# Linear Models are Robust Optimal Under Strategic Behavior

**Wei Tang**[†], **Chien-Ju Ho**[†], **and Yang Liu**[*]
[†]Washington University in St. Louis, [*]UC Santa Cruz
{w.tang, chienju.ho}@wustl.edu, yangliu@ucsc.edu

## Abstract

There is an ubiquitous use of algorithms to inform decisions nowadays, 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 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 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 contributions are two-fold. First, we 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. Second, we explore the computational problem of searching for the robust optimal decision rule and demonstrate its connection to distributionally robust optimization. We believe our results promote the use of simple linear decisions with uncertain individual manipulations.

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. 

## 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 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 transparent to offer "right to explanation" (Goodman and Flaxman, 2017). 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 creates opportunities for individuals to strategically respond to 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 strategic behavior, there has been a recent flurry of work in studying 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

[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.

tractable, almost all the works explicitly assume the decision maker has the *full knowledge* of agents' action space and the corresponding costs for agents to manipulate their features. The above knowledge enables a game theoretic analysis that characterizes agents' best responses when offered a particular decision rule.

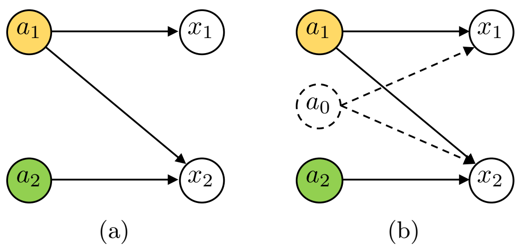



Figure 1: An instance of student evaluation problem.

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 either study ($a_1$) or copy 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, the teacher can design a decision rule (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** Our contributions are two-fold. First, we formalize the problem of robust strategic decision-making and characterize the robust optimal decision rules. We show that under mild conditions, for any robust optimal decision rule, there exists a linear one that is equally robust optimal. Our result implies that, to find robust optimal decision rules, it suffices to search over the space of linear decision rules. Second, we explore the computational problem of searching for the robust optimal $f^*$. While the problem is NP-hard in general (since non-robust strategic decision-making is only solvable in restricted settings but is generally NP-hard (Kleinberg and Raghavan, 2019)), we investigate the additional complexity introduced by our robustness desiderata, through adapting techniques from distributionally robust optimization (Delage and Ye, 2010). Our results inform efficient algorithms especially in settings when non-robust strategic decision-making problem is efficiently solvable.

## 1.1 Related Work

Our problem closely connects to the recent literature in machine learning in the presence of strategic manipulation (Hardt et al., 2016; Brückner et al., 2012; Brückner and Scheffer, 2011). 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 settings in which the costs for manipulation differ for different groups and explore the societal impacts. There are also works directly utilizing the decision rule as an incentive device to induce desired behavior (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 extend 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 studies the strategic interplay between two parties with misaligned interests. 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. Our work 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 by Dütting et al. (2019); Babaioff et al. (2006); Ho et al. (2016); Babaioff et al. (2010). Our work also 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. Our work promotes the usage of linear models: In addition to interpretability and good generalization, linear models are also robust to unknown strategic manipulation.

## 2 A Model of Robust Strategic Decision-making

In this section, we formalize our model for robust strategic decision-making. Agent features are represented by a vector $\mathbf{x} = (x_1, \dots, x_n)$, which takes value in a bounded compact set $\mathcal{X} \subseteq \mathbb{R}^n$. The agent can take actions to alter the features. An action of the agent can be represented by the outcome (i.e., the distribution of agent features after the action) and the cost of the action. We use a pair $(\mathbb{P}, c) \in \Delta(\mathcal{X}) \times \mathbb{R}_+$ to denote an action, where $\mathbb{P}$ is the outcome, i.e., the distribution of the agent features after action, and $c$ is the associated cost. 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. Informally, 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 $f : \mathcal{X} \to \mathbb{R}_{\geq 0}$ is a mapping from the agent's features to a decision, where the decision domain of $f$ is normalized to be non-negative and directly represents the value of the decision to the agent. 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 $h : \mathcal{X} \to \mathbb{R}_{\geq 0}$. This function characterizes the utility that the agent brings to the designer. For example, it could be a qualification function, assuming the decision maker aims to increase the chance that the agent passes the qualification, and the agent's effort in changing their features may lead to self-improvement, thus in their true qualifications. Assume that there's an upper bound $\bar{C} > 0$ of $f(\mathbf{x})$ for any $\mathbf{x} \in \mathcal{X}$. In addition, we 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., $f(\mathbf{x}) = \boldsymbol{\omega}^\top \mathbf{x} + \beta$ for $\boldsymbol{\omega} \in \mathbb{R}^n$ and $\beta \in \mathbb{R}$. Let $\mathcal{G}^{lin} = \{(\boldsymbol{\omega}, \beta) \in \mathbb{R}^n \times \mathbb{R} : f(\mathbf{x}) = \boldsymbol{\omega}^\top \mathbf{x} + \beta \in [0, \bar{C}], \forall \mathbf{x} \in \mathcal{X}\}$ be the space of parameter pair $(\boldsymbol{\omega}, \beta)$.*

