Charles University in Prague

Faculty of Social Science

Institute of Economic Studies

Bachelor Thesis

Duopoly modelling with agent-based computational economics

Author: Mgr. Stanislav Kovalčin
Consultant: PhDr. Petr Švarc
Academic year: 2008/2009

I hereby declare that I have elaborated this thesis on my own and that I have used only the sources listed.

Prague, 25 May 2008 Stanislav Kovalčin

Acknowledgment

I am deeply indebted to my supervisor PhDr. Petr Švarc, who revealed me interesting topic of agent-based computational economics, gave me lot of important and helpful impulses, was always free to help me and without who I would not be able to successfully finish this topic.

I would like to thank to my parents and my whole family as well, for unlimited support, love and patience they were dealing me with in hard times of making this thesis.

Title: Duopoly modelling with agent-based computational economics

Author: Mgr. Stanislav Kovalčin

Consultant: PhDr. Petr Švarc

Abstract: Existing duopoly models as Cournot duopoly or Stackelberg duopoly, when firms compete on quantities, does not explain in the real world observed phenomenon of collusive behaviour. We try to simulate and explain such behaviour with agent-based computational economics. Expansion of this model by adding possibility of endogenous timing of production is also examined and arise of simultaneous or sequential plays is observed. Both models are fully implemented by JAVA programming language and resulting data are analysed through graphical representation.

Keywords: agent-based computational economics, Cournot duopoly, Stackelberg duopoly, endogenous timing of production, collusive behaviour

Názov práce: Modelovanie duopolu pomocou agent-based výpočetnej ekonómie

Autor: Mgr. Stanislav Kovalčin

Konzultant: PhDr. Petr Švarc

Abstrakt: Existujúce modely duopolov ako Cournotov duopol anebo Stackelbergov duopol, kde si firmy konkurují množstvami, nevysvetľujú jav kolúzie, ktorý je pozorovaný v bežnom svete. Pokúšame se simulovat a vysvetliť takéto chovanie pomocou agent-based výpočetní ekonomie. Tento model je rozšírený o možnosť endogénneho načasovania produkcie a je skúmaný vznik simultánnych alebo sekvenčných ťahov. Oba modely sú plne naimplementované v programovacom jazyku JAVA and výsledné data sú analyzované pomocour grafickej reprezentácie.

Kľúčové slová: agent-based výpočetná ekonómia, Cuornotov duopol, Stackelbergov duopoly, endogénne načasovanie produkcie, kolúzia

Table of Contents

1. Introduction 6

1.1 Agent-based computational economics 7

1.2 Q-learning model 8

2. Duopoly models 12

2.1 Cournot duopoly 12

2.2 Stackelberg duopoly 15

2.4 Collusive behaviour 18

2.5 Endogenous timing in duopoly games 19

2.5.1 Extended game with observable delay 20

2.5.2 Extended game of action commitment 22

2.5.3 Conclusion of their research 23

3. The market with one period 25

3.1 Setup of computer simulation 26

3.2 Application of Q-learning 27

3.3 The agent with a memory 29

3.3.1 Result of the computer simulations 30

3.4 The agent without a memory 32

3.4.1 Result of the computer simulations 33

4 The market with two periods 37

4.1 The agent without a memory 38

4.1.1 Result of the computer simulations 39

4.2 The agent with a memory 42

4.2.1 Result of the computer simulations 44

5. Conclusion 52

Bibliography 54

1. Introduction

Economic models are created to capture and describe the outside economical world as good as possible. Nowadays, we can see nowadays a huge effort in mainstream economy to model and formalize economic situations quantitatively. However, this thesis is out of scope to answer the question whether this is a right approach or not. What we can say about this approach is that we need quite strict conditions to be able to solve models and to obtain reasonable results. Liberalization of these conditions leads to analytically intractable models.

These aforementioned conditions include perfect information, profit maximization and rational choices. These conditions create benign environment suitable for further analysis with mathematical analytical tools. Unfortunately, under these conditions we face the problem that we do not model reality, but rather a situation close to reality. Therefore we are sometimes experiencing situations in which theories predict outcomes that are not observed in the real world.

Nevertheless, research in recent years on the field of computer science brought us ability to simulate artificial societies, allowing us to analyze economic models in the mean of making simulations. This approach is called Agent-based Computational Economics (ACE).

Aim of the work is to present this modelling approach on the example of firms acting in a duopoly market. Structure of this thesis is following. After a brief explanation of Agent-based computational economics and Q-learning we chose to use as a learning reinforcement algorithm, we explain theoretical bases of duopoly and other models, used in this thesis, in the second section. The third and the fourth section is concerned about two different market models, they theoretical bases, implementation and analysis of results we obtained from simulations.

1.1 Agent-based computational economics

One of the leading researchers on the field of ACE, Leigh Tesfatsion claims ACE to be “the computational study of economic processes modelled as dynamic systems of interacting agents” [8]. Formally, ACE as a field of study is supposed to be somewhere in between cognitive science, computer science and evolutionary economics. It creates artificial societies of interacting agents and combines approaches and views from social science, economics and computer science.

Speaking about agents in ACE, we define agent as an individual entity with limited rationality and decision-making capabilities. At the beginning of the simulation, we have a population of agents with preset possible interactions, some behavioural rules and goals, which they are trying to achieve. This population is located into an environment with some characteristics. ACE simulate interactions between agents, looking at their adaptation to the other agents and environment, and observing output. Thus this basic inter-agent interaction creates the structure as agents are trying to achieve their goals. As we are not studying overall outcome, but observing at the very first stage of creation of economic model, we are able to figure out global patterns, interactions which are hidden in classical approach. ACE simulation can provide us an inner look in an equilibrium selection, which we are looking in this thesis.

One of the features of ACE is creation of modelled systems from the bottom up. It means that we are concerned about proper construction of agents in the first place and final system is made and defined by these agents. Secondly, we are able to relax some of the strict conditions provided by classical models. There are several aspects, in which conditions can be relaxed.

While classical economic models assume perfect information, profit maximization and rational choices, ACE can assume “bounded” rationality of agents. In classical models, there are usually entities interacting in economical models perfect information and their decision-making is made based on this perfect information. Clearly, this does not describe the real world perfectly, as we cannot have all information needed for our decision-making process. On the other hand, by so-called bounded rationality, we describe situation in which agents deal with “limited time, limited resources and incomplete information” and “the fact that decision making makers, often do not have the computational capabilities and the information needed to make a perfectly rational decisions” [1]. In the case of perfect information it means usually we do not have all information needed about market environment or moves and choices of agent acting on this market, whose can be taken into consideration as important for our decision. In the case of profit maximization it means that as we do not have perfect information about environment and other agents, it is difficult to know how to maximize profit. In the case of rational choices it means that as we have limited time, resources and incomplete information, we are not able to examine all available cases and decide which one is the best one.

As we say above, we can use different means of modelling interactions and decision-making process. While we are examining simple interaction of two agents in a duopoly market, our focus was reduced to the selection of appropriate mean of modelling the decision-making process. As they are many algorithms for modelling decision-making process, we chose Q-learning as Waltman and Kaymak [9] denote that it gives an interesting result of collusive behaviour of agents in a Cournot game in duopoly market. This behaviour is common in the real world, but it is predicted by almost no other model of learning behaviour of individual economic agents. Only exception is a trial-and-error model by [6].

1.2 Q-learning model

Q-learning is a reinforcement learning algorithm of agent’s behaviour and decision-making process developed by Watkins in 1989. It does not need a model of its environment and it is suitable for repetitive games against unknown opponents. Model is based on so-called Q-values and probabilities which are assigned to these Q-values. One of advantages of this model is that it can compare the expected outcome of available actions without having information about the external environment, which is suitable as normal economic actors rarely have complete information about the external environment.

