1. Modern Games Theory

1.1. Normal Form Games

Definition 1.1.1 (Normal Form Game): A *Norma Form Game* is a tuple $(\mathcal{N}, \boldsymbol{A}, \boldsymbol{u})$, where $\mathcal{N}$ is finite set of *players*. Then $\boldsymbol{A} = \times_{i \in \mathcal{N}} A_i$ are tuples of *actions* of all players, where A_i are actions of player i. The $\boldsymbol{u} = (u_i)_{i \in \mathcal{N}}$ are players *utility functions* such that $u_i : \boldsymbol{A} \to \mathbb{R}$.

If there are only two players we can fully describe the game with a table where each row and column represents the actions of player 1 and player 2 respectively and the value in the cell is the tuple of utility of the players.

Definition 1.1.2 (Constant sum game): The normal form game $(\mathcal{N}, \boldsymbol{A}, \boldsymbol{u})$ is called *constant sum game* if for all $\boldsymbol{a} \in \boldsymbol{A}$ holds that $\sum_{i \in \mathcal{N}} u_i(\boldsymbol{a}) = c$, for some constant $c \in \mathbb{R}$.

Definition 1.1.3 (Zero sum game): The constant sum game $(\mathcal{N}, \boldsymbol{A}, \boldsymbol{u})$ is called *zero sum game* if for all $\boldsymbol{a} \in \boldsymbol{A}$ holds that $\sum_{i \in \mathcal{N}} u_i(\boldsymbol{a}) = 0$.

Definition 1.1.4 (Pure Strategy): We will call the action $a_i \in A_i$ a *pure strategy* of player i.

Definition 1.1.5 (Mixed Strategy): The *mixed strategy* of player i is a probability distribution $\pi_i \in \Delta(A_i)$ over the set of actions A_i.

Definition 1.1.6 (Strategy Profile): The *strategy profile* is a tuple $\boldsymbol{\pi} = (\pi_i)_{i \in \mathcal{N}}$, where $\pi_i \in \Delta(A_i)$ is the mixed strategy of player i.

We will use the symbol $\boldsymbol{\pi}_{-i}$ to denote the strategy profile of all players except player i.

Since all players are choosing their actions independently of others, the probability probability of action $\boldsymbol{a} \in \boldsymbol{A}$ happening, given strategy profile $\boldsymbol{\pi}$, is $P^{\boldsymbol{\pi}}(\boldsymbol{a}) = \prod_{i \in \mathcal{N}} \pi_i(a_i)$. Thus the expected value for player i is $R_i^{\boldsymbol{\pi}} = \sum_{\boldsymbol{a} \in \boldsymbol{A}} \pi_i(\boldsymbol{a}) u_i(\boldsymbol{a})$.

Definition 1.1.7 (Best Response): The *best response* of player i to the strategy profile $\boldsymbol{\pi}_{-i}$ of all players except player i is $\arg\max_{\pi_i \in \Delta(A_i)} R_i^{\pi_i, \boldsymbol{\pi}_{-i}}$. We will use $\mathbb{BR}_i(\boldsymbol{\pi}_{-i})$ to denote the set of all best responses of player i to the strategy profile $\boldsymbol{\pi}_{-i}$.

Note that in two player zero-sum games, player maximizing his expected utility is minimizing the expected utility of the other player.

$$\arg\max_{\pi_i \in \Delta(A_i)} R_i^{\pi_i, \boldsymbol{\pi}_{-i}} = \arg\min_{\pi_i \in \Delta(A_i)} R_{-i}^{\pi_i, \boldsymbol{\pi}_{-i}}$$

Definition 1.1.8 (Support): The *support* of the mixed strategy π_i is the set of actions with non-zero probability $\{a_i \in A_i \mid \pi_i(A_i) > 0\}$.

Lemma 1.1.1 (Best Response Lemma): For any best response strategy $\pi_i \in \mathbb{BR}_i(\boldsymbol{\pi}_{-i})$ to the strategy profile $\boldsymbol{\pi}_{-i}$ it holds that the actions in the support of π_i have the same expected value.

Proof: If this was not the case, then there would be two actions that have different expected value. Player i would then be able to increase his expected value by moving the probability from the action with lower expected value to the action with higher expected value. This would mean that the strategy π_i is not the best response to the strategy profile $\boldsymbol{\pi}_{-i}$. □

Lemma 1.1.2 (Best Reponse Set is Convex): The set $\mathbb{BR}_i(\boldsymbol{\pi}_{-i})$ is a convex set.