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 $(\mathbb{P}, c) \in \mathcal{A}_a$ to respond to $f$; (3) the agent features are then moved to $\mathbf{x} \sim \mathbb{P}$; (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 $\mathbb{E}_{\mathbb{P}}[f(\mathbf{x})] - c$ for taking action $(\mathbb{P}, c)$. 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

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

corresponding utility:

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

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

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

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 $\mathcal{A}_d$:

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

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:

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

# 3 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.

**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 $f(\mathbf{x}) = \beta + \boldsymbol{\omega}^\top \mathbf{x}$, for some $\boldsymbol{\omega} \in \mathbb{R}^n, \beta \in \mathbb{R}$.*

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 following, we provide the proof sketch and use an example to demonstrate our results. Our result and analysis extend the work of robust contract design (Carroll, 2015) to deal with situations in which both the decision rule and the utility of the decision maker can take more general function forms (instead of restricting to one-dimensional contract as decision rule, and additive utility for decision maker). 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 Characterize the worst-case utility $V_d(f)$

Before we move to the main analysis, consider a trivial case that the decision maker chooses to post no decision rule (i.e., $f(\mathbf{x}) = 0, \forall \mathbf{x}$). Since this is also a linear decision rule, if the robust optimal decision rule is to post no decision rule, Theorem 1 is trivially correct. In the following discussion, we focus the discussion on the cases in which the decision maker can benefit from posting some decision rule (otherwise, she can choose to post no decision rules). In particular, we define rational decision rules as follows.

**Definition 2** (Rational decision rule)**.** *A decision rule $f$ is* rational *for the decision maker if $V_d(f) > V_d(0)$,where $V_d(0)$ represents the utility of the decision maker when she publishes no decision rules.*

We first characterize the worst-case utility guarantee for any given rational decision rule.

**Lemma 1.** *Let $f$ be any rational decision rule. Define a set $\Gamma = \{\mathbb{P} \in \Delta(\mathcal{X}) : \mathbb{E}_{\mathbb{P}}[f(\mathbf{x})] \geq V_a(f|\mathcal{A}_d)\}$. Then one of the following two cases occurs:*

$$(i) \quad V_d(f) = \min_{\mathbb{P}\in\Gamma} \mathbb{E}_{\mathbb{P}}[h(\mathbf{x})]; \quad (4)$$
$$or \quad (ii) \max_{\mathbb{P}\in\Delta(\mathcal{X})} \mathbb{E}_{\mathbb{P}}[f(\mathbf{x})] = V_a(f|\mathcal{A}_d). \quad (5)$$

*Moreover, for $\mathbb{P}$ attaining the minimum in* (4)*, the inequality in $\Gamma$ will reduce to equality at $\mathbb{P}$.*

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 $\mathbb{P}$ 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.