Basically, Q-learning reinforcement model is based on two assumptions:

1. The agent knows the outcome for every possible strategy based on his previous experience with this strategy. This value is referred to as Q-value and in very simple way it can be described as a weighted average of outcomes the agent has obtained by choosing this strategy in the past. Moreover, the newest strategy is considered to be the most important one, whereas the oldest one is considered to be the least important one.
2. The agent probabilistically chooses an action he wants to play and the probability with which he chooses a certain action is dependent on Q-value of this action. This his decision making behaviour is modelled by a logit model, which is widely use in economic models.

Having these two assumptions, the agent's goal is to maximize his profit by choosing appropriate action in a current state. As we said, he is rewarded by outcome and this is only available information he gets from the external environment. He starts from the scratch and he has to learn and adapt to the best strategy he is able to find. What is important to us, is that as long as it is probabilistic he does not converge to the same result. Final results are similar, but often the same. But it gives a good representation of the real world, as we as human beings often face the situations in which we would make different decisions having an opportunity to solve this situation repeatedly. We know the two basic assumptions of Q-learning reinforcement model, now we can explain it more formally.

Fundamentally, problem is defined as a problem of action selection of the agent in repeated game in round t, so as the agent is in state s_t and he has to choose action a_t in the round t. So we have a finite set of available states of agent S and a finite set of agent's available actions A. By choosing action a_t

when the agent is in the state s_t, the agent moves into the new state s_{t+1}. Additionally, the agent is rewarded or punished by some outcome. Therefore we have function Q, which defines the Quality of state-action combination. This is where name Q-learning comes from. This function Q is defined as

$$Q : S \times A \rightarrow R$$

Therefore we call *Q(s,a)* as being Q-value for combination of state *s* and action *a*. At the beginning are all Q-values are set to some default, predefined value. In case of our use we put all values at the beginning to be zero.

After each move to a new state the agent is rewarded by some outcome, which can be positive or negative. Having this reward, Q-values are recalculated each round and for the next decision making the agent has a new set of Q-values. Therefore the core of Q-learning algorithm is a simple value update by iteration. Update rule is given by

$$Q_{t+1}(s,a)\begin{cases}\underbrace{Q_t(s,a)}_{\text{old value}} + \underbrace{\alpha(s,a)}_{\text{learning rate}} \times \left[\overbrace{\underbrace{\pi_t}_{\text{profit}} + \underbrace{\gamma}_{\text{discount factor}} \underbrace{\max_{a'\in A} Q(s_t,a')}_{\text{max future value}} - Q_t(s,a)}^{\text{expected discounted reward}}\right] & \text{if } a = a_t, \\ Q_t(s,a) & \text{otherwise.}\end{cases}$$

Where $0 < \alpha \leq 1$ is learning rate and $0 \leq \gamma < 1$ is discount factor. As we can see, we update just that Q-value, which represents state-action combination chosen by the agent in current round *t*. All other Q-values remain the same. The new Q-value for state-action combination is made as conjunction of its old value and value *(learning rate)x(expected discounted reward).*

Learning rate determines how much does the agent learn. The higher the learning rate is, the more attention the agent pays to the new information and forgets the past. The lower the learning rate is, the more the past is stressed and the agent tends to learn from new information less rapidly. Therefore, value zero of the learning factor means no learning at all, stickiness to the past, and not accepting anything new. Whereas value one of the learning factor means that the agent takes into consideration only the most recent information into consideration.

The discount factor involved in expected discounted reward determines how important future rewards are to the agent. Therefore we can say that the agent with a value of discount factor zero means that this agent just takes current profit into consideration, whereas the higher discount factor implies that the agent is more future reward driven.

2. Duopoly models

In this chapter, we focus on theoretical foundation of well-known duopoly models competing on quantity – Cournot duopoly and Stackelberg duopoly. We also provide foundation of other theoretical concepts used in this thesis as collusive behaviour and endogenous timing in duopoly games.

2.1 Cournot duopoly

In this thesis we assume a Cournot duopoly model has the following conditions:

- Firms (agents) produce single goods
- Firms are solely choosing the quantity of produced goods
- Both firms have same cost function with constant marginal costs
- Price on market is determined by linear function
  Our inverse demand function is given by

$$p(q_1 + q_2) = \max\left(u - v(q_1 + q_2), 0\right)$$

where $u > 0$ denotes maximum price, $v > 0$ denotes the slope of inverse demand function and q_1,q_2 denote quantities produced by firms.

Because of the condition that both firms have the same cost function with constant marginal costs, the total cost for firm *i* is given by:

$$C(q_i) = wq_i$$

where *w* is the firm's marginal cost of production and q_i is the quantity produced by firm *i*. We are able to compute the profit acquired by a firm *i* on in this model as

$$\pi_i = p(q_1 + q_2)q_i - C(q_i) = q_i\left(u - w - v(q_1 + q_2), -w\right)$$

Having in mind, that the agent tries to maximize its profit and we assume positive profit is available, we know that he will not choose quantity, for which he knows he will obtain negative profit. Negative profit is defined for Q-learning purposes in our simulations, because for bad selection of quantity we have to punish the agent in order to encourage the learning behaviour. We can assume, purely for formal reasons that its profit is defined as

$$\pi_i = p(q_1 + q_2) \cdot q_i - C(q_i)$$

An important feature in a Cournot duopoly model is that the firm thinks about opponent's quantity as a fixed quantity and it tries to maximize its quantity with respect to opponent's quantity. To maximize profit in a Cournot duopoly, we have to equal first derivation of profit with respect to firm's quantity to zero. Therefore profit maximization is given by equation:

$$\frac{\partial \pi_i}{\partial q_i} = p(q_1 + q_2) + \frac{\partial p(q_1 + q_2)}{\partial q_1} - \frac{\partial C(q_i)}{\partial q_i} = 0$$

Having these equations (for q_1, q_2), we can express q_1 as some formula with q_2 and vice versa. We have to realize that it basically defines the best response of the firm if it knows the quantity chosen by its opponent and the response by which the firm has to maximize its profit. We call this an expression reaction function. Reaction function for firm 1 dependent on firm 2's quantity is $R_1(q_2)$ and reaction function for firm 2 dependent on firm 1's quantity is $R_2(q_1)$. We can see both reaction functions on Figure 1 in the space of firms' quantities q_1,q_2.

Cournot equilibrium is a combination of the firms' quantities q_1,q_2, for which each quantity is the best response for the other. Therefore no firm would like to change selection of its quantity, because it would lose profit. According to the theory, if a Cournot duopoly game is played as sequential game, both firms are adjusting their quantities as reaction to the opponent's quantity and selection of quantities finally converge to Cournot equilibrium. For one turn game is Cournot equilibrium also sustainable if we assume that both firms have information opponent's reaction function and think about equilibrium in advance. If the firm is rational, what we presume, is it should predict behaviour of opponent and choose the Cournot equilibrium right at the beginning. On Figure 1 it is the intersection point of both reaction functions, what means that a Cournot equilibrium is when combination of firms' quantities is q_1^*,q_2^*.

What is need to mention for further use is, that the Cournot equilibrium as we defined is also the Nash equilibrium. Nash equilibrium is defined as the set of strategies, when no player can obtain higher profit by unilaterally changing his strategy. We can see both profit functions π_1,π_2 as functions dependent on both quantities $\pi_1(q_1,q_2),\pi_2(q_1,q_2)$. Formally, we can say that for the Nash equilibrium in case of two firms playing Cournot duopoly game is given by:

$$\pi_1\left(q_1^*,q_2^*\right) \geq \pi_2\left(q_1,q_2^*\right) \wedge \pi_2\left(q_1^*,q_2^*\right) \geq \pi_2\left(q_1^*,q_2\right)$$

where q_1^*,q_2^* denotes both the Cournot and Nash equilibriums. That is the reason why Cournot equilibrium is often called in papers Cournot – Nash equilibrium.