Proof:

Let $\pi_i^a, \pi_i^b \in \mathbb{BR}_i(\boldsymbol{\pi}_{-i})$ be two best response strategies to the strategy profile $\boldsymbol{\pi}_{-i}$ and $\pi_i^c = \lambda\pi_i^a + (1-\lambda)\pi_i^b$ be their convex combination. It simply follows that

$$\begin{aligned} R_i^{\pi_i^c, \boldsymbol{\pi}_{-i}} &= \sum_{\boldsymbol{a} \in \boldsymbol{A}} \left(\lambda\pi_i^a(a_i) + (1-\lambda)\pi_i^b(a_i)\right)\boldsymbol{\pi}_{-i}(\boldsymbol{a}_{-i}) \\ &= \lambda \sum_{\boldsymbol{a} \in \boldsymbol{A}} \pi_i^a(a_i)\boldsymbol{\pi}_{-i}(\boldsymbol{a}_{-i}) + (1-\lambda) \sum_{\boldsymbol{a} \in \boldsymbol{A}} \pi_i^b(a_i)\boldsymbol{\pi}_{-i}(\boldsymbol{a}_{-i}) \\ &= \lambda R_i^{\pi_i^a, \boldsymbol{\pi}_{-i}} + (1-\lambda) R_i^{\pi_i^b, \boldsymbol{\pi}_{-i}}. \end{aligned}$$

It is clear that for each strategy $\pi \in \mathbb{BR}_i(\boldsymbol{\pi}_{-i})$ we get the same expected value so even the convex combination of them is the same. In fact this should be proven for every finite set of strategies, but it simply follows using induction. □

Definition 1.1.9 (Nash Equilibrium): The strategy profile $\boldsymbol{\pi}$ forms a *Nash Equilibrium* if:

$$\forall i \in \mathcal{N}, \forall \pi_i^a \in \Delta(A_i) : R_i^{\pi_i, \boldsymbol{\pi}_{-i}} \geq R_i^{\pi_i^a, \boldsymbol{\pi}_{-i}}$$

In other words for no player it is beneficial to deviate from their strategy in the profile $\boldsymbol{\pi}$.

Definition 1.1.10 (Dominance): We say that the strategy π_i^a *strictly dominates* the strategy π_i^b if for all $\boldsymbol{\pi}_{-i}$ holds that $R_i^{\pi_i^a,\boldsymbol{\pi}_{-i}} > R_i^{\pi_i^b,\boldsymbol{\pi}_{-i}}$. The strategy π_i^a *weakly dominates* the strategy π_i^b if for all $\boldsymbol{\pi}_{-i}$ holds that $R_i^{\pi_i^a,\boldsymbol{\pi}_{-i}} \geq R_i^{\pi_i^b,\boldsymbol{\pi}_{-i}}$ and for at least one $\boldsymbol{\pi}_{-i}$ the inequality is strict. Also the strategy π_i^a is *weakly/strongly dominated* if there exists a strategy π_i^b such that π_i^b weakly/strongly dominates π_i^a. The strategies π_i^a, π_i^b are intransitive if one neither dominates nor is dominated by the other.

Observation 1.1.1: In two players game a strategy π_i^a of player i, is weakly dominated by strategy π_i^b if and only if for every pure strategy $\boldsymbol{\pi}_{-i}$ of the oponent the expected reward $R_i^{\pi_i^b,\boldsymbol{\pi}_{-i}} \geq R_i^{\pi_i^a,\boldsymbol{\pi}_{-i}}$ and for at least one it holds that $R_i^{\pi_i^b,\boldsymbol{\pi}_{-i}} > R_i^{\pi_i^a,\boldsymbol{\pi}_{-i}}$.

```
GetWeaklyDominatedPureStrategy(game, player):
for $\pi^a$ in $A_i$:                          // Where $A_i$ are actions of player
    for $\pi^b$ in $A_i \setminus \{\pi^a\}$:        // We are checking only pure strategies
        if $\pi^b$ weakly dominates $\pi^a$:    // Based on Observation 1.1.1
            return $\pi^a$
return fail
```

We can see that if we are checking that the strategy π^a is dominated only by pure strategies, we can be possibly missing cases where the strategy π^a is dominated by mixed strategy. As can be seen in the following example[1]:

$$\begin{pmatrix} (3,-1) & (-1,1) \\ (0,0) & (0,0) \\ (-1,0) & (2,-1) \end{pmatrix}$$