*Proof Sketch.* The full proof is in Appendix 6, and we provide a sketch here. For any action set $\mathcal{A}_a \supseteq \mathcal{A}_d$ the agent has, and any optimal action $(\mathbb{P}, c)$ he chooses under $\mathcal{A}_a$ and the rational decision rule $f$, the expected utility the agent gets from $f$ must satisfy:

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

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 $\Gamma$. Furthermore, the decision maker's utility $V_d(f|\mathcal{A}_a) = \mathbb{E}_{\mathbb{P}}[h(\mathbf{x})]$ is at the least the minimum given by Eqn. (4). Thus, we have $V_d(f) \geq \min_{\mathbb{P}\in\Gamma} \mathbb{E}_{\mathbb{P}}[h(\mathbf{x})]$. 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 $\min_{\mathbb{P}\in\Gamma} \mathbb{E}_{\mathbb{P}}[h(\mathbf{x})]$. 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 Improve nonlinear rule to a linear one

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) rational decision rule $f$, there exists a linear one $f'$ such that: $V_d(f') \geq V_d(f)$.*

*Proof Sketch.* The full proof is in Appendix 7. 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 $f$.

In more detail, given a decision rule $f$, consider a point $(\mathbb{E}_{\mathbb{P}}[\mathbf{x}], \mathbb{E}_{\mathbb{P}}[f(\mathbf{x})])$ generated by any possible action $(\mathbb{P}, c)$. 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 $\Gamma$ 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. □

### 3.3 Wrapping up

We have shown that any rational 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. $(\boldsymbol{\omega}, \beta) \in \mathcal{G}^{\text{lin}}$, this guarantees that $V_d(f)$ has a maximum over the compact set $\mathcal{G}^{\text{lin}}$. We defer the proof to Appendix 8.

### 3.4 Illustrating example: Student evaluation

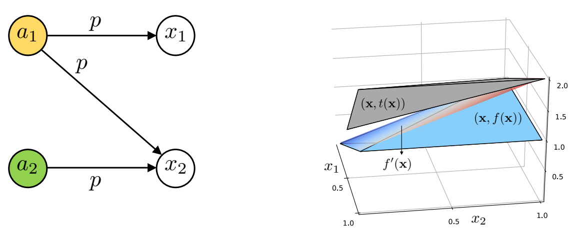
We now use the example of student evaluation to demonstrate the intuitions of our results and analysis. We first illustrate the application of Lemma 2: For a particular nonlinear decision rule, we show how to find an improved linear decision rule. Then, we compute the worst case utility for both decision 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). Assume the cost of actions are the same, the student needs to decide a distribution over the actions. Using the terminology by Kleinberg and Raghavan (2019), we say the student needs to allocate their effort budget of 1 to two actions, with $e_j$ denoting the effort of (i.e., the probability of choosing) action $a_j$. The effort-feature conversion obeys the following rule: $\Pr(x_i = 1) = \sum_j w_{j,i} \cdot e_j$, where $w_{j,i} \in [0, 1]$ is the 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. 2a.

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 $a_1$). This could correspond to the teacher setting her utility function as $h(\mathbf{x}) = \boldsymbol{\omega}_h^\top \mathbf{x} + \beta_h$, where $\boldsymbol{\omega}_h = (1, 0)$ and $\beta_h$ is a small positive value[3]. 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 $p$.

[3] $\beta_h$ can be used to guarantee the decision rules are rational.

Note that $f$ is a (non-robust) optimal decision rule for maximizing $h$. We now show that by leveraging the constructive proof in Lemma 2, we can find a linear rule that weakly improves the worst-case utility. In particular, 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, as illustrated in Fig. 2b. From Lemma 2, $f'$ weakly improves over $f$ in terms of worst-case utility. Below we compute the worst-case utility for $f'$ and $f$ for confirmation. For $f'$, by Eqn. (4) in Lemma 1, $V_d(f') = \min_{\mathbb{P}\in\Delta(\mathcal{X})} \mathbb{E}_{\mathbb{P}}[h(\mathbf{x})] = \min_{\mathbb{P}\in\Delta(\mathcal{X})} \mathbb{E}_{\mathbb{P}}[x_1] + \beta$, where $\mathbb{P}$ satisfies $\mathbb{E}_{\mathbb{P}}[f'(\mathbf{x})] = \mathbb{E}_{\mathbb{P}}[x_1 + x_2] \geq V_a(f'|\mathcal{A}_d)$. Observe that, when the student's available action set $\mathcal{A}_d$ is depicted as in Fig. 2a, $V_a(f'|\mathcal{A}_d) = 2p$. Since when $\mathbb{P}$ attains the minimum of $V_d(f')$, the inequality must bind. Thus, we have $\min_{\mathbb{P}\in\Delta(\mathcal{X})} \mathbb{E}_{\mathbb{P}}[x_1] = 2p - 1$, 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 $V_d(f')$.