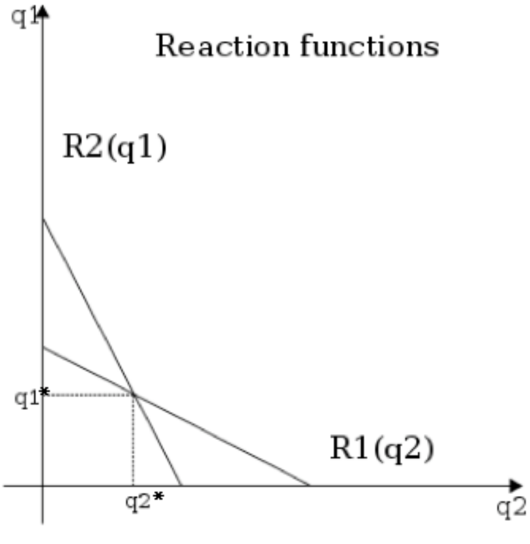

Figure 1 Reaction functions in a Cournot duopoly and quantities in a Cournot equilibrium.

We are interested in joint profit of both firms, as we have firms which outcome is not firmly defined, but is dependent on probability and is therefore changing each time the simulation is launched. Above conditions and assumptions about a simple Cournot model, they imply that the Nash equilibrium is given by

$$q_i^* = \frac{2(u - w)}{3v}$$

where all parameters refer to the parameters used above. Consequently, as we have equilibrium, we can compute the firm's joint profit in the Nash equilibrium, which is equal to

$$\pi^* = \frac{2(u - w)^2}{9v}$$

However, in the Nash equilibrium both firms maximize their own profit: they do not maximizing their joint profit. Firms can maximize their joint profit in the manner that they will together act as one monopolist and in that case they maximize their joint profit. Therefore quantity produced in joint profit will be lower than in the case of Nash equilibrium, because they try to increase the price of a good.

2.2 Stackelberg duopoly

Cournot duopoly is a game when both firms competing on quantity in one market choose quantities simultaneously. Stackelberg duopoly, when we consider just two firms in this competition, is similar to Cournot duopoly. Only exception is that instead of choosing quantities simultaneously firms choose sequentially. As reader surely knows, the firm that has an advantage of choosing quantity as first is called Stackelberg leader. On the other hand, the firm that choose quantity as second is called Stackelberg follower.

We assume conditions same as in Cournot duopoly mentioned above:

- Firms (agents) produce single goods
- Firms are choosing the quantity of produced goods
- Both firms have same cost function with constant marginal costs

- Price on market is determined by linear function

Inverse demand function, total costs and profit remain the same as in Cournot duopoly, so it means, inverse function is given by

$$p(q_L + q_F) = \max(u - v(q_L + q_F),0)$$

where q_L, respectively q_F is quantity of Stackelberg leader, respectively Stackelberg follower. These terms are explained below. Total costs for firm *i* are given by

$$C(q_L) = wq_L, C(q_F) = wq_F$$

Profits for purposes of simulations are given by

$$\pi_L = p(q_L + q_F)q_L - C(q_L) = q_L(u - w - v(q_L + q_F),-w)$$

$$\pi_F = p(q_L + q_F)q_F - C(q_F) = q_F(u - w - v(q_L + q_F),-w)$$

When π_L, respectively π_F is the profit of Stackelberg leader, respectively Stackelberg follower. Profits for purposes of analysis are given by

$$\pi_L = p(q_L + q_F) \cdot q_L - C(q_L)$$

$$\pi_F = p(q_L + q_F) \cdot q_F - C(q_F)$$

Condition, which Stackelberg originally assumed to assure that one of firms, can be Stackelberg, or market leader is an information asymmetry. However, we assume this by a firm's commitment to some quantity. Therefore, Stackelberg followers know, that if he chooses any quantity, The Stackelberg leader is not able to react to his selection and cannot change his quantity.

Knowing this, the Stackelberg leader has a significant advantage, because he knows that Stackelberg follower will react on his quantity with quantity, which is the best response on follower's quantity. Thus, Stackelberg leader can obtain at least same profit as in Cournot equilibrium or better. But still must be fulfil important condition of quantity commitment, because if Stackelberg leader can change his selection of quantity, Stackelberg follower can through choosing quantities not maximizing its profit push Stackelberg leader to Cournot equilibrium.

We can compute Stackelberg equilibrium. Stackelberg follower simply reacts to Stackelberg leader's quantity, so he is using reaction function as we know from Cournot duopoly. This reaction function is $R_F(q_L)$ and is derived from the equation

$$\frac{\partial \pi_F}{\partial q_F} = p(q_F + q_L) + \frac{\partial p(q_F + q_L)}{\partial q_F} - \frac{\partial C(q_F)}{\partial q_F} = 0$$

Stackelberg leader's profit is therefore given by

$$\pi_L = p(q_L + R_F(q_L)) \cdot q_L - C(q_L)$$

So we have to find a maximum of π_L with respect to q_L, it means that we have to derive π_L with respect to q_L and equal it to zero for maximization

$$\frac{\partial \pi_L}{\partial q_L} = \frac{\partial p(q_L + q_F)}{\partial q_F} \cdot \frac{\partial R_F(q_L)}{\partial q_L} + p(q_L + R_F(q_L)) - \frac{\partial C(q_L)}{\partial q_L} = 0$$

If we compute quantity chosen by Stackelberg leader with parameters from our model, we obtain that quantity of Stackelberg leader in equilibrium of Stackelberg duopoly for two firms is given by

$$q_L^* = \frac{u - w}{2v}$$

Quantity chosen by Stackelberg follower in equilibrium of Stackelberg duopoly for two firms is given by

$$q_F^* = \frac{u - w}{4v}$$

Similarly, as in Cournot duopoly, we are interested if collusive behaviour emerges in case of sequential selection of quantities or joint profit of firms will be on the level of Stackelberg duopoly's equilibrium. Therefore, when we put chosen quantities in equilibrium of Stackelberg duopoly into corresponding profit functions, we obtain joint profit of firms in Stackelberg duopoly's equilibrium as

$$\pi^* = \frac{3(u-w)^2}{16v}$$

2.4 Collusive behaviour

Collusive behaviour is a situation in which firm in oligopoly or duopoly make an agreement, either through some communication channel or tacitly about prices, produced quantities or division of markets. We call pure collusive behaviour when firms instead of maximizing their individual profit maximize their joint profit, therefore they act as a single monopolist.

For a Cournot duopoly model, as we have here joint production of both firms in collusive behaviour is equal to $\frac{u-w}{2v}$ and joint profit is equal to $\frac{(u-w)^2}{4v}$. While both firms can increase their own profit by producing more in respect to its goal to maximize their profit, this collusive equilibrium is not stable by the theory, because in collusive behaviour can the agent always improve its profit by producing more and thus push equilibrium into the Cournot equilibrium.

Things can be changed, if we realise that we will try to simulate a Cournot duopoly as a repeated game. If we have agents with the ability to remember outcomes of their previous actions and can be looking at the future expected reward, some interesting results could arise. An important thing to remember is that in case if firms' joint production is higher than the joint production in the Nash equilibrium, it is a sign of a collusive behaviour.

2.5 Endogenous timing in duopoly games

In the real world situations simultaneous actions of both firms are observed uniquely. It is simply the consequence of the fact that it is extremely difficult to match the action movement and make it happens at the same time. However, many economic models, same as for a Cournot duopoly game, assume that production and action taking is made at the same time. Whereas other models, such as a Stackelberg duopoly game (which is quite similar to a Cournot duopoly game), work with the assumption that firms are not moving simultaneously but rather sequentially.

As Hamilton and Slutsky [5] observed, much of the traditional analysis premise that either game is simultaneous (it develops into a Cournot game) or it is sequential (it develop into a Stackelberg game), is supposed to be set from the external environment, so it is exogenously given. They remarked that there is a recent recognition that "whether duopolists play a simultaneous or a sequential move game should not be exogenous but should result from the firm's decision".

Therefore, they suggest distinguishing all duopoly models according to these two conditions:

1. Both firms are moving simultaneously.
2. Firms are moving sequentially, the first mover is called leader and the second mover is called follower.

