Linear Models are Robust Optimal Under Strategic Behavior: Supplementary Materials

6 Proof for Lemma 1

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

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

Thus, 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 the (4). This implies the following guarantee of worst-case utility $V_d(f)$:

$$V_d(f) \geq \min_{\mathbb{P}\in\Delta(\mathcal{X})} \mathbb{E}_{\mathbb{P}}[h(\mathbf{x})] \quad \text{s.t.} \quad \mathbb{E}_{\mathbb{P}}[f(\mathbf{x})] \geq V_a(f|\mathcal{A}_d). \tag{17}$$

We now show that (17) is tight. Let $\texttt{supp}(\mathbb{P})$ denote the support of distribution $\mathbb{P}$. Let $\mathbb{P}_0$ be a distribution attaining the minimum in (4) and also satisfying the constraint. We consider following two cases:
Case 1: $\texttt{supp}(\mathbb{P}_0) \not\subset \arg\max_{\mathbf{x}} f(\mathbf{x})$. Then let $\mathbb{P}_1$ be a distribution which achieves a higher value of $\mathbb{E}_{\mathbb{P}}[f(\mathbf{x})]$. Let $\mathbb{P}'$ be a mixture distribution $\mathbb{P}' = (1-\epsilon)\mathbb{P}_0 + \epsilon\mathbb{P}_1$, with a small positive ϵ. Then we have $\mathbb{E}_{\mathbb{P}'}[f(\mathbf{x})] = (1-\epsilon)\mathbb{E}_{\mathbb{P}_0}[f(\mathbf{x})] + \epsilon\mathbb{E}_{\mathbb{P}_1}[f(\mathbf{x})] > \mathbb{E}_{\mathbb{P}_0}[f(\mathbf{x})]$. Now take $\mathcal{A}'_a = \mathcal{A}_d \cup \{(\mathbb{P}', 0)\}$, then the agent's unique optimal action under $\mathcal{A}'_a$ is $(\mathbb{P}', 0)$. This brings the decision maker with utility of $V_d(f|\mathcal{A}'_a) = (1-\epsilon)\mathbb{E}_{\mathbb{P}_0}[h(\mathbf{x})] + \epsilon\mathbb{E}_{\mathbb{P}_1}[h(\mathbf{x})]$. Since $V_d(f|\mathcal{A}'_a) \geq V_d(f)$, we further have

$$V_d(f) \leq V_d(f|\mathcal{A}'_a) = (1-\epsilon)\mathbb{E}_{\mathbb{P}_0}[h(\mathbf{x})] + \epsilon\mathbb{E}_{\mathbb{P}_1}[h(\mathbf{x})]. \tag{18}$$

When $\epsilon \to 0$, the RHS in (18) will converge to $\mathbb{E}_{\mathbb{P}_0}[h(\mathbf{x})]$.This implies $V_d(f) \leq \mathbb{E}_{\mathbb{P}_0}[h(\mathbf{x})]$ when $\epsilon \to 0$. Recall our definition of $\mathbb{P}_0$, and together with the lower bound we have shown for $V_d(f)$ in (17), we can conclude our results in (4) for this case.

Case 2: $\texttt{supp}(\mathbb{P}_0) \subset \arg\max_{\mathbf{x}} f(\mathbf{x})$. For this case, we discuss following two situations.
(i): $\mathbb{E}_{\mathbb{P}_0}[f(\mathbf{x})] > V_a(f|\mathcal{A}_d)$, we now consider action set $\mathcal{A}'_a = \mathcal{A}_d \cup \{(\mathbb{P}_0, 0)\}$. Since $\mathbb{E}_{\mathbb{P}_0}[f(\mathbf{x})] > V_a(f|\mathcal{A}_d)$, then the agent will uniquely chose action $(\mathbb{P}_0, 0)$ for f under the action set $\mathcal{A}'_a$. This brings the decision maker with the utility of $V_d(f|\mathcal{A}'_a) = \mathbb{E}_{\mathbb{P}_0}[h(\mathbf{x})]$. Again, with the fact that $V_d(f|\mathcal{A}'_a) \geq V_d(f)$ and the definition of $\mathbb{P}_0$, we have now proved (4).
(ii): $\mathbb{E}_{\mathbb{P}_0}[f(\mathbf{x})] = V_a(f|\mathcal{A}_d) = \max f(\mathbf{x})$, this situation can only be satisfied when $\mathcal{A}_d$ contains some action of the form $(\mathbb{P}', 0)$ with $\texttt{supp}(\mathbb{P}') \subset \arg\max f(\mathbf{x})$. Thus, we define

$$\mathcal{G} := \{(\mathbb{P}', 0) \in \mathcal{A}_d : \texttt{supp}(\mathbb{P}') \subset \arg\max f(\mathbf{x})\} \neq \emptyset.$$