(a) (b)

Figure 2: (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 (where $w_{0,1} = p - \epsilon$ and $w_{0,2} = p + \epsilon$ for some $\epsilon \in (0, 1-p)$). 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) 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 $p$.

# 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 $f^*$. Below 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.*
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 10. The key idea is to first formulate the problem of computing $f^*$ as an optimization problem. We then demonstrate 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 $f_{(\boldsymbol{\omega},\beta)} \equiv \boldsymbol{\omega}^\top \mathbf{x} + \beta$, where $(\boldsymbol{\omega}, \beta) \in \mathcal{G}^{\text{lin}}$ (see Definition 1). We use `SO` to denote the corresponding (non-robust) 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):

$$\arg\max_{(\boldsymbol{\omega},\beta)\in\mathcal{G}^{\text{lin}}} \mathbb{E}_{\mathbb{P}}[h(\mathbf{x})], \tag{SO}$$

$$\text{s.t. } (\mathbb{P}, c) \in \arg\max_{(\mathbb{P},c)\in\mathcal{A}_d} \mathbb{E}_{\mathbb{P}}[f_{(\boldsymbol{\omega},\beta)}(\mathbf{x})] - c, \tag{6}$$

where $\mathbb{E}_{\mathbb{P}}[\cdot]$ is the expectation taken with respect to the random vector $\mathbf{x}$ given that it follows the probability distribution $\mathbb{P}$. Note that while there exist efficient algorithms to solve this (non-robust) decision making under uncertainty (`SO`) in restricted cases, the problem is known to be NP-hard in general (Hansen et al., 1992; Kleinberg and Raghavan, 2019). Therefore, instead of investigating the complexity of solving the robust variant, we aim to understand the *additional* complexity of requiring robustness in (non-robust) decision making under uncertainty.

With slight abuse of notation, for any $(\boldsymbol{\omega}, \beta) \in \mathcal{G}^{\text{lin}}$, we use $(\mathbb{P}_{\boldsymbol{\omega}}, c_{\boldsymbol{\omega}}) \in \mathcal{A}_d$ to denote the solution to the constraint (6) in SO and let $C_{\boldsymbol{\omega}} = \mathbb{E}_{\mathbb{P}_{\boldsymbol{\omega}}}[f_{(\boldsymbol{\omega},\beta)}(\mathbf{x})] - c_{\boldsymbol{\omega}}$. Then according to Lemma 1, we can compute $f^*$ by solving:

$$\arg\max_{(\boldsymbol{\omega},\beta)\in\mathcal{G}^{\text{lin}}} \min_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[h(\mathbf{x})], \tag{R-SO}$$

where $\mathcal{P}$ can be expressed as follows:

$$\mathcal{P} \equiv \left\{ \mathbb{P} \in \Delta(\mathcal{X}) \,\middle|\, \begin{array}{l} \mathbf{x} \sim \mathbb{P}, \\ \mathbb{E}_{\mathbb{P}}[\boldsymbol{\omega}^\top \mathbf{x}] = C_{\boldsymbol{\omega}} - \beta, \\ \Pr(\mathbf{x} \in \mathcal{X}) = 1 \end{array} \right\}. \tag{7}$$

Different from the problem in SO, after identifying the agent's best response $(\mathbb{P}_{\boldsymbol{\omega}}, c_{\boldsymbol{\omega}}) \in \mathcal{A}_d$ under $f_{(\boldsymbol{\omega},\beta)}$, our problem in R-SO will have an additional layer of optimization over the set $\mathcal{P}$. It is easy to see that this is a linear program with equality constraint, where the decision variables are a probability simplex over $\mathcal{X}$. Therefore, the computation of R-SO can be decomposed into the computation of SO and a linear program. This decomposition enables us to complete the proof.