They were interested in the question of endogenous determining which firm moves first. Therefore they constructed two different extended models, on which they tried to figure out, what position a firm would prefer – being a Stackelberg leader, being a Stackelberg follower or to move simultaneously as in a Cournot game. They developed two games which extended basic models and claimed them to be the most suitable for studying timing of choosing actions:

- Extended game with observable delay – firms announce at which time they will choose an action and they are committed to do so. They do not have to choose an action in the first period. After this first period we have got either Cournot simultaneous game or a Stackelberg leader and follower.

- Extended game of action commitment – if the firm chooses to be a leader, it must announce an action as a leader and then it is committed to fulfil this action. If the firm chooses to be a follower, it can change its action after learning its rival's decision.

In both cases is the basic duopoly game extended by having both firms choose a quantity as in the basic game and then a time when to produce this quantity.

2.5.1 Extended game with observable delay

We keep formal definitions of extended game with observable delay same as it is defined in original paper of Slutsky and Hamilton [5]. Formally, we define the extended game with observable delay as $\Gamma_1 = \left(N, S^1, P^1\right)$. We define *N* to be the set of both firms $N = \{A, B\}$ and α, respectively β to be a compact, convex intervals in R^1, from whose firm A, respectively firm B chooses possible quantities to produce. Profit depends on combination of both firm's quantities and payoff functions are defined to be $a : \alpha \times \beta \to R^1$ for firm A and $b : \alpha \times \beta \to R^1$ for firm B. The set of possible timing, either firm chooses to be first or second which can be defined to be $T = \{F, S\}$. We call $S_A^1 = \{F, S\} \times \Phi_A$ to be the set of strategies of firm A. We use Φ_A , what is the set of functions, whose map the set $\{(F,F), (F,S), (S,F) \times \beta, (S,S)\}$ into interval of all possible quantities α. Similarly we define $S_B^1 = \{F, S\} \times \Phi_B$ to be the set of strategies for firm B and Φ_B to be the set of functions, whose map the set $\{(F,F), (F,S), (S,F) \times \alpha, (S,S)\}$ into interval of all possible quantities of firm B β. Define $s_A = (c, \phi_A) \in S^A$, where *c* is firm A's selection of the timing and $\phi_A \in \Phi_A$. Similarly $s_B = (d, \phi_B) \in S^B$, where *d* is firm B's selection of the timing and

$\phi_B \in \Phi_B$. Having defined state of both firms, we can define payoff functions for current state as

$$p_A(s) = \begin{cases} a\big(\phi_A(c,d),\phi_B(c,d)\big) & \text{if } (c,d) \in \{(F,F),(S,S)\} \\ a\big(\phi_A(F,S),\phi_B(F,S,\phi_A(F,S))\big) & \text{if } (c,d) = (F,S) \\ a\big(\phi_A(S,F,\phi_B(S,F)),\phi_B(S,F)\big) & \text{if } (c,d) = (F,S) \end{cases}$$

and

$$p_B(s) = \begin{cases} a\big(\phi_A(c,d),\phi_B(c,d)\big) & \text{if } (c,d) \in \{(F,F),(S,S)\} \\ a\big(\phi_A(F,S),\phi_B(F,S,\phi_A(F,S))\big) & \text{if } (c,d) = (F,S) \\ a\big(\phi_A(S,F,\phi_B(S,F)),\phi_B(S,F)\big) & \text{if } (c,d) = (F,S) \end{cases}$$

Figure 2 displays how is process of endogenous timing made. We can see, that first two links are just means how to decide whether play simultaneous Cournot game or sequential Stackelberg duopoly. We start to play the basic game on the nodes d$_1$ to d$_4$, when we already know what game do we play.

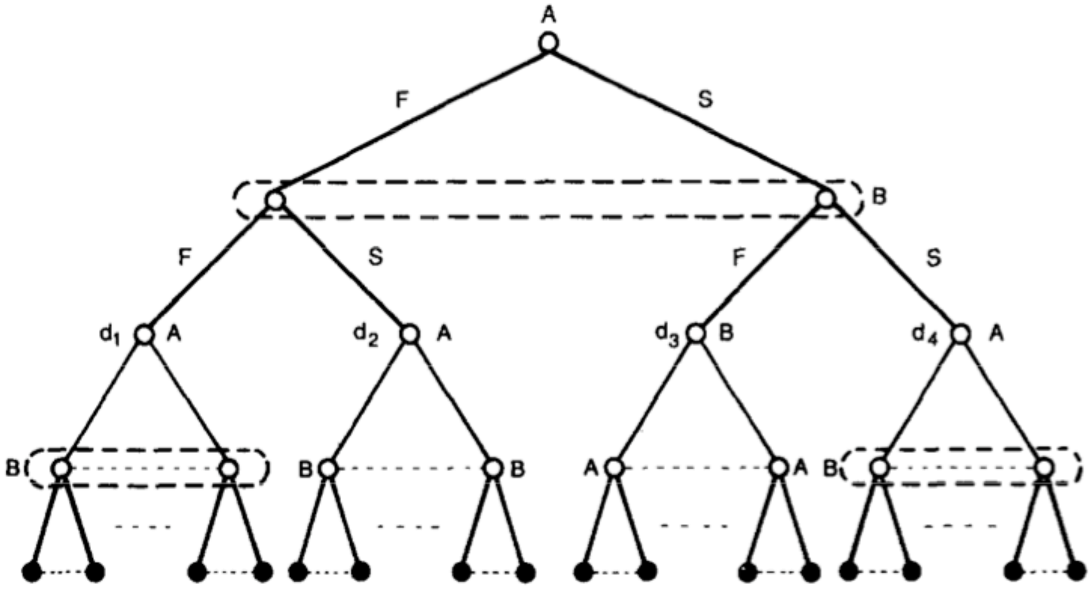

Figure 2 Extended game with observable delay [5].

2.5.2 Extended game of action commitment

Similarly, as in the case above, we do not change the formal definitions of the extended game of action commitment against original paper of Slutsky and Hamilton [5]. We call $\Gamma_2 = \left(N, S^2, P^2\right)$ to be extended game of action commitment. The set of both firms $N = \{A, B\}$ remains the same so as α, respectively β, a compact, convex intervals in R^1, from whose firm A, respectively firm B chooses possible quantities to produce.

What we have in this new game a process of action commitment. Define W to be the action of waiting until the second period to choose an action from α, respectively β. The set $S_A^2 = \{\alpha_i, W \times \psi_A(\beta_i)\}$ is the set of firm A's strategies, where $\alpha_i \in \alpha$ is a selection of quantity from all firm A's possible quantities and ψ_A denotes the set of functions that project $W \cup \beta$ into the interval α. Similarly, firm B's set of strategies is defined as $S_B^2 = \{\beta_i, W \times \psi_B(\alpha_i)\}$, where $\beta_i \in \beta$ is a selection of quantity from all firm B's possible quantities and ψ_B denotes the set of functions that project $W \cup \alpha$ into the interval β.

We use the above definition of payoff functions in the extended game with observable delay. Define strategies chosen by firm A and firm B as $s_A \in S_A^2$ and $s_B \in S_B^2$; denote $s = (s_A, s_B)$. Define a selection of quantities for firm A to be

$$\hat{\alpha}_i = \begin{cases} \alpha_i & \text{if } s_A = \alpha_i \\ \psi_A(\beta_i) & \text{if } s_A = W \times \psi_A(\beta_i), \psi_A \in \Psi_A \end{cases}$$

and similarly define a selection of quantities for firm B to be

$$\hat{\beta}_i = \begin{cases} \beta_i & \text{if } s_B = \beta_i \\ \psi_B(\alpha_i) & \text{if } s_B = W \times \psi_B(\alpha_i), \psi_B \in \Psi_B \end{cases}$$

Having defined a selection of quantities for both firms, in which is included waiting in the case of choosing second period, we can define now payoffs using payoff functions from the case of the extended game with observable delay and define firm A's payoff as $p_A(s) = a\left(\hat{\alpha}_i,\hat{\beta}_i\right)$ and firm B's payoff as $p_B(s) = b\left(\hat{\alpha}_i,\hat{\beta}_i\right)$.