Then, under action set $\mathcal{A}_d$, the agent will choose an action in $\mathcal{G}$ which would benefit decision maker (according to the tie-breaking assumption, 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 $V_d(f|\mathcal{A}_d) = \max_{(\mathbb{P},0)\in\mathcal{G}} \mathbb{E}_{\mathbb{P}}[h(\mathbf{x})] \geq V_d(f)$. In this scenario, the unique optimal action for the agent under any action set $\mathcal{A} \supseteq \mathcal{A}_d$ is some $(\mathbb{P}, 0) \in \mathcal{G}$. However, the agent would stick to the same action even under zero decision rule (recall our tie-breaking assumption), leading the decision maker's utility $V_d(0|\mathcal{A}) = \max_{(\mathbb{P},0)\in\mathcal{G}} \mathbb{E}_{\mathbb{P}}[h(\mathbf{x})] = V_d(0)$. This implies $V_d(0) \geq V_d(f)$, which contradicts our rationality assumption.

Now we establish the equality claims. Without loss of generality, we may assume the agent has a costless action $(\delta_{\underline{\mathbf{x}}}, 0)$ in $\mathcal{A}_d$ where $h(\underline{\mathbf{x}}) = 0$.[4] Recall that we have $\mathbb{E}_{\mathbb{P}_0}[h(\mathbf{x})] = V_d(f) > V_d(0) > 0$ by our assumption on $\mathbb{P}_0$ and DM's rationality. If we have $\mathbb{E}_{\mathbb{P}_0}[f(\mathbf{x})] > V_a(f|\mathcal{A}_d)$ strictly, then replace $\mathbb{P}_0$ by a mixture distribution $\mathbb{P}' = (1-\epsilon)\mathbb{P}_0 + \epsilon\delta_{\underline{\mathbf{x}}}$ for small ϵ. Consider $\mathcal{A}'_a = \mathcal{A}_d \cup \{(\mathbb{P}', 0)\}$, then the agent's utility by taking the action $(\mathbb{P}', 0)$ is given by $V_a(f|\mathcal{A}'_a) = (1-\epsilon)\mathbb{E}_{\mathbb{P}_0}[f(\mathbf{x})] + \epsilon f(\underline{\mathbf{x}})$, then one can always 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 decision maker with a utility of $V_d(f|\mathcal{A}'_a) = (1-\epsilon)\mathbb{E}_{\mathbb{P}_0}[h(\mathbf{x})] + \epsilon h(\underline{\mathbf{x}}) = (1-\epsilon)\mathbb{E}_{\mathbb{P}_0}[h(\mathbf{x})]$. Since $V_d(f|\mathcal{A}'_a) \geq V_d(f)$, given any positive ϵ, this implies that $V_d(f) \leq (1-\epsilon)\mathbb{E}_{\mathbb{P}_0}[h(\mathbf{x})] < \mathbb{E}_{\mathbb{P}_0}[h(\mathbf{x})]$, which contradicts the minimality of $\mathbb{P}_0$. Thus we have $\mathbb{E}_{\mathbb{P}_0}[f(\mathbf{x})] = V_a(f|\mathcal{A}_d)$. Finally, if $\mathbb{P}_0 \in \arg\max_{\mathbb{P}\in\Delta(\mathcal{X})} \mathbb{E}_{\mathbb{P}}[f(\mathbf{x})]$, and $\mathbb{E}_{\mathbb{P}_0}[f(\mathbf{x})] = V_a(f|\mathcal{A}_d)$, 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* $(\mathbb{P}, 0)$. *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 realistic models that can explicitly incorporate 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* $\mathbb{P}$ *at the lowest allowable cost.*