### 4.1 Solving R-SO when $\mathcal{X}$ is Infinite

So far, we demonstrate that the additional complexity of solving robust decision making under uncertainty (R-SO) compared with the non-robust version (SO) can be characterized by a linear program. When the space of agent features $\mathcal{X}$ is finite, this additional complexity is polynomial. However, in some applications, the space of agent features could be infinite, e.g., with real-valued features. In this subsection, we investigate the situation when $\mathcal{X}$ is infinite, through adapting the techniques from distributional robust optimization (Delage and Ye, 2010; Jiang and Guan, 2016; Ben-Tal et al., 2013).

To highlight the additional complexity of requiring robustness, in the following discussion, we assume that there exists an oracle that can provide agent's best response $(\mathbb{P}_{\boldsymbol{\omega}}, c_{\boldsymbol{\omega}})$ and compute the value $C_{\boldsymbol{\omega}}$ for any $(\boldsymbol{\omega}, \beta) \in \mathcal{G}^{\text{lin}}$ in time polynomial in $n$. Equipped with such an oracle, the problem defined in R-SO resembles the spirit of distributionally robust optimization (in short DRO), which aims to evaluate optimal solutions under the worst-case expectation with respect to a family of probability distributions of the uncertain parameters. The key concept in DRO is the ambiguity set, a family of measures consistent with the prior knowledge about uncertainty. In our formulation, the ambiguity set $\mathcal{P}$ is specified via a hyperplane (see Eqn. (7)).

While our discussion so far applies for arbitrary utility functions $h(\cdot)$, analyzing the additional complexity of R-SO is challenging when the agent feature space $\mathcal{X}$ is infinite. Therefore, in the rest of this subsection, we focus on a general set of concave and piecewise utility functions as defined below.

$$h(\mathbf{x}) = \min_{k\in[K]} h_k(\mathbf{x}). \tag{8}$$

Note that this set of utility functions is general since many commonly-seen utility functions are concave and can usually be approximated using simple piecewise functions, such as the piecewise linear functions: $h_k(\mathbf{x}) = \mathbf{a}_k^\top \mathbf{x} + b_k$, where for all $k \in [K]$, $\mathbf{a}_k \in \mathbb{R}^n$ and $b_k \in \mathbb{R}$ and the piecewise quadratic functions: $h_k(\mathbf{x}) = \min_{k\in[K]} \mathbf{x}^\top \mathbf{A}_k \mathbf{x} + \mathbf{b}_k^\top \mathbf{x} + c_k$ where for all $k \in [K]$, $\mathbf{A}_k \succeq 0$ and $\mathbf{A}_k \in \mathbb{R}^{n\times n}$, $\mathbf{b}_k \in \mathbb{R}^n$ and $c_k \in \mathbb{R}$.

**Theorem 3.** *Given that $\mathcal{X}$ is ellipsoidal, i.e., $\mathcal{X} = \{\mathbf{x} : (\mathbf{x} - \mathbf{x}_0)^\top \boldsymbol{\Theta}(\mathbf{x} - \mathbf{x}_0) \le 1\}$, where $\boldsymbol{\Theta}$ has at least one strictly positive eigenvalue, the objective of the problem R-SO is the same as the optimal value of the following optimization problem:*

*When $h(\cdot)$ is a piecewise linear function:*

$$\begin{aligned} &\arg\min_{(\boldsymbol{\omega},\beta)\in\mathcal{G}^{lin},\alpha,\lambda,\boldsymbol{\tau}} \quad \alpha + \lambda(C_{\boldsymbol{\omega}} - \beta) \qquad (9)\\ &s.t. \quad \begin{bmatrix} \tau_k \boldsymbol{\Theta} & \frac{\lambda\cdot\boldsymbol{\omega} + \mathbf{a}_k - 2\tau_k \mathbf{x}_0^\top \boldsymbol{\Theta}}{2} \\ \frac{\lambda\cdot\boldsymbol{\omega}^\top + \mathbf{a}_k^\top - 2\tau_k \boldsymbol{\Theta}\mathbf{x}_0}{2} & \alpha + b_k + \tau_k(\mathbf{x}_0^\top \boldsymbol{\Theta}\mathbf{x}_0 - 1) \end{bmatrix} \succeq 0, \forall k \\ &\qquad \alpha \in \mathbb{R}, \lambda \in \mathbb{R}, \tau_k \ge 0, \forall k \in [K]. \end{aligned}$$