Figure 3 displays how the extended game with action commitment differs from the extended game with observable delay displayed on Figure 2 Instead of two links from nodes, when firms decide whether choose first or second period, we have one additional link of an waiting action in the tree of extended game with action commitment. While white nodes stand for timing and quantity choosing decisions, basic games of Cournot doupoly or Stackelberg duopoly start in the black nodes.

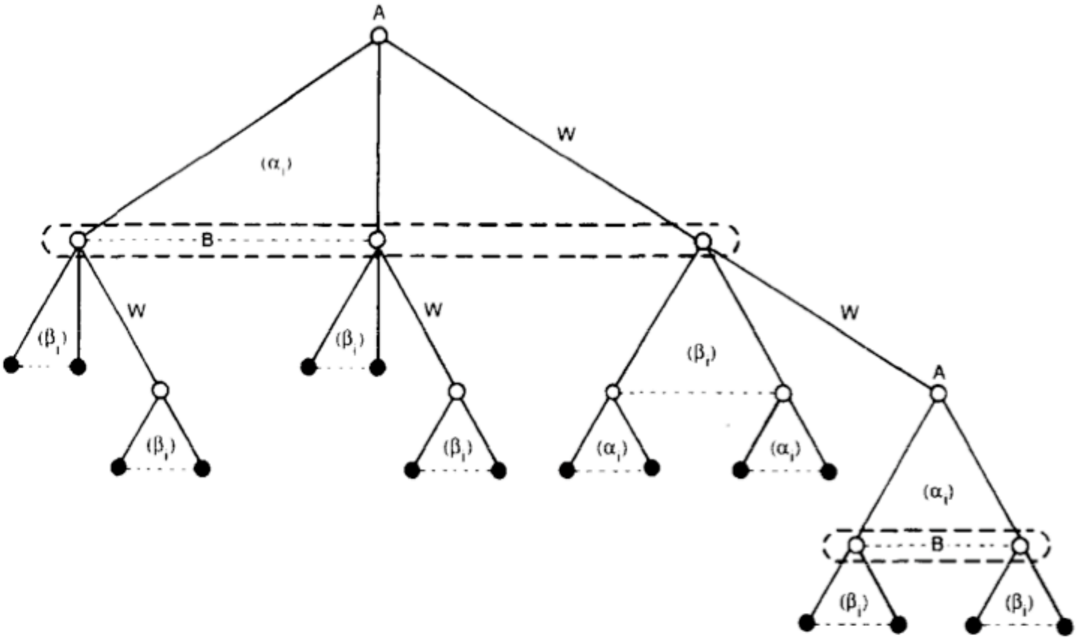

Figure 3 Extended game with action commitment [5].

2.5.3 Conclusion of their research

As a conclusion of their research, they found out that in case of extended game of action commitment "both sequential play subgames are outcomes in undominated strategies; the simultaneous play subgame uses dominated strategies" [5]. So they proved, that if the game is constructed as an extended game of action commitment, there is equilibrium in using simultaneous moves.

When it comes to the extended game with observable delay, their conclusion was that “either the simultaneous play subgame is an outcome of the unique equilibrium in the extended game or one of the sequential play subgames is the unique outcome of the extended game” [5]. While there are no analytical results of how equilibrium is determined in such game, we can use agent-based computational economics to try to simulate this game and look what kind of outcome we will obtain.

By the combination of two different questions which arose above, we can be interested in following topics:

- Will the result of ACE simulation of a Cournot game rather be collusive behaviour or a Nash equilibrium? And if it will be collusive behaviour, what are factors in our model that influence such behaviour and when is joint profit of such behaviour maximized?
- What will be the result of the extended game with observable delay in our case, when the agent can choose between producing in the first or in the second period? And if a result of such a game will be Stackelberg’s sequential game, will there be collusive behaviour or not?

To answer these questions we modelled two types of markets. The first type is the market with one period which successfully simulate a Cournot duopoly game, in which both agents choose quantities they produce simultaneously. Secondly, we provided a model of the market with two periods in one turn, which successfully simulates the extended game with observable delay and can thus lead to either a Cournot duopoly game or a Stackelberg duopoly game.

3. The market with one period

We got the following assumptions about the market and agents acting on this market:

- Each round of the market has only one period
- Number of agents acting on the market is fixed to 2
- Both agents produce a homogenous product
- Both agents have market power, therefore each agent's output decision affects price of good on the market
- Price of good is estimated every round by inverse demand function
- Agents compete in quantities and there is only one market period, these quantities are chosen simultaneously
- There is no exogenous agreement about collusion of agents
- Agents are economically rational and try to maximize their profit

As we have previous assumptions, we achieve a definition of Cournot duopoly, which is defined by exactly these conditions. Assumption of one period in each market round is extremely important, because having more periods there arise lot of other problems. For example, having two periods we can successfully duplicate Stackelberg duopoly, where one of agents can be the quantity leader and another can be the quantity follower.

As we mentioned, the price is estimated every round and is derived from inverse demand function. We use following formula to estimate the price [9]:

$$price = \max\left(u - v\left(q_1 + q_2\right), 0\right)$$

where u > 0 denotes maximum price, v > 0 denotes slope of inverse demand function and q_1,q_2 denote quantities produced by agents.

Then profit for agent i is derived from the formula given by:

$$profit_i = \max\left(q_i \cdot \left(price - \text{cost}_i\right), -q_i \cdot \text{cost}_i\right)$$

where *price* is the price estimated on market for current round and *cost* is the agent's constant marginal cost.

3.1 Setup of computer simulation

While we have provided all formulas, we are using for the estimation of market price and the agent's profit on market, we have not provided constants that we used for market simulation. For estimating the price we used the same constants as were used in the computer simulation paper by Waltmam and Kaymak [9]. Constants for price calculation are following:

- $u = 40$ – denotes maximum possible price
- $v = 1$ – denotes slope of inverse demand function
- $w = 4$ – denotes firm's marginal cost

With these constants we can calculate the joint profit of firms in a Nash equilibrium, which was said to be given by

$$\pi^* = \frac{2(u-w)^2}{9v}$$

thus the joint profit of firms in a Nash equilibrium for our model is 288. Once again it is important to say, that any joint profit of firms above this joint profit is an evidence of collusive behaviour.

We can also calculate the joint profit of firms in a Stackelberg equilibrium, which was said to be given by

$$\pi^* = \frac{3(u-w)^2}{16v}$$

thus the join profit of firms in a Stackelberg duopoly equilibrium for our model is 243. Please note that in case of choosing not simultaneous, but sequential timing, every joint profit higher than 243 means collusive behaviour in a Stackelberg duopoly.

Similarly, as we computed joint profits for agents in a Nash equilibrium and a Stackelberg duopoly equilibrium for parameters which characterize setup of market, we should compute joint profit arisen in the case of pure collusive behaviour. This happens when both agents are trying to instead of maximization of their own profit maximize joint profit they can obtain. Thus they behave as a single entity – monopolist on a market. As we mentioned before, their joint profit for collusive behaviour is given by

$$\frac{(u-w)^2}{4v}$$

what gives us joint profit of 324 for setup of the market we used for simulations.

As agents try to fulfil their own goal which is to maximize their individual profit, some kind of collusion would be interesting result. We can talk about inclination to collusive behaviour when joint profit of agents is somewhere between 288 and 324.

3.2 Application of Q-learning

Q-learning is applied to agents as follows. We are playing repeated game on a market. During each market round the agent must choose the quantity he is going to produce. At the beginning of each round the agent is in some state s_t. The agent appeared in this state by taking into account his and his opponent's actions in a previous round $t-1$. Being in state s_t the agent has to choose some action a_t from the set of all available actions. This choice depends on his state and is made probabilistically based on his Q-values for the state he is in. Choosing action a_t, the agent obtains some profit π_t, which is dependent on the market price and he moves from a state s_t to a new state s_{t+1}. Then the agent updates appropriate Q-values using information about profit π_t, chosen action a_t and last state s_t. Which means he actually modifies the way how he chooses actions in next rounds.

