Trembling-hand Perfect (THP) Equilibria
Felipe Balmaceda
Microeconomics II

1 Definition

In a Nash equilibrium each player's equilibrium strategy is a best response to the other player's equilibrium strategies. In a Trembling Hand Perfect (THP) equilibrium, there must also be arbitrarilly small perturbations of all players' strategies such that every pure strategy gets strictly positive probability and each player's equilibrium strategy is still a best response to the other players' perturbed strategies. The definition of a THP equilibrium is the following.

Definition 1 Strategy profile σ is a trembling hand perfect (THP) equilibrium if there exists a sequence of totally mixed strategy profiles $\sigma^n \to \sigma$ such that, for all i,

$$u_i\left(\sigma_i, \sigma_{-i}^n\right) \geq u_i\left(s_i, \sigma_{-i}^n\right) \text{ for all } s_i \in S_i.$$

2 Intuition

In a THP equilibrium, the optimality of a player's strategy choice does not depend on an assumption that some pure strategies are getting zero probability in an equilibrium. Thus, THP helps to get rid of some strange equilibria, such as (T,L) in the example below, in which a player is playing a weakly dominated strategy.

3 Example

Consider the following normal form game

	L	R
U	$2,2$	$2,2$
D	$1,0$	$3,1$

This game has two N.E. in pure strategies: (U,R) and (D,R) and a continuum of mixed strategy equilibria: player 1 plays U and player 2 randomizes with $\sigma_2(L) \geq \frac{1}{2}$.
Observe that that in the N.E. (U,L), player 2 plays a weakly dominated strategy.
Consider first the equilibria (D,R), where $\sigma_1(U) = 0$ and $\sigma_2(L) = 0$. To prove that this is THP, we need to construct a totally mixed strategy that converges to $\boldsymbol{\sigma}$. Let $\boldsymbol{\sigma}^\varepsilon$ be the totally mixed strategy profile: $\sigma_1^\varepsilon(D) = 1-\varepsilon$ and $\sigma_2^\varepsilon(R) = 1-\varepsilon$.
Observe that as $\varepsilon \to 0$, $\boldsymbol{\sigma}^\varepsilon \to \boldsymbol{\sigma}$.
Next, observe that

$$u_1\left(U, \sigma_2^\varepsilon\right) = 2 \text{ and } u_1\left(D, \sigma_2^\varepsilon\right) = 1\varepsilon + 3(1-\varepsilon) = 3 - 2\varepsilon.$$

Thus, $u_1(D, \sigma_2^\varepsilon) \geq u_1(U, \sigma_2^\varepsilon)$ if and only if $3 - 2\varepsilon \geq 2 \Rightarrow \varepsilon \leq \frac{1}{2}$. Thus, there is a sequence in which $\varepsilon = \frac{1}{2n}$ in which $\sigma_1(D) = 1$ is a best response when player 2 plays the totally mixed strategy σ_2^ε.
Next, observe

$$u_2(\sigma_1^\varepsilon, L) = 2\varepsilon + 0(1-\varepsilon) \text{ and } u_2(\sigma_1^\varepsilon, R) = 2\varepsilon + 1(1-\varepsilon) = 1 + \varepsilon.$$

Thus, $u_2(\sigma_1^\varepsilon, R) \geq u_2(\sigma_1^\varepsilon, L)$ if and only if $1 + \varepsilon \geq 2\varepsilon \Rightarrow \varepsilon \leq 1$. Thus, there is a sequence in which $\varepsilon = \frac{1}{2n}$ in which $\sigma_2(R) = 1$ is a best response when player 1 plays the totally mixed strategy σ_1^ε.
Consider next the equilibria (U,L), where $\sigma_1(U) = 1$ and $\sigma_2(L) = 1$. To prove that this is THP, we need to construct a totally mixed strategy that converges to $\boldsymbol{\sigma}$. Let $\boldsymbol{\sigma}^\varepsilon$ be the totally mixed strategy profile: $\sigma_1^\varepsilon(U) = 1-\varepsilon$ and $\sigma_2^\varepsilon(L) = 1-\varepsilon$. Observe that as $\varepsilon \to 0$, $\boldsymbol{\sigma}^\varepsilon \to \boldsymbol{\sigma}$.

$$u_1(U, \sigma_2^\varepsilon) = 2 \text{ and } u_1(D, \sigma_2^\varepsilon) = 1(1-\varepsilon) + 3\varepsilon = 1 + 2\varepsilon.$$

Thus, $u_1(U, \sigma_2^\varepsilon) \geq u_1(D, \sigma_2^\varepsilon)$ if and only if $2 \geq 1 + 2\varepsilon \Rightarrow \varepsilon \leq \frac{1}{2}$. Thus, there is a sequence in which $\varepsilon = \frac{1}{2n}$ in which $\sigma_1(U) = 1$ is a best response when player 2 plays the totally mixed strategy σ_2^ε.
Next, observe

$$u_2(\sigma_1^\varepsilon, L) = 2(1-\varepsilon) + 0(1-\varepsilon) \text{ and } u_2(\sigma_1^\varepsilon, R) = 2(1-\varepsilon) + 1\varepsilon = 2 - \varepsilon.$$

Thus, $u_2(\sigma_1^\varepsilon, L) \geq u_2(\sigma_1^\varepsilon, R)$ if and only if $2(1-\varepsilon) \geq 2 - \varepsilon \Rightarrow \varepsilon \leq 0$. Thus, there is no sequence in which $\sigma_2(L) = 1$ is a best response when player 1 plays the totally mixed strategy σ_1^ε.
In fact note that as long as player 1 plays D with positive probability, player 2 is not willing to play L with positive probability and thus no mix strategy equilibrium in which $\sigma_2(L) > 0$ is a THP equilibrium