*When $h(\cdot)$ is piecewise quadratic function, the first constraint will be replaced by the following*

$$\begin{bmatrix} \mathbf{A}_k & \frac{\lambda\cdot\boldsymbol{\omega} + \mathbf{b}_k}{2} \\ \frac{\lambda\cdot\boldsymbol{\omega}^\top + \mathbf{b}_k^\top}{2} & \alpha + c_k \end{bmatrix} \succeq -\tau_k \begin{bmatrix} \boldsymbol{\Theta} & -\boldsymbol{\Theta}\mathbf{x}_0 \\ -\mathbf{x}_0^\top \boldsymbol{\Theta} & \mathbf{x}_0^\top \boldsymbol{\Theta}\mathbf{x}_0 - 1 \end{bmatrix}, \forall k$$

*Proof.* To solve R-SO, we first reformulate it as a minimization problem:

$$\min_{(\boldsymbol{\omega},\beta)\in\mathcal{G}^{\text{lin}}} \max_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}\left[\max_{k\in[K]} -h_k(\mathbf{x})\right]. \tag{10}$$

For every $(\boldsymbol{\omega}, \beta) \in \mathcal{G}^{\text{lin}}$, let $\rho(\boldsymbol{\omega}, \beta)$ denote the inner supremum problem in (10) over $\mathcal{P}$:

$$\rho(\boldsymbol{\omega}, \beta) \equiv \max_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}\left[\max_{k\in[K]} -h_k(\mathbf{x})\right]. \tag{11}$$

We can now recast the inner supremum problem $\rho(\boldsymbol{\omega}, \beta)$ as a minimization problem, which can be performed jointly with the outer minimization over $\mathcal{G}^{\text{lin}}$. Introducing dual variables $\alpha, \lambda$ that correspond to the respective probability and expectation constraints in (7), we have the following dual of $\rho(\boldsymbol{\omega},\beta)$

$$\begin{aligned}\rho_{\texttt{dual}}(\boldsymbol{\omega},\beta)\equiv\min_{\alpha,\lambda}\ &\alpha+\lambda(C_{\boldsymbol{\omega}}-\beta) \qquad (12)\\ \text{s.t. }\ &\alpha+\lambda\cdot\boldsymbol{\omega}^\top\mathbf{x}\geq -h(\mathbf{x}),\forall\mathbf{x}\in\mathcal{X}\\ &\alpha\in\mathbb{R},\lambda\in\mathbb{R},\end{aligned}$$

which provides an upper bound on $\rho(\boldsymbol{\omega},\beta)$. Indeed, consider any $\mathbb{P}\in\mathcal{P}$ and any feasible solution $(\alpha,\lambda)$ in problem (12); the robust counterpart in the dual implies that

$$\mathbb{E}_{\mathbb{P}}[-h(\mathbf{x})]\leq\mathbb{E}_{\mathbb{P}}\left[\alpha+\lambda\cdot\boldsymbol{\omega}^\top\mathbf{x}\right]=\alpha+\lambda(C_{\boldsymbol{\omega}}-\beta).$$

Thus, we have that weak duality holds: $\rho(\boldsymbol{\omega},\beta)\leq\rho_{\texttt{dual}}(\boldsymbol{\omega},\beta)$. Furthermore, the strong duality also holds since the problem (11) is a linear optimization problem (with respect to $\mathbb{P}$). Having established the dual of (11) and its strong duality, we can formulate the problem (11) via a min-min operation that can be performed jointly over the constraint involving $h(\mathbf{x})$ decomposes.

$$\begin{aligned}\min_{(\boldsymbol{\omega},\beta)\in\mathcal{G}^{\text{lin}},\alpha,\lambda}\ &\alpha+\lambda(C_{\boldsymbol{\omega}}-\beta) \qquad (13)\\ \text{s.t. }\ &\alpha+\lambda\cdot\boldsymbol{\omega}^\top\mathbf{x}+h_k(\mathbf{x})\geq 0,\forall\mathbf{x}\in\mathcal{X},k\in[K] \qquad (14)\\ &\alpha\in\mathbb{R},\lambda\in\mathbb{R}.\end{aligned}$$

