Bounded Rationality and Robust Mechanism Design: An Axiomatic Approach

By LUYAO ZHANG AND DAN LEVIN

Online Appendices

I. Appendix A

The crucial part of the proof is that (i) implies (ii). First, since Axiom 1, 4, 5 implies von Neumann-Morgenstern’s three axioms on lotteries, it follows directly from their theory (and the fact that $F^c$ and $Z$ are isomorphic) that there exists an affine function $u: Z \rightarrow R$, such that for all $p, q \in F^c: p \succeq q$ iff $u(p) \geq u(q)$. Moreover, $u$ is cardinally unique. By Axiom 2, $u$ is not a constant function. For any constant act $f \in F^c$, $V(f) = u(f)$, satisfying (1) for any $a(f) \in [0, 1]$. So $V(f)$ calibrates the preference on $F^c$.

For any $f \in F \setminus F^c$, pick constant acts $f^{\text{best}}, f^{\text{worst}} \in F^c$ that always generate the most and least preferred outcomes given $f$ is chosen. Formally, $f^{\text{best}} \in \{p|p \succeq q, \forall q \in C(f)\}$ and $f^{\text{worst}} \in \{h|h \preceq q, \forall q \in C(f)\}$. For $f \in F \setminus F^c$, by the definition of $F^c$, $f^{\text{best}} \sim f^{\text{worst}}$ which implies $u(f^{\text{best}}) = u(f^{\text{worst}})$ and by Axiom 2, $f \sim f^{\text{best}} \sim f^{\text{worst}}$. So $V(f) = u(f^{\text{best}}) = u(f^{\text{worst}})$ satisfying (1) for any $a(f) \in [0, 1]$. Hence $V(f)$ also calibrates the preference on $F^c$.

Finally, for $f \in F \setminus F^c$, by the definition of $F^e$, $f^{\text{worst}} \prec f^{\text{best}}$. And by Axiom 3, $f^{\text{worst}} \preceq f \preceq f^{\text{best}}$.

**LEMMA 1**: for $f \in F \setminus F^e$, Axiom 2-5 imply there exists a unique $\beta^* \in [0, 1]$ such that $f \sim \beta^* f^{\text{best}} + (1 - \beta^*) f^{\text{worst}}$.

**PROOF**:

First since $u[\beta f^{\text{best}} + (1 - \beta) f^{\text{worst}}] = \beta u(f^{\text{best}}) + (1 - \beta) u(f^{\text{worst}})$, so for $0 \leq a < b \leq 1$, $bf^{\text{best}} + (1 - b) f^{\text{worst}} \succeq af^{\text{best}} + (1 - a) f^{\text{worst}}$. Then it ensures that if $\beta^*$ exists, it is unique.

If $f \sim f^{\text{best}}$, then $\beta^* = 1$ works. The same way around, if $f \sim f^{\text{worst}}$, then $\beta^* = 0$ works. Otherwise, $f^{\text{worst}} \prec f \prec f^{\text{best}}$. Define

$$\beta^* = \sup\{\beta \in [0, 1] : f \succeq \beta f^{\text{best}} + (1 - \beta) f^{\text{worst}}\}.$$
Since $\beta = 0$ is in the set, we aren’t taking a sup over an empty set. By the definition of $\beta^*$ if $1 \geq \beta > \beta^*$, then $f \prec \beta f_{\text{best}} + (1 - \beta) f_{\text{worst}}$. Moreover, by the same argument to prove uniqueness above, if $0 \leq \beta < \beta^*$, then $f \succ \beta f_{\text{best}} + (1 - \beta) f_{\text{worst}}$. To see this, note that if $0 \leq \beta < \beta^*$, then there exists $\beta'$ such that $0 \leq \beta < \beta' \leq \beta^*$ and $f \succ \beta' f_{\text{best}} + (1 - \beta') f_{\text{worst}}$ by the definition of $\beta^*$. And $\beta < \beta'$ implies that $f \succ \beta' f_{\text{best}} + (1 - \beta') f_{\text{worst}} \succ \beta f_{\text{best}} + (1 - \beta) f_{\text{worst}}$. 

There are three possibilities to consider.