7 Proof for Lemma 2

Proof. Our proof structure is similar to Carroll (2015), with the key difference on how to define the two disjoint convex sets. Suppose 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. A graph illustration of our proof is presented in Figure 3.
We now claim that $\mathcal{S}$ and $\mathcal{T}$ are disjoint. To see this, suppose $\mathcal{S}$ and $\mathcal{T}$ are not disjoint, then there exists a distribution $\mathbb{P} \in \Delta(\mathcal{X})$ such that $\mathbb{E}_{\mathbb{P}}[f(\mathbf{x})] > \mathbb{E}_{\mathbb{P}}[t(\mathbf{x})]$. In particular, we have

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

and also

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

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

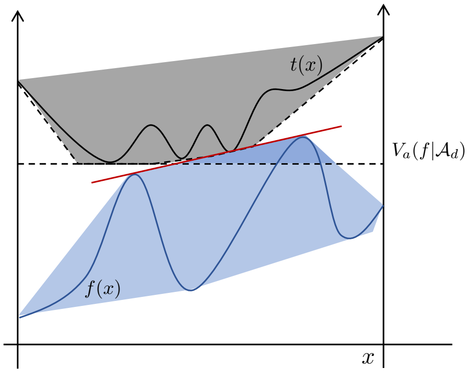

Figure 3: Illustrate $\mathcal{S}$ and $\mathcal{T}$ when $n = 1$. The blue line is $f(\mathbf{x})$ and its associated convex hull in blue shaded region (the top blue triangle is the set Γ). Black line is $t(\mathbf{x})$. The black shaded region is the convex hull for all points $(\mathbf{x}, z)$ where $z > t(\mathbf{x})$. The red line is the hyperplane to separate $\mathcal{S}$ and $\mathcal{T}$.

The disjointness and convexity of $\mathcal{S}$ and $\mathcal{T}$ enable us to apply the separating hyperplane theorem: There exists a

[4] This assumption is merely an additive normalization of the decision maker's utility and it can be relaxed to a more general scenario where our reulsts still hold (see our discussion at the end of the Appendix 7). Earlier 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 $(\delta_0, 0) \in \mathcal{A}_d$.

vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$ and constants μ, v such that

$$\sum_i \lambda_i x_i + \mu z \le v, \quad \forall (\mathbf{x}, z) \in \mathcal{S} \tag{19}$$

$$\sum_i \lambda_i x_i + \mu z \ge v, \quad \forall (\mathbf{x}, z) \in \mathcal{T} \tag{20}$$

and $\boldsymbol{\lambda}$ is a non-zero vector. Note that (19) and (20) 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 (19) gives $\sum_i \lambda_i x_i \le v$ and (20) 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 (19) 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

$$f'(\mathbf{x}) = \frac{v - \sum_i \lambda_i x_i}{\mu}, \quad \forall \mathbf{x} \in \mathcal{X}. \tag{21}$$

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 $(\mathbb{P}_0, c_0)$ 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

$$V_a(f'|\mathcal{A}_a) \ge V_a(f'|\mathcal{A}_d) \ge V_a(f|\mathcal{A}_d). \tag{22}$$

Let $(\mathbb{P}, c)$ be the action that the agent chooses under f' and action set $\mathcal{A}_a$. Then (20) implies

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}}[t(\mathbf{x})] &\ge \frac{v - \sum_i \lambda_i \mathbb{E}_{\mathbb{P}}[x_i]}{\mu} \\
&= \mathbb{E}_{\mathbb{P}}[f'(\mathbf{x})] && (23) \\
&= V_a(f'|\mathcal{A}_a) + c \\
&\ge V_a(f'|\mathcal{A}_a) && (c \in \mathbb{R}_+) \\
&\ge V_a(f|\mathcal{A}_d). && \text{(by (22))}
\end{aligned}$$

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

$$\mathbb{E}_{\mathbb{P}}[t(\mathbf{x})] = \mathbb{E}_{\mathbb{P}}[h(\mathbf{x})] + \mathbb{E}_{\mathbb{P}}[f(\mathbf{x})] - V_d(f). \tag{24}$$

So we have

$$\begin{aligned}
V_d(f'|\mathcal{A}_a) = \mathbb{E}_{\mathbb{P}}[h(\mathbf{x})] &= \mathbb{E}_{\mathbb{P}}[t(\mathbf{x})] - \mathbb{E}_{\mathbb{P}}[f(\mathbf{x})] + V_d(f) \\
&\ge \mathbb{E}_{\mathbb{P}}[t(\mathbf{x})] - \mathbb{E}_{\mathbb{P}}[f'(\mathbf{x})] + V_d(f) && \text{(by definition of } f'\text{)} \\
&\ge V_d(f). && \text{(by 23)}
\end{aligned}$$

