{A}_a$ induced by student's effort $\mathbf{e} = (e_0, e_1, 1 - e_0 - e_1)$. e_1, e_0 are the efforts decided by the student for actions a_1 and a_0 where $e_1 + e_0 \in [0, 1]$.

E Missing proof and the Algorithm for Theorem 2

Proof. According to case (i) in Lemma 1, given $f_{(\alpha,\beta)}$, for any distribution F attaining the minimum in (4), we know that the inequality in Γ must bind at F. Let $(F_0, c_0) \in \mathcal{A}_d$ be the solution to the

Algorithm 1 Find the optimal robust decision rule

```
Input: Decision maker's knowledge $\mathcal{A}_d$, linear decision space $\mathcal{G}^{\text{lin}}$, objective function $h$.
Initial $f^* \in \mathcal{F}^{\text{lin}}$ arbitrarily and $V_d(f^*) = 0$.
for every $(\alpha, \beta) \in \mathcal{G}^{\text{lin}}$ do
    Let $(F_0, c_0) \in \arg\max_{(F,c)\in\mathcal{A}_d} \mathbb{E}_F\left[\alpha^\top \mathbf{x} + \beta\right] - c$;
    Solve the set $\mathcal{E} = \left\{F : \alpha^\top \left(\mathbb{E}_{F_0}[\mathbf{x}] - \mathbb{E}_F[\mathbf{x}]\right) = c_0, F \in \Delta(\mathcal{X})\right\}$;
    Compute $V_d\left(f_{(\alpha,\beta)}\right) = \min_{F\in\mathcal{E}} \mathbb{E}_F[h(\mathbf{x})]$;
    if $V_d\left(f_{(\alpha,\beta)}\right) > V_d(f^*)$ then
        $f^* \leftarrow \alpha^\top \mathbf{x} + \beta$.
    end if
end for
Output Robust optimal decision: $f^*$.
```

constraint in `Strategic-opt`. Then we can compute f^* by solving:

$$\arg\max_{(\alpha,\beta)\in\mathcal{G}^{\text{lin}}} \min_{F\in\mathcal{E}} \mathbb{E}_F[h(\mathbf{x})], \tag{Robust-strategic-opt}$$

$$\text{s.t. } \mathcal{E} = \left\{F' : \mathbb{E}_{F'}[f_{(\alpha,\beta)}(\mathbf{x})] = \mathbb{E}_{F_0}[f_{(\alpha,\beta)}(\mathbf{x})] - c_0, F' \in \Delta(\mathcal{X})\right\}, \tag{16}$$

where we refer to the set $\mathcal{E}$, as the *worst-action set*, since we choose the worst action among it to minimize the expected utility $\mathbb{E}_F[h(\mathbf{x})]$. Different from the problem in `Strategic-opt`, after identifying the agent's best response $(F_0, c_0) \in \mathcal{A}_d$ under $f_{(\alpha,\beta)}$, our problem in `Robust-strategic-opt` first turns to characterizing a worst-action set $\mathcal{E}$. Then the searching of f^* will hinge on maximizing $\mathbb{E}_F[h(\mathbf{x})]$ in each $\mathcal{E}$ over $\mathcal{G}^{\text{lin}}$. This implies that to make our problem tractable, one may first need to guarantee the corresponding strategic decision-making problem tractable. Furthermore, given a linear $f_{(\alpha,\beta)}$, the additional computational complexity in `Robust-strategic-opt` is due to the robustness concern in minimizing $\mathbb{E}_F[h(\mathbf{x})]$ over set $\mathcal{E}$. It is easy to see that this is a linear programming with equality constraint, where the decision variables are a probability simplex over $\mathcal{X}$.

$$\min_{F\in\mathcal{E}} \mathbb{E}_F[h(\mathbf{x})], \quad \text{s.t. } \mathcal{E} = \left\{F' : \alpha^\top \mathbb{E}_{F'}[\mathbf{x}] = \alpha^\top \mathbb{E}_{F_0}[\mathbf{x}] - c_0, F' \in \Delta(\mathcal{X})\right\}. \tag{$\mathcal{E}$-LP}$$

Inside the optimization, for every $(\alpha, \beta) \in \mathcal{G}^{\text{lin}}$, our problem `Robust-strategic-opt` has one more induced Linear programming $\mathcal{E}$-`LP` to solve compared with the standard problem `Strategic-opt`.

As it will in general be hard to optimize arbitrary non-concave functions, we may consider assuming a concave h. However, as pointed out by other studies (Kleinberg and Raghavan, 2019; Alon et al., 2020), there exist concave functions h that are NP-hard to solve the problem `Strategic-opt` (via a reduction from the maximum independent set problem), which naturally leads the hardness of our problem. In particular, back to our student evaluation setting, let $F(\mathbf{e})$ be the induced feature distribution if the agent's effort profile is $\mathbf{e}$. As a result, the decision maker's goal on maximizing $h(\mathbf{x})$ can be reduced to maximizing $h(F(\mathbf{e}))$. When $h(F(\mathbf{e})) = \|\mathbf{e}\|_0$, solving the problem `Strategic-opt` is then NP-hard. □