Formally, denote value Q(s, a), where $s \in S$ is the state of the agent in round *t* and $a \in A$ is the action chosen in round *t-1*, as an agent's Q-value for that combination of state and action (s,a). Lets clarify, that S is a finite set of all possible states of the agent and A is a finite set of all possible actions of the agent. The Q-value for certain combination of state and action can help us decide which action we choose next time. The higher the Q-value is, the higher the probability of choosing that action. Formally, there is a probability assigned to each Q-value and the probability that the agent chooses action *a* in the current state is given by the following formula:

$$\Pr(a) = \frac{\exp\left(\left(Q_t(s_t,a) - \max_{a' \in A} Q_t(s_t,a')\right)/\beta\right)}{\sum_{a' \in A} \exp\left(\left(Q_t(s_t,a') - \max_{a'' \in A} Q_t(s_t,a'')\right)/\beta\right)}$$

In this formula s_t denotes the agent's state at the beginning of round *t*. Parameter $\beta > 0$ denotes experimentation tendency. The higher parameter β is, the higher is probability that the agent chooses to experiment, what means not to choose an action that has the highest Q-value. Whereas, closer parameter β is to zero, the more probable it is that the agent chooses an action with the highest Q-value.

Action choosing, as we introduced here, is known in literature as the Boltzmann exploration strategy. We use it to model agent's action choice behaviour, because it corresponds to the logit model, which is widely used in economic applications.

As we know, a high value of parameter β means experimentation and a small value of parameter β means choosing action with the highest Q-value. Therefore, we need β to be large at the beginning and converges to zero at the end. We used the function for β estimation proposed in Waltman and Kaymak [9] and it is given by the formula:

$$\beta = 1000 \cdot 0{,}99999^t$$

where *t* denotes the current round number. It is obvious, that β converges to zero for *t* converging to the positive infinity.

The question is: How exactly are the agent's Q-values updated? There is a different method to do it for the agent with a memory than for the agent a without memory.

3.3 The agent with a memory

All we have mentioned above is applicable to the agent with a memory. Although we have to clarify what exactly does state of agent in this particular case means exactly. We call state *s* to be a pair *(my quantity, opponent's quantity)*, where *my quantity* is a quantity chosen in the last round by an agent, who this state belongs to and *opponent's quantity* is a quantity chosen in the last round by his opponent. Therefore, we formally use state *s*, but be aware, that it is a symbol for a pair *(my quantity, opponent's quantity).*

What we have to specify now is the way how to update Q-values. In our simulations Q-values are updated according to

$$Q_{t+1}(s,a) = \begin{cases} (1-\alpha)Q_t(s,a) + \alpha\left(\pi_t + \gamma \max\limits_{a' \in A} Q_t(s,a')\right) & \text{if } s = s_t \text{ and } a = a_t, \\ Q_t(s,a) & \text{otherwise.} \end{cases}$$

where $0 < \alpha \leq 1$ is learning rate, $0 \leq \gamma < 1$ is the discount factor and π_t stands for the profit obtained in round *t* on a market. The learning rate in this formula establishes how quick the agent learns or better said, how much does the agent cares for new obtained profit compared to recent history, which is codified in Q-value. While the learning rate concerns about history, discount factor focuses on the future on the other hand, as it determines the importance of the future choice. If we have discount a factor of the zero value, the agent becomes "opportunistic" and is just focused on the current profit. Whereas, having the discount factor close to the value one the agent is focused on obtaining long-term high profit.

As mentioned in Waltman and Kaymak [9] above update rule *"has the appealing property that when there is only one learning agent (either because there is only one agent or because all other agents use fixed* *strategies), the update rule allows the agent, under certain conditions, to learn to behave optimally".*

3.3.1 Result of the computer simulations

As we tried to achieve accurate and plausible results, every simulation of the market and therefore convergence to some joint profit was the result of one million turns in one market simulation. The same number was used by in Waltman and Kaymak [9] their simulation and in respect of specified values, it is a sufficient number as far as last 100 000 turns the joint profit of firms changes very gently if even at all. Finally, we made a final joint profit as an average of the last 10 000 turns of joint profit for each simulation, what we consider to be sufficient number.

As we have two different parameters to examine, we had to make a market simulation for each combination of pair *(learning rate, discount factor)*. We wanted to provide credible and significant results, so for every pair combination we ran 100 market simulations.

Having pair *(learning rate, discount factor)* and increment for each value is 0,1, so as we started at value 0,05 and finished at value 0,95, it is 10 values for each parameter, what give us together 100 combinations of pair *(learning rate, discount factor)*. It would not be good for anything to put here all 100 values in a table, as we would have too big table and the amount of data would be overwhelming. So we decided to display all these values in a graphical way. However, we provided all generated data for every combination of pair *(learning rate, discount factor)* and you can find this values on attached CD.

We can notice on Chart 1, we can say that in general increasing of the learning rate and decreasing of the discount factor leads to increasing of joint profit, so it means increases collusive behaviour of both agents. As we will see for the agent without a memory, increasing of the learning rate will also emerge in amplification of collusive behaviour. Moreover, we are dealing with the discount factor.

Generally, we can say that for all levels of the learning rate we can see significant improvement of joint profit at the moment of movement from the point when joint profit of agents shows the worst performance. From the generated data we can see, that for the pair *(learning rate, discount factor)* from set {[0,05;0,95], [0,05;0,85], [0,15;0,95], [0,15;0,85], [0,25;0,95]} is joint profit either equal or slightly less than a joint profit in Nash equilibrium. As we will see below, for the agent without memory, when we are changing just learning rate, there will not arise situation when average agents' joint profit will be less than joint profit of agents in Nash equilibrium.

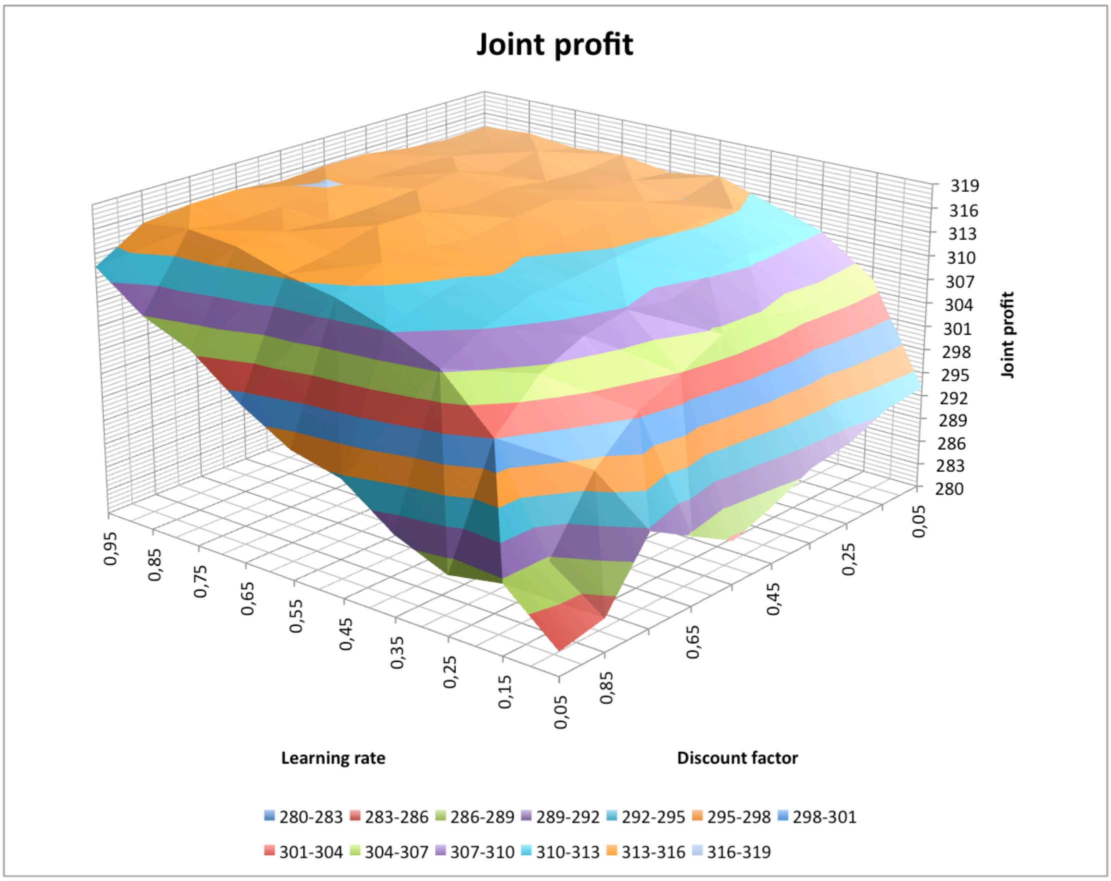

Chart 1 Joint profit of agents with a memory in the one turn market.

