Signaling Games

A *signaling* (or *sender-receiver*) game is a two-player game of incomplete information in which players interact as described by the following stages:

1. Nature selects the type of player 1.
2. Player 1 observes his type and then chooses an action.
3. Player 2 observes the action taken by player 1 and then chooses his own action.
4. Payoffs are realized.

Thus, the primitives of a signaling game are $(T_1, A_1, A_2, u_1, u_2, p)$, where

- T_1 is the set of possible types of player 1;
- $p \in \triangle T_1$ is the probability distribution according to which nature selects the type of player 1;
- A_1 and A_2 are the sets of actions of players 1 and 2, respectively;
- u_1 and u_2, are the players' payoff functions, where $u_i : T_1 \times A_1 \times A_2 \to \mathbb{R}$.

In a signaling game, a mixed strategy for player 1 is a mapping

$$\sigma_1 : T_1 \to \triangle A_1,$$

and a mixed strategy for player 2 is a mapping

$$\sigma_2 : A_1 \to \triangle A_2.$$

For $t_1 \in T_1$ and $a_1 \in A_1$ we write $\sigma_1(a_1 \mid t_1)$ for the probability with which player 1 takes action a_1 when his type is t_1. Likewise, for for $a_1 \in A_1$ and $a_2 \in A_2$ we write $\sigma_2(a_2 \mid a_1)$ for the probability with which player 2 takes action a_2 after observing that player 1 took action a_1.

Assume that the sets T_1, A_1, and A_2 are finite, and players select their actions according the the strategies (σ_1, σ_2). The payoff to player 1 if his type is $t_1 \in T_1$ and he takes action $a_1 \in A_1$ is

$$U_1(a_1, \sigma_2 \mid t_1) = \sum_{a_2 \in A_2} \sigma_2(a_2 \mid a_1) u_1(a_1, a_2, t_1).$$

Likewise, the payoff to player 1 if he takes action $a_2 \in A_2$ when he observed action $a_1 \in A_1$ is

$$U_2(a_2 \mid a_1) = \sum_{t_1 \in T_1} \mu(t_1 \mid a_1) u_2(a_1, a_2, t_1),$$

where $\mu(t_1 \mid a_1)$ is the probability with which player 2 beliefs player 1's type is t_1 after observing action a_1.

A solution concept appropriate for signaling games is *Perfect Bayesian Nash equilibrium* (PBNE).

A profile (σ_1^*, σ_2^*) is a PBNE if it satisfies:

PBNE.1. For all $t_1 \in T_1$ and $a_1, a_1' \in A_1$

$$\sigma_1^*(a_1 \mid t_1) > 0 \Rightarrow U_1(a_1, \sigma_2^* \mid t_1) \geq U_1(a_1', \sigma_2^* \mid t_1).$$

PBNE.2. For all $a_1 \in A_1$ and $a_2, a_2' \in A_2$

$$\sigma_2^*(a_2 \mid a_1) > 0 \Rightarrow U_2(a_2 \mid a_1) \geq U_1(a_2' \mid a_1).$$

PBNE.3. For all $\bar{t}_1 \in T_1$ and $a_1 \in A_1$

$$\sum_{t_1 \in T_1} p(t_1)\sigma_1^*(a_1 \mid t_1) > 0 \Rightarrow \mu(\bar{t}_1 \mid a_1) = \frac{p(\bar{t}_1)\sigma_1^*(a_1 \mid t_1)}{\sum_{t_1 \in T_1} p(t_1)\sigma_1^*(a_1 \mid t_1)}.$$

Conditions PBNE.1 and PBNE.2 require, respectively, that players 1 and 2 behave optimally. Condition PBNE.3 requires that the beliefs of player 2 about the type of player 1 be consistent with the behavior of player 1.