Intuitive criterion for signaling game equilibria

Signaling games tend to have many sequential equilibria. Some of them can be made to seem implausible via **forward induction** arguments, arguments of the form "player 2 should know that if he sees someone have beer for breakfast, it probably is the surly type, as the wimpy type could not possibly improve his payoff by switching to beer. Therefore, player 2 should not fight someone who has beer for breakfast, and therefore an equilibrium in which both types have quiche for breakfast makes no sense, as the surly type will want to deviate".

There are several different equilibrium refinements that formalize this logic. We will focus on the one most commonly applied, the **intuitive criterion**.

- Start with a sequential equilibirum of a signaling game, (σ, μ). You would find these equilibria though usual methods.
- If every message is played by some type in the equilibrium, the intuitive criterion is satisfied.
- If there is an unused message, eliminate all types whose equilibrium payoff $u(\sigma)$ is greater than *any* conceivable payoff from switching to the unused message.
- For the remaining types, determine the set of conceivable best responses for player 2 should the unused message be sent by one of the remaining types.
- Iterate these previous two steps as necessary until no more types or strategies are dominated.
- If there is at least one type whose equilibrium payoff $u(\sigma)$ is less than the lowest conceivable payoff from the unused message, after removing all dominated strategies for 2, (σ, μ) fails the intuitive criterion.
- If it is at least possible that switching to the unused message could lower a type's payoff from his equilibrium payoff $u(\sigma)$, and if this is true for every unused message, the equilibrium satisfies the intuitive criterion.

Mathematically, for a sequential equilibrium (σ, μ), let $u_1^*(t)$ be the equilibrium payoff received by type t. For each unused message m, let

$$D(m) = \{t(m) : u_1^*(t) > \max_{r \in BR(T(m),m))} u_1(t, m, r)\} \text{ be the set of types for whom message } m \text{ is dominated}$$

If, for some unused message m and some type t,

$$u_1^*(t) < \min_{r \in B_2(T(m) \setminus D(m),m)} u_1(t, m, r)\}$$

then (σ, μ) fails the intuitive criterion. For simplicity, the mathematical definition above does not mention iteration, but this is not hard to add.