The discount factor says determines how important is maximal future value important for agent, so as how he is focused on a long-term profit in comparison with short term profit. The agent who prefers rather long-term profit is characterized by value of the discount factor close to 1 and the agent who prefers rather short-term is characterized by value of discount factor close to 0. According to Chart 1 we might do the conclusion, that combination of the low learning rate and high preference on long-term profit does not lead to collusive behaviour in the interaction of two agents. Even more, we can say, that it did not result in an average joint profit at the level of joint profit of agents in Nash equilibrium. If we consider the value from opposite side of a scale, value of discount factor 0, we will obtain quite similar results as for the agent without a memory.

In the average we obtain collusive behaviour, if we have joint profit between joint profit of agents in Nash equilibrium and joint profit of agents in pure collusive behaviour, what is in our case 288, respectively 324. Even thought we found in every 100 market rounds at least one occurrence of pure collusive behaviour, so that joint profit of agents was 324, we did not approach to this joint profit in average for any combination of parameters. We can easily see from the Chart 1 that significant majority of final joint profits are higher than 288 and are somewhere between 304 and 316. Therefore, we can consider it to be as collusive behaviour and since average of 288 and 324 is 306, we can say that for chosen parameters and model agents more tend to collusive behaviour.

3.4 The agent without a memory

We denote the agent to be an agent without memory in the case that his decision process about which action should be chosen is based only on the knowledge of the Q-values for each action, but does not depend on the state he is currently in. So instead of having the Q-value $Q_t(s, a)$ dependent on both state *s* and action *a*, the Q-value for the agent without memory depends only on action, and is therefore reduced to $Q_t(a)$. In a previous case, *s* was a symbol for pair *(my quantity, opponent's quantity)*; in this case *s* is symbol for *(null)*, as there is not any memory.

The reason for having the agent without a memory is that most of the reinforcement learning models studied in economic literature do not have a possibility of remembering things. So we examine both; the agent with and the agent without a memory. Having the agent without a memory, the formula of choosing certain Q-value is changed to:

$$\Pr(a) = \frac{\exp\left(\left(Q_t(a) - \max_{a''\in A} Q_t(a'')\right)/\beta\right)}{\sum_{a'\in A} \exp\left(\left(Q_t(a') - \max_{a''\in A} Q_t(a'')\right)/\beta\right)}$$

We have a quite good idea how the market where are simulations are being executed, looks like. We need to take a closer look at the updating process of the Q-values. The formula for updating the Q-values is given by:

$$Q_{t+1}(a) = \begin{cases} (1-\alpha)Q_t(a) + \alpha\pi_t & \text{if } a = a_t, \\ \mathrm{Q}_t(a) & \text{otherwise}. \end{cases}$$

We can see, that compared to the update rule applied to the agent with a memory, we are missing the discount factor. According to the Waltman and Kaymak [9] this is because *"an agent without a memory cannot take into account the consequences of the action plays in the current stage game on the payoffs it obtains in future state games. For such an agent, the discount factor must therefore equal zero."*

3.4.1 Result of the computer simulations

Similarly, as with the simulation of the agent with a memory, we made simulation of 1 000 000 turns for every simulation of the market. We did 100 market simulations for every fixed value of the learning rate. As argued above, this number is sufficient, because joint profit of last 100 000 turns is fluctuating just mildly, therefore we are able to obtain quite stable average. Then result joint profit for one market simulation was created as an average of last 10 000 joint profits of agents.

While we have only one parameter, which can be changed in case of the market with one period and the agent with a memory, our analysis will be concerned about changing this parameter and will analyse results that we obtained by simulations. As we have 1 000 000 turns in one market simulation, we decided to run 100 of this market simulations to ensure higher credibility of results and avoid some random effects, as we are dealing with probabilistic model of decision-making of agents in ACE and results of simulations are dependent on random numbers that are used. So for every setup of parameter we provide 100 market simulations.

We can see in Table 1 results of this simulations, where by the result we mean average of 100 joint profits, whose are results of a market simulation for one fixed value. The corresponding standard deviation for each set of 100 market simulations is provided in parentheses. There is also provided a joint profit of firms whose are in a Nash equilibrium, which we have calculated to be 288 for our setup of market model.

Learning rate	Nash	α = 0,05	α = 0,15	α = 0,25	α = 0,35	α = 0,45
Joint Profit	288	300,6 (10,6)	306,6 (12,8)	310,3 (12,3)	312,3 (9,4)	313,7 (7,4)
Max Profit		324	324	324	324	324
Learning rate	α = 0,55	α = 0,65	α = 0,75	α = 0,85	α = 0,95	
Joint Profit	312,9 (8,1)	314,5 (6,9)	314,6 (6,8)	315,5 (6,0)	314,6 (6,3)	
	324	324	324	324	324	

Table 1 Joint profit of agents without a memory in one turn market according to the learning rate.

Looking to the Table 1 we can say, that even for the lowest value of the learning rate there is collusive behaviour emerging. This is really interesting result, because it can give us insight into the problem, why is collusive behaviour emerging also on the market with firm, whose do not communicate with each other. As agent has only goal to maximize its profit, as it is a common situation observed in a real world, because all entities on the market are driven by self-interest as had already noted Smith in 1776 in his *The Wealth of Nations*.

We also know, that only information the agent has from external environment is profit he obtains every round and his decision-making about choosing proper action in a future is changed and evaluated just according the information about the profit.

We can better see the mean how is joint profit dependent on the learning rate, if we put results we obtained in the chart as we did in Chart 2.

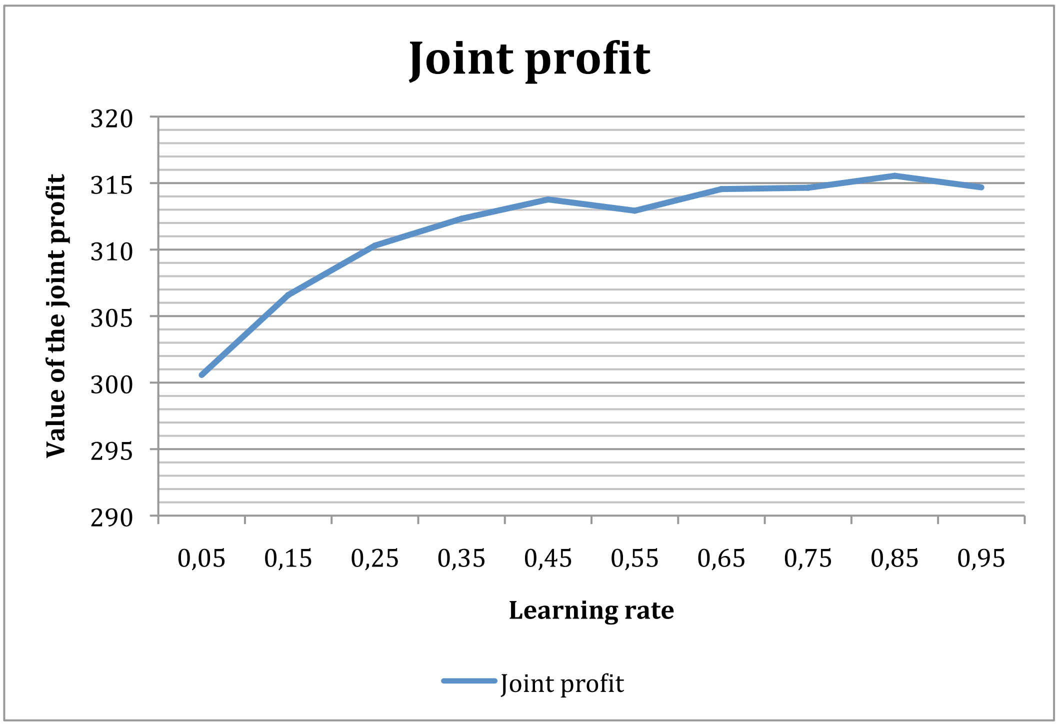

Chart 2 Joint profit of agents without a memory in the one turn market.