On the other hand, if $\mathbb{E}_{\mathbb{P}}[t(\mathbf{x})] = V_a(f|\mathcal{A}_d)$. 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 $(\mathbb{P}_0, c_0)$, 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: $V_d(f'|\mathcal{A}_a) \ge \mathbb{E}_{\mathbb{P}_0}[h(\mathbf{x})] =$

$V_d(f|\mathcal{A}_d) \geq V_d(f)$, 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 $\mathbb{P}$ such that

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

This contradicts our statement (5) in Lemma 1. Using the same arguments as in case 1, we find a vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$ and constants μ, v such that (19) and (20) hold, and we can still guarantee that $\mu > 0$. Again, we define a linear decision rule f' by (21); from (19) we know that $f' \geq f$ pointwise. Consider the agent's behavior under decision rule f', for any action $(\mathbb{P}, c)$ chosen by the agent under any possible action set, we have

$$\mathbb{E}_{\mathbb{P}}[f'(\mathbf{x})] - c = f'(\mathbb{E}_{\mathbb{P}}[\mathbf{x}]) - c \leq V_a(f|\mathcal{A}_d). \qquad \text{(by (20))}$$

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 $\mathbb{P}$, 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: $V_d(f) = \min_{\mathbb{P} \in \Delta(\mathcal{X})} \mathbb{E}_{\mathbb{P}}[h(\mathbf{x})]$, s.t. $\mathbb{E}_{\mathbb{P}}[f(\mathbf{x})] - \underline{c} \geq V_a(f|\mathcal{A}_d)$ or $\max_{\mathbb{P} \in \Delta(\mathcal{X})} \mathbb{E}_{\mathbb{P}}[f(\mathbf{x})] - \underline{c} = V_a(f|\mathcal{A}_d)$. 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.

8 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 rational 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 $f_{(\boldsymbol{\omega},\beta)}(\mathbf{x}) = \boldsymbol{\omega}^\top \mathbf{x} + \beta$. Then it suffices to show that the guaranteed worst-case utility $V_d(f)$ is an upper semi-continuous function of $(\boldsymbol{\omega}, \beta) \in \mathcal{G}^{\text{lin}}$. Now fix a sequence $(\boldsymbol{\omega}^1, \beta^1), (\boldsymbol{\omega}^2, \beta^2), \ldots$ in $\mathcal{G}^{\text{lin}}$ converging to some $(\boldsymbol{\omega}^\infty, \beta^\infty)$ in $\mathcal{G}^{\text{lin}}$. Then it suffices to show that $V_d(f_{(\boldsymbol{\omega}^\infty,\beta^\infty)}) \geq \limsup_k V_d(f_{(\boldsymbol{\omega}^k,\beta^k)})$. To prove this, first note that by replacing the sequence $((\boldsymbol{\omega}^k, \beta^k))$ with a subsequence along which $V_d(f((\boldsymbol{\omega}^k, \beta^k)))$ converges to its lim sup on the original sequence, thus, we can assume that $V_d(f_{(\boldsymbol{\omega}^k,\beta^k)})$ converges to $\limsup_k V_d(f_{(\boldsymbol{\omega}^k,\beta^k)})$. Now for any action set $\mathcal{A}_a$, and let $(\mathbb{P}^k, c^k)$ be the agent's chosen action under $\mathcal{A}_a$ and the decision rule $f_{(\boldsymbol{\omega}^k,\beta^k)}$. Then if necessary, by extracting a further subsequence, we can assume that the sequence $(\mathbb{P}^k, c^k)$ converges to some $(\mathbb{P}^\infty, c^\infty) \in \mathcal{A}_a$. Since the agents' utility are continuous in $(\boldsymbol{\omega}, \beta)$, then $(\mathbb{P}^\infty, c^\infty)$ is an optimal action for the agent under $f_{(\boldsymbol{\omega}^\infty,\beta^\infty)}$, and its utility to the decision maker is the limit of the corresponding utility of $(\mathbb{P}^k, c^k)$ under $f_{(\boldsymbol{\omega}^k,\beta^k)}$. We thus have

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

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

9 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 $\mathcal{A}_d$.