In this scenario there is no pure strategy that dominates another pure strategy. But it is obious that the strategy $\pi_1^a = \left(\frac{1}{2}, 0, \frac{1}{2}\right)$ strictly dominates the strategy $\pi_2^b = (0,1,0)$. Thus the function GetWeaklyDominatedPureStrategy is not sufficient to find all dominated strategies, but it is good enough for us.

```
IteratedEliminationOfDominatedStrategies(game):
while $s \leftarrow$ GetWeaklyDominatedPureStrategy (game, player 1 or player 2):
    remove $s$ from game
if only one strategy for each player left in game:
    return the only left pure strategies left for player 1 and 2
else:
    return fail
```

[1] https://en.wikipedia.org/wiki/Strategic_dominance#Iterated_elimination_of_strictly_dominated_strategies_(IESDS)

The strategies returned by the function IteratedEliminationOfDominatedStrategies or Ieds will be *weakly dominant* for each player. The *weakly dominant* strategy will be a Nash equilibrium, but beware that it does not have to be the only one. Consider the following minimal example[2]:

$$\begin{pmatrix} (1,1) & (0,0) \\ (0,0) & (0,0) \end{pmatrix}$$

If we will consider the pure strategy profiles $((1,0),(1,0))$ and $((0,1),(0,1))$, then both of those are Nash equilibria, but only the first one is weakly dominant.

1.2. Extensive Form Games

To use the formalism of the *extensive form games* we will need the following definitions.

Let $\mathcal{N} = \{1, 2, ..., n\}$ be a finite set of *players* and also let $\mathcal{N}_c = \mathcal{N} \cup \{c\}$ bet the set of players and a *chance* player c. The finite set $\mathcal{H}$ will denote the set of *histories* representing sequences of actions. It holds that $\forall(h,a) \in \mathcal{H} \Rightarrow h \in \mathcal{H}$ and we call a an *action*. We also use $h' \sqsubseteq h$ to denote that h' is equal to or the prefix of h. The terminal histories $\mathcal{Z}$ are histories, that are not prefixes of any other history. The set of actions $\mathcal{A}(h) = \{a \mid (h,a) \in \mathcal{H}\}$ is the set of actions available at a non-terminal history h. The function $p : \mathcal{H} \setminus \mathcal{Z} \to \mathcal{N}_c$ assigns each non-terminal history a player or the chance player. It creates a partition over the non-terminal histories. We will denote each of the partitions $\mathcal{H}_i$ as *histories of player* i. The function $\pi_c : h \in \mathcal{H}_c \mapsto \Delta(\mathcal{A}(h))$ which associates with each history of the chance player a probability distribution over the available actions. Let the *information partition* be $\mathcal{I} = (\mathcal{I}_i)_{i\in\mathcal{N}}$, where for each player i the set $\mathcal{I}_i$ is a partition of the player histories $\mathcal{H}_i$. The set $I \in \mathcal{I}$ is the *information set*. Each two histories $h, h' \in I$ are indistinguishable for player i, so $\mathcal{A}(h) = \mathcal{A}(h')$ and we use $\mathcal{A}(I)$ to denote the set of actions available at information set I. Let $\boldsymbol{u} = (u_i)_{i\in\boldsymbol{\pi}}$, where $u_i : \mathcal{Z} \to \mathbb{R}$ is the utility function of player i.

The *extensive form game* is defined as the following tuple $(\mathcal{H}, \mathcal{Z}, \mathcal{A}, \mathcal{N}, p, \pi_c, \boldsymbol{u}, \mathcal{I})$.

For the following discussion it needs to be mentioned that game in this form, forms a tree since the finiteness condition on the histories enforces the graph of the game to be acyclic.

The *behavioral strategy* of player i is a mapping $\pi_i : I \in \mathcal{I}_i \mapsto \Delta(\mathcal{A}(I))$ that for each information set $I \in \mathcal{I}_i$ assigns a probability distribution over the available actions at I. We note that the π_c is a behavioral strategy of the chance player. The *behavioral profile* is $\boldsymbol{\pi} = (\pi_i)_{i\in\mathcal{N}}$. It is freely extended with the chance player's strategy π_c when needed and in that case we denote the behavioral profile as $\boldsymbol{\pi}_c$. We use the notation $\pi_i(I,a)$ to denote the probability of action a at information set I under the behavioral policy π_i. Also we use $\boldsymbol{\pi}(I,a) = \pi_{p(I)}(I,a)$.

The *reach probability* of a history $h \in \mathcal{H}$ given behavioral profile $\boldsymbol{\pi}$ is $P^{\boldsymbol{\pi}}(h) = \prod_{(h',a)\sqsubseteq h} \boldsymbol{\pi}(h',a)$. With this we are able to define the *expected utility* of player i given behavioral profile $\boldsymbol{\pi}$ as $R_i^{\boldsymbol{\pi}} = \sum_{h\in\mathcal{Z}} P^{\boldsymbol{\pi}}(h) u_i(h)$.