(1). Suppose $\beta^* f_{\text{best}} + (1 - \beta^*) f_{\text{worst}} \succ f \succ f_{\text{worst}}$, then by Axiom 5 there exists $b \in (0, 1)$ such that $b [\beta f_{\text{best}} + (1 - \beta^*) f_{\text{worst}}] + (1 - b) f_{\text{worst}} = \beta f_{\text{best}} + (1 - b \beta^*) f_{\text{worst}} \succ f$. But $\beta^* \prec \beta^*$, so by the previous argument $f \succ \beta^* f_{\text{best}} + (1 - b \beta^*) f_{\text{worst}}$. Contradiction.

(2). Suppose instead that $f_{\text{best}} \succ f \succ \beta^* f_{\text{best}} + (1 - \beta^*) f_{\text{worst}}$. Then by Axiom 5, there exists $a \in (0, 1)$ such that $f \succ a [\beta f_{\text{best}} + (1 - \beta^*) f_{\text{worst}}] + (1 - a) f_{\text{best}} = (1 - a(1 - \beta^*)) f_{\text{best}} + a(1 - \beta^*) f_{\text{worst}}$. Since $(1 - a(1 - \beta^*)) > \beta^*$, we have from above that $(1 - a(1 - \beta^*)) f_{\text{best}} + a(1 - \beta^*) f_{\text{worst}} \succ f$. Contradiction.

(3). This leaves us with the third possibility (which is what we want) namely that $f \sim \beta^* f_{\text{best}} + (1 - \beta^*) f_{\text{worst}}$.

Proof of lemma 1 ends.

Follows the argument of lemma 1, then $V (f) = V [\beta^* f_{\text{best}} + (1 - \beta^*) f_{\text{worst}}]$. Since $[\beta^* f_{\text{best}} + (1 - \beta^*) f_{\text{worst}}] \in F^c$,

$$V [\beta^* f_{\text{best}} + (1 - \beta^*) f_{\text{worst}}] = u [\beta^* f_{\text{best}} + (1 - \beta^*) f_{\text{worst}}]$$

Moreover, since $u$ is affine,

$$u [\beta^* f_{\text{best}} + (1 - \beta^*) f_{\text{worst}}] = \beta^* u (f_{\text{best}}) + (1 - \beta^*) u (f_{\text{worst}}).$$

Then, by the definition of $f_{\text{best}}$ and $f_{\text{worst}},$

$$\min_{p \in C(f)} u(p) = u (f_{\text{worst}}) < u(f_{\text{best}}) = \max_{p \in C(f)} u(p).$$

So

$$u [\beta^* f_{\text{best}} + (1 - \beta^*) f_{\text{worst}}] = \beta^* \max_{p \in C(f)} u(p) + (1 - \beta^*) \min_{p \in C(f)} u(p).$$

Then

$$V (f) = \beta^* \max_{p \in C(f)} u(p) + (1 - \beta^*) \min_{p \in C(f)} u(p).$$

So $\alpha(f) = \beta^*$ works and is uniquely determined.

II. Appendix B

$\implies$ If $s_x^*$ is an obviously dominant strategy, then by (2) and the obvious monotonicity axiom, (3) is satisfied.
(\iffalse) If (3) holds, assume by contradiction that \( s_i^* \) is not an obviously dominant strategy. Then there exists an information set \( I \in \partial I(s_i^*) \), a deviating strategy \( s_i' \in S_i(I)[s_i^*(I)]^c \) such that

\[
\inf_{(s_{-i}, r_n) \in [I]} u_i \left( s_i^*, s_{-i}, \omega_n \right) < \sup_{(s_{-i}, r_n) \in [I]} u_i \left( s_i', s_{-i}, \omega_n \right).
\]

Then we can find an obvious preference represented by (1) with \( \alpha(s_i^*) = 0 \) and \( \alpha(s_i') = 1 \) such that \( V(s_i^*) < V(s_i') \). So \( s_i^* \prec_{[I]} s_i' \). Contradiction.

III. Appendix C

Since \( u_i \left( s_i^*, s_{-i}^*, \omega_n \right) \geq \inf_{\omega_n \in \Omega_N} u_i \left( s_i^*, s_{-i}, \omega_n \right) \) and \( u_i \left( s_i', s_{-i}^*, \omega_n \right) \leq \sup_{\omega_n \in \Omega_N} u_i \left( s_i', s_{-i}, \omega_n \right) \)

for any \( s_i' \in S_i(I)[s_i^*(I)]^c \) and \( \omega_n \in \Omega_N \), (4) implies (5).