Note that when $\mathcal{X}$ has infinite elements, i.e., $\mathbb{P}$ is a measure with infinite support over $\mathcal{X}$, there will be infinite-many constraints in (14). However, with our assumption on function $h(\cdot)$ and leveraging the geometry of $\mathcal{X}$, we can reduce the above optimization problem with infinite-many constraints to the problem with tractable finite number of constraints. In particular, when $\mathcal{X}$ is ellipsoidal, i.e., $\mathcal{X}=\{\mathbf{x}:(\mathbf{x}-\mathbf{x}_0)^\top\boldsymbol{\Theta}(\mathbf{x}-\mathbf{x}_0)\leq 1\}$, and $\boldsymbol{\Theta}$ has at least one one strictly positive eigenvalue, we can apply S-Lemma (cf., Theorem 2.2 in Pólik and Terlaky (2007)) for any given $k\in[K]$ to replace Constraint (14), which enforces that

$$\nexists\mathbf{x}\in\mathbb{R}^n\text{ s.t. }\alpha+\lambda\cdot\boldsymbol{\omega}^\top\mathbf{x}+h_k(\mathbf{x})<0\ \bigwedge\ (\mathbf{x}-\mathbf{x}_0)^\top\boldsymbol{\Theta}(\mathbf{x}-\mathbf{x}_0)\leq 1 \qquad (15)$$

with the equivalent constraint that

$$\exists\tau_k\geq 0\text{ s.t. }\forall\mathbf{x}\in\mathbb{R}^n,\ \alpha+\lambda\cdot\boldsymbol{\omega}^\top\mathbf{x}+h_k(\mathbf{x})\geq -\tau_k\left((\mathbf{x}-\mathbf{x}_0)^\top\boldsymbol{\Theta}(\mathbf{x}-\mathbf{x}_0)-1\right). \qquad (16)$$

When $h_k(\mathbf{x})=\mathbf{a}_k^\top\mathbf{x}+b_k$, then one can further use Schur's complement to replace Constraint (14) by an equivalent linear matrix inequality for any $k\in[K]$:

$$\begin{bmatrix}\tau_k\boldsymbol{\Theta} & \frac{\lambda\cdot\boldsymbol{\omega}+\mathbf{a}_k}{2}-\tau_k\mathbf{x}_0^\top\boldsymbol{\Theta}\\ \frac{\lambda\cdot\boldsymbol{\omega}^\top+\mathbf{a}_k^\top}{2}-\tau_k\boldsymbol{\Theta}\mathbf{x}_0 & \alpha+b_k+\tau_k(\mathbf{x}_0^\top\boldsymbol{\Theta}\mathbf{x}_0-1)\end{bmatrix}\succeq 0.$$

The problem can therefore be reformulated as:

$$\begin{aligned}\min_{(\boldsymbol{\omega},\beta)\in\mathcal{G}^{\text{lin}},\alpha,\lambda,\boldsymbol{\tau}}\ &\alpha+\lambda(C_{\boldsymbol{\omega}}-\beta)\\ \text{s.t. }\ &\begin{bmatrix}\tau_k\boldsymbol{\Theta} & \frac{\lambda\cdot\boldsymbol{\omega}+\mathbf{a}_k}{2}-\tau_k\mathbf{x}_0^\top\boldsymbol{\Theta}\\ \frac{\lambda\cdot\boldsymbol{\omega}^\top+\mathbf{a}_k^\top}{2}-\tau_k\boldsymbol{\Theta}\mathbf{x}_0 & \alpha+b_k+\tau_k(\mathbf{x}_0^\top\boldsymbol{\Theta}\mathbf{x}_0-1)\end{bmatrix}\\ &\alpha\in\mathbb{R},\lambda\in\mathbb{R},\tau_k\geq 0,\forall k\in[K]\end{aligned}$$

where $\boldsymbol{\tau}=(\tau_1,\ldots,\tau_K)$. When $h_k(\mathbf{x})=\min_{k\in[K]}\mathbf{x}^\top\mathbf{A}_k\mathbf{x}+\mathbf{b}_k^\top\mathbf{x}+c_k$ where for all $k\in[K]$, $\mathbf{A}_k\succeq 0$ and $\mathbf{A}_k\in\mathbb{R}^{n\times n},\mathbf{b}_k\in\mathbb{R}^n$ and $c_k\in\mathbb{R}$. We then have following equivalent linear matrix inequality for Constraint (14) for any $k\in[K]$:

$$\begin{bmatrix}\mathbf{A}_k & \frac{\lambda\cdot\boldsymbol{\omega}+\mathbf{b}_k}{2}\\ \frac{\lambda\cdot\boldsymbol{\omega}^\top+\mathbf{b}_k^\top}{2} & \alpha+c_k\end{bmatrix}\succeq -\tau_k\begin{bmatrix}\boldsymbol{\Theta} & -\boldsymbol{\Theta}\mathbf{x}_0\\ -\mathbf{x}_0^\top\boldsymbol{\Theta} & \mathbf{x}_0^\top\boldsymbol{\Theta}\mathbf{x}_0-1\end{bmatrix}.$$

The problem can therefore be reformulated as:

$$\begin{aligned}\min_{(\boldsymbol{\omega},\beta)\in\mathcal{G}^{\text{lin}},\alpha,\lambda,\boldsymbol{\tau}}\ &\alpha+\lambda(C_{\boldsymbol{\omega}}-\beta)\\ \text{s.t. }\ &\begin{bmatrix}\mathbf{A}_k & \frac{\lambda\cdot\boldsymbol{\omega}+\mathbf{b}_k}{2}\\ \frac{\lambda\cdot\boldsymbol{\omega}^\top+\mathbf{b}_k^\top}{2} & \alpha+c_k\end{bmatrix}\succeq -\tau_k\begin{bmatrix}\boldsymbol{\Theta} & -\boldsymbol{\Theta}\mathbf{x}_0\\ -\mathbf{x}_0^\top\boldsymbol{\Theta} & \mathbf{x}_0^\top\boldsymbol{\Theta}\mathbf{x}_0-1\end{bmatrix}\\ &\alpha\in\mathbb{R},\lambda\in\mathbb{R},\tau_k\geq 0,\forall k\in[K]\end{aligned}$$

where $\boldsymbol{\tau}=(\tau_1,\ldots,\tau_K)$. □

In both linear and quadratic $h_k(\mathbf{x})$, we show that we can reformulate the original problem `R-SO` to the problem with a finite number of tractable linear matrix inequalities, instead of infinitely many constraints with $\mathcal{X}$. This formulation provides a tool for us to analyze the additional complexity of `R-SO` compared with `SO`. For example, the problem in (9) could be possibly efficiently solvable when $C_{\boldsymbol{\omega}}$ of the agent's best response exhibits nice behaviors to retain a convexity of the objective in (9), e.g., when $C_{\boldsymbol{\omega}}$ is a linear form of $\boldsymbol{\omega}$, then (9) is a semi-definite program. Then it is known that an interior point algorithm can be used to solve the above SDP with the polynomial time, i.e., the above problem can be solved to any precision $\epsilon$ in time polynomial in $\log(1/\epsilon)$ and the sizes of the problem. We leave the full characterization of conditions for the problem to be efficiently solvable for future work.

# 5 Discussions and Future Work

Linear models, one of the "white-box" 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 models to deploy. While we demonstrate that finding the robust optimal decision rule is generally hard, our analysis in decomposing the problem could provide 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.

### Acknowledgements

This work is supported in part by the Office of Naval Research Grant N00014-20-1-2240, National Science Foundation grant IIS-1939677, and Amazon.

## 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.

Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2): 341–357, 2013.

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.

Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations research*, 58(3):595–612, 2010.

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. In *29th International Joint Conference on Artificial Intelligence*, 2020.

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

Pierre Hansen, Brigitte Jaumard, and Gilles Savard. New branch-and-bound rules for linear bilevel programming. *SIAM Journal on scientific and Statistical Computing*, 13(5):1194–1217, 1992.

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.

Ruiwei Jiang and Yongpei Guan. Data-driven chance constrained stochastic program. *Mathematical Programming*, 158(1-2):291–327, 2016.

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.

Imre Pólik and Tamás Terlaky. A survey of the s-lemma. *SIAM review*, 49(3):371–418, 2007.

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.