We will also introduce the notion of *sequence* for each player i the $\sigma_i(h)$, which is the sequence of actions of player i to a history $h \in \mathcal{H}$, disregarding the actions of the other players.

Player i has *perfect recall* if for each two histories $h, h' \in I, I \in \mathcal{I}_i$ it holds that $\sigma_i(h) = \sigma_i(h')$. In a game of *perfect recall* each player has perfect recall. Since each non-terminal history is assigned a specific player we will use σ_h instead of $\sigma_i(h)$ freely. Also in game of perfect recall for each $h, h' \in I, \sigma_h = \sigma_{h'}$ so we will use σ_I to denote the sequence of actions at information set I. It is clear

[2] https://en.wikipedia.org/wiki/Strategic_dominance#Dominance_and_Nash_equilibria

that in such games each sequence is uniquely determined by the last move in the last information set I, so $\sigma = \sigma_I a$. This also holds for any non-empty history $h \in \mathcal{H}, \sigma_h = \sigma_I a$ for some information set I.

From that we can define the *realization probability* $x_\pi(\sigma_h)$ given a behavioral policy π as $x_\pi(\sigma_h) = \prod_{\sigma_{h'} a \sqsubseteq \sigma_h} \pi(h', a)$. It is clear that such function must satisfy the folllowing constraints $x(\emptyset) = 1$ and $x_\pi(\sigma_h) = \sum_{a \in \mathcal{A}(h)} x(\sigma_h a)$. Function satisfying these constraints also gives us a the behavioral policy π, since $\pi(h, a) = \frac{x_\pi(\sigma_{ha})}{x_{\pi(\sigma_h)}}$ is a well defined probability distribution over actions $a \in \mathcal{A}(h)$ for each history $h \in I, I \in \mathcal{I}_i$. Also the reach probability has simple equivalent in the realization probabilities, $P^\pi(h) = \prod_{i \in \mathcal{N}_c} x_{\pi_i}(\sigma_h)$. From this we can see that the expected utility of player i may be written as

$$
\begin{aligned}
R_i^{\boldsymbol{\pi}} &= \sum_{h \in \mathcal{Z}} P^{\boldsymbol{\pi}}(h) u_i(h) \\
&= \sum_{h \in \mathcal{Z}} \left(\prod_{j \in \mathcal{N}_c} x_{\pi_j}(\sigma_h) \right) u_i(h).
\end{aligned}
$$

To simplify some of the future equations we will use the following notation, in the following form corresponds to the *counterfactual reach* for player i, given a historyr h.

$$
P^{\boldsymbol{\pi}_{-i}}(h) = \prod_{j \in \mathcal{N}_c \setminus \{i\}} x_{\pi_j}(\sigma_h).
$$

So we can write $P^{\boldsymbol{\pi}}(h) = x_{\pi_i}(\sigma_h) P^{\boldsymbol{\pi}_{-i}}(h)$.

We are now interested in averaging of two policies π_i^a and π_i^b for player i. The property that we want from the averaging is for any behavioral profile $\boldsymbol{\pi}$ to hold $R_i^{\pi_i^c, \boldsymbol{\pi}_{-i}} = 0.5 R_i^{\pi_i^a, \boldsymbol{\pi}_{-i}} + 0.5 R_i^{\pi_i^b, \boldsymbol{\pi}_{-i}}$. With behavioral policies this is tricky to do right, but with realization probabilities it is fairly simple. We will create a new averaged policy π_i^c from policies π_i^a and π_i^b of player i as policy π_i^c, for which the following holds $x_{\pi_i^c}(\sigma_h) = 0.5 x_{\pi_i^a}(\sigma_h) + 0.5 x_{\pi_i^b}(\sigma_h)$. Since this satisfy the consraints of the realization probabilities it also represents a behavioral policy. We can quickly inquire that the required property really holds for any behavioral profile $\boldsymbol{\pi}$.

$$
\begin{aligned}
R_i^{\pi_i^c, \boldsymbol{\pi}_{-i}} &= \sum_{h \in \mathcal{Z}} P^{\boldsymbol{\pi}}(h) u_i(h) \\
&= \sum_{h \in \mathcal{Z}} x_{\pi_i^c}(\sigma_h) P^{\boldsymbol{\pi}_{-i}}(h) u_i(h) \\
&= \sum_{h \in \mathcal{Z}} \left(\frac{1}{2} x_{\pi_i^a}(\sigma_h) + \frac{1}{2} x_{\pi_i^b}(\sigma_h) \right) P^{\boldsymbol{\pi}_{-i}}(h) u_i(h) \\
&= \frac{1}{2} R_i^{\pi_i^a, \boldsymbol{\pi}_{-i}} + \frac{1}{2} R_i^{\pi_i^b, \boldsymbol{\pi}_{-i}}
\end{aligned}
$$