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

Table 1: All possible distributions $\mathbb{P}$ in $\mathcal{A}_d$ 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, and $e_1 + e_0 \in [0,1]$.

10 Missing proof and the Algorithm for Theorem 2

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 $f^* \in \mathcal{G}^{\text{lin}}$ arbitrarily and $V_d(f^*) = 0$.
for every $(\boldsymbol{\omega}, \beta) \in \mathcal{G}^{\text{lin}}$ do
    Let $(\mathbb{P}_0, c_0) \in \arg\max_{(\mathbb{P},c)\in\mathcal{A}_d} \mathbb{E}_{\mathbb{P}}\left[\boldsymbol{\omega}^\top \mathbf{x} + \beta\right] - c$;
    Solve the set $\mathcal{P} = \left\{\mathbb{P} : \boldsymbol{\omega}^\top (\mathbb{E}_{\mathbb{P}_0}[\mathbf{x}] - \mathbb{E}_{\mathbb{P}}[\mathbf{x}]) = c_0, \mathbb{P} \in \Delta(\mathcal{X})\right\}$;
    Compute $V_d\left(f_{(\boldsymbol{\omega},\beta)}\right) = \min_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[h(\mathbf{x})]$;
    if $V_d\left(f_{(\boldsymbol{\omega},\beta)}\right) > V_d(f^*)$ then
        $f^* \leftarrow \boldsymbol{\omega}^\top \mathbf{x} + \beta$.
    end if
end for
Output Robust optimal decision: $f^*$.
```

Proof. According Lemma 1, given $f_{(\boldsymbol{\omega},\beta)}$, for any distribution $\mathbb{P}$ attaining the minimum in (4), we know that the inequality in Γ must bind at $\mathbb{P}$. Let $(\mathbb{P}_{\boldsymbol{\omega}}, c_{\boldsymbol{\omega}}) \in \mathcal{A}_d$ be the solution to the constraint in SO. Then we can compute f^* by solving:

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

$$\text{s.t. } \mathcal{P} = \left\{\mathbb{P}' : \mathbb{E}_{\mathbb{P}'}[f_{(\boldsymbol{\omega},\beta)}(\mathbf{x})] = \boldsymbol{\omega}^\top \mathbb{E}_{\mathbb{P}_{\boldsymbol{\omega}}}[\mathbf{x}] - c_{\boldsymbol{\omega}} = C_{\boldsymbol{\omega}}, \mathbb{P}' \in \Delta(\mathcal{X})\right\}, \tag{25}$$

where we refer to the set $\mathcal{P}$, as the *worst-action set*, since we choose the worst action among it to minimize the expected utility $\mathbb{E}_{\mathbb{P}}[h(\mathbf{x})]$. 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 first turns to characterizing a worst-action set $\mathcal{P}$. Then the searching of f^* will hinge on maximizing $\mathbb{E}_{\mathbb{P}}[h(\mathbf{x})]$ in each $\mathcal{P}$ 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 $f_{(\boldsymbol{\omega},\beta)}$, the additional computational complexity in R-SO is due to the robustness concern in minimizing $\mathbb{E}_{\mathbb{P}}[h(\mathbf{x})]$ over set $\mathcal{P}$. It is easy to see that this is a linear programming with equality constraint, where the decision variables are a probability simplex over $\mathcal{X}$.

$$\min_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[h(\mathbf{x})], \quad \text{s.t. } \mathcal{P} = \left\{\mathbb{P}' : \boldsymbol{\omega}^\top \mathbb{E}_{\mathbb{P}'}[\mathbf{x}] = C_{\boldsymbol{\omega}} - \beta, \mathbb{P}' \in \Delta(\mathcal{X})\right\}. \tag{26}$$

Inside the optimization, for every $(\boldsymbol{\omega}, \beta) \in \mathcal{G}^{\text{lin}}$, our problem R-SO has one more induced Linear programming to solve compared with the standard problem SO.

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 SO (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 $\mathbb{P}(\mathbf{e})$ 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 $h(\mathbb{P}(\mathbf{e}))$. When $h(\mathbb{P}(\mathbf{e})) = \|\mathbf{e}\|_0$, solving the problem SO is then NP-hard. □