We will represent $v_i^{\boldsymbol{\pi}}(h) = \sum_{ha, a \in \mathcal{A}(h)} \boldsymbol{\pi}(h, a) v_i^{\boldsymbol{\pi}}(ha)$ for non-terminal history h and $v_i^{\boldsymbol{\pi}}(h) = u_i(h)$ for terminal ones. We could also write $v_i^{\boldsymbol{\pi}}(h) = \sum_{z \in \mathcal{Z}} P^{\boldsymbol{\pi}}(z \mid h) u_i(z)$. Where $P^{\boldsymbol{\pi}}(z \mid h)$ is the probability of reaching the terminal history z given we are at history h and we are following the behavioral profile $\boldsymbol{\pi}$. It is easy to see that $R_i^{\boldsymbol{\pi}} = v_i^{\boldsymbol{\pi}}(\emptyset)$. We will call $v_i^{\boldsymbol{\pi}}(h)$, the *expected future reward* under behavioural profile $\boldsymbol{\pi}$ to player i, given the history h.

We focus on a games of a perfect recall. For a player i let $I \in \mathcal{I}_i$, π_i be theirs behavioral policy and $\boldsymbol{\pi}_{-i}$ be the behavioral profile of the other players. Set $\boldsymbol{\pi} = (\pi_i, \boldsymbol{\pi}_{-i})$. It is clear that the contribution of the state h to the value $v_i^{\boldsymbol{\pi}}(\emptyset)$ is $P^{\boldsymbol{\pi}}(h) v_i^{\boldsymbol{\pi}}(h)$. Also since we are in game of perfect recall it holds that each two histories in the information set I have disjoint subtrees, meaning that their contribution to the value $v_i^{\pi}(\emptyset)$ is independent of each other and we can sum them up. Since for each $h, h' \in I$ it holds that $\sigma_h = \sigma_{h'}$, we get that $x_\pi(\sigma_h) = x_\pi(\sigma_{h'})$ and we will use $x_\pi(\sigma_I)$ to denote that realization probability of information set I. So the contribution of the information set I to the value $v_i^{\pi}(\emptyset)$ is

$$\begin{aligned} P^{\pi}(h)v_i^{\pi}(I) &= \sum_{h \in I} P^{\pi}(h)v_i^{\pi}(h) \\ &= \sum_{h \in I} x_{\pi_i}(\sigma_h)P^{\pi_{-i}}(h)v_i^{\pi}(h) \\ &= x_{\pi_i}(\sigma_I)\sum_{h \in I} P^{\pi_{-i}}(h)v_i^{\pi}(h). \end{aligned}$$

.

That means that each history in the information set I contributes equally to the value $v_i^{\pi}(\emptyset)$ from the point of view of player i, since he can't make any of the two histories more likely to be reached. We can also see that if we want to maximize the players i expected future reward at information set I we can maximize the two values $x_{\pi_i}(\sigma_I)$ and $\sum_{h \in I} P^{\pi_{-i}}(h)v_i^{\pi}(h)$ separately. We note that the second value is dependent only on the subtree induced by the information set I, the behavioral profile π_{-i} of the other players and player i actions in following histories. Thus if we want to find the best policy for player i we can form it bottom up, by first finding the best policy at the lower information sets and then using the best policies at the lower information sets to find the best policy at the higher information sets.

We will also use the $q_i^{\pi}(h, a)$, which is the expected future reward of player i given the history h and choosing the action a under the behavioral profile π. It is defined as $q_i^{\pi}(h, a) = v_i^{\pi}(ha)$, which would not be in itself interesting, but in the perfect information games we can define it for the information set $I \in \mathcal{I}_i$ as $q_i^{\pi}(I, a) = \sum_{h \in I} q_i^{\pi}(h, a)$. Now when deciding what action is the best response in information set I to the behavioral profile π is we can succintly ask $a^\star = \arg\max_{\{a \in \mathcal{A}(I)\}} q_i^{\pi}(I, a)$.

1.2.1. Counterfactual regret

Now we have already defined the counterfactual reach $P^{\pi_{-i}}$, using this we can ask the following question. Given that player i tries to reach the information set $I \in \mathcal{I}_i$, what is his expected future reward at this information set I? And what would be his expected future reward if he was to choose the action a at this information set I? The first question is answered by $v_{i,c}^{\pi} =$