We can see, that we have evidence of collusive behaviour throughout the whole interval of possible values that the learning rate can have. However, we can see, that there with an increasing learning rate is also increasing joint profit, what means increasing collusive behaviour of agents. Even for the lowest learning rate we ran our simulations for, is collusive behaviour quite obvious since joint profit of agents in a Nash equilibrium is 288 and average joint profit for the learning rate α = 0,05 is 300,6.

There is a significant increase in joint profit of agents when increasing the learning rate. We can observe difference of value 6 between the agent with the learning rate α = 0,05 and α = 0,15. The difference of value 4 is between the agent with the learning rate α = 0,15 and α = 0,25. There are some remarkable differences when we move between learning rates from α = 0,25 to α = 0,35 and α = 0,35 to α = 0,45. After this is the joint profit more less stable with some error variance, which is evocated by use of random numbers and probabilistic model of decision making.

However we obtained the maximal value for collusive behaviour of 324 in every set of 100 simulations for fixed learning rate, we did not experience the average joint profit to be as high as in a pure collusion. Therefore we can say, that there is not a pure collusion of agents in average.

The learning rate in the Q-learning reinforcement model expresses how much is new information obtained from the market relevant in comparison with current Q-value, which as we know contains the performance of state-action combination in past turns. Therefore with increasing learning rate we put more stress on current profit.

In our model is joint profit of agents approximately same from the learning rate of value 0,45 and basically not changed to the value 0,95. Thus our conclusion in a case of interaction of two agents without a memory in the market with one round is that collusive behaviour emerges as soon as agents have possibility to learn from the external environment. Although joint profit in this collusive behaviour is higher than the joint profit in Nash equilibrium, it is not that high as it is in a pure collusion. Full ability to maximize joint profit and therefore maximal value of collusion is reached for the learning rate higher than 0,45. We can conclude that if the agent puts weigh current profit more than on 45%, there is an evidence of maximal collusive behaviour in our examined case.

4 The market with two periods

Having a market with two periods we need to change our model a little bit. Firstly, we have to consider that instead of just one possible way of choosing quantities, we have three different situations to choose from:

- Both agents choose to produce in a first period.
- The agents choose to produce in different periods.
- Both agents choose to produce in a second period.

The expansion of the previous model brings us more information that agents can obtain. Generally, they obtain information about the opponent's choice of pair (period, quantity). Therefore, agents can obtain the following information about opponent's choice:

- (null, null) – in the case, when the agent chooses to produce in the first period, he does not know the opponent's choice of either period or quantity.
- (period, quantity) – in the case, when the agent chooses to produce in the second period and the opponent chooses to produce in the first period. As we do not demand commitment to produce quantity in advance, the agent, who chooses the second period, is able to observe the quantity produced by his opponent.
- (period, null) – in the case, when both agents choose to produce in the second period. Although both agents are able to observe that the opponent did not produce in the first period, they are not able to observe the opponent's choice, as they have to choose and produce the quantity simultaneously.

There are more differences of this model in following models of agents either with or without memory.

4.1 The agent without a memory

There is no different logic in evaluating probabilities, however we make them clear to create order. The probability for choosing a period is given by the following formula:

$$\Pr(p) = \frac{\exp\left(\left(Q_t(p) - \max_{p''\in P} Q_t(p'')\right)/\beta\right)}{\sum_{p'\in P} \exp\left(\left(Q_t(p') - \max_{p''\in P} Q_t(p'')\right)/\beta\right)}$$

where *p* is a period, $Q_t(p)$ is the Q-value for period *p* in turn *t,* and β is an experimentation tendency as we know it. Symbol *P* stands for finite set of all possible periods, which is two in our case.

Similarly, as we had state of the agent being *(null)* for the agent without a memory in case of a market with one period, now we have the state of the agent being *(null)* in a case of a market with two periods. However, we should realize and keep in mind, that there are two different sets of Q-values.

One set of Q-values is needed in case if the agent chooses in the first period and the other one set is needed in case if the agent chooses in the second period. This distinction is particularly important, as choosing in the second period means to obtain some additional information about the opponent's choice, therefore we have a pair *(period, quantity)* or *(period,null)*. While choosing in the first period do not bring us any additional information.

Formally, we say that the Q-values for quantities are defined as:

$$Q_t(o,a) = \begin{cases} Q_{t,1}(a) & \text{for first period,} \\ Q_{t,2}(p_o,q_o,a) & \text{for second period.} \end{cases}$$

where *o* is the opponent's choice for the current period and it is a pair *(period, quantity).* For choosing the first period we will obtain a pair *(null, null)* for the opponent's choice. Therefore the Q-value in round *t* $Q_t(o,a) = Q_t((null,null),a)$ is reduced to $Q_{t,1}(a), a \in A$, where *A* is a finite set of all possible actions of the agent. As far as the second period is concerned, the Q-value in round *t* is

$Q_t(o,a) = Q_{t,2}(p_o,q_o,a)$, $p_o \in P$. The symbol *P* stands for a finite set of all possible periods, $q_o \in Q \cup \{null\}$ where *Q* is a finite set of all possible quantities which the opponent can choose and also *null* value in case the opponent's period is unknown.

According to this definition, the probabilities are given by:

$$\Pr(Q_t(o,a)) = \frac{\exp\left(\left(Q_t(o,a) - \max_{a''\in A} Q_t(o,a'')\right)/\beta\right)}{\sum_{a'\in A} \exp\left(\left(Q_t(o,a') - \max_{a''\in A} Q_t(o,a'')\right)/\beta\right)}$$

The update rule for Q-values for quantities is given by:

$$Q_{t+1}(o,a) = \begin{cases} (1-\alpha)Q_t(o,a) + \alpha\pi_t & \text{if } a = a_t, \\ Q_t(o,a) & \text{otherwise.} \end{cases}$$

However, we got a different update rule for Q-values in round *t* for periods:

$$Q_{t+1}(o,a) = \begin{cases} (1-\alpha)Q_t(o,a) + \alpha\gamma \max_{a'\in A}\left(Q_t(o,a')\right) & \text{if } a = a_t, \\ Q_t(o,a) & \text{otherwise.} \end{cases}$$

where $0 \le \alpha < 1$, $0 < \gamma \le 1$ is the learning rate and the discount factor respectively. Both these variables have been explained above.

4.1.1 Result of the computer simulations

As we tried to achieve accurate and plausible results, every simulation of the market and therefore convergence to some joint profit was the result of one million turns in one market simulation. It is a sufficient number as far as last 100 000 turns the joint profit of firms changes does not change too much. Finally, we made a final joint profit as an average of the last 10 000 turns of joint profit for each simulation, what we consider to be sufficient number.

As we have two different parameters to examine, we had to make a market simulation for each combination of pair *(learning rate, discount* *factor)*. We wanted to provide credible and significant results, so for every pair combination we ran 100 market simulations.

Having pair *(learning rate, discount factor)* and increment for each value is 0,1, so as we started at value 0,05 and finished at value 0,95, it is 10 values for each parameter, what give us together 100 combinations of pair *(learning rate, discount factor)*. As above, we decided to display all these values in a graphical way. However, we provided all generated data for every combination of pair *(learning rate, discount factor)* and you can find this values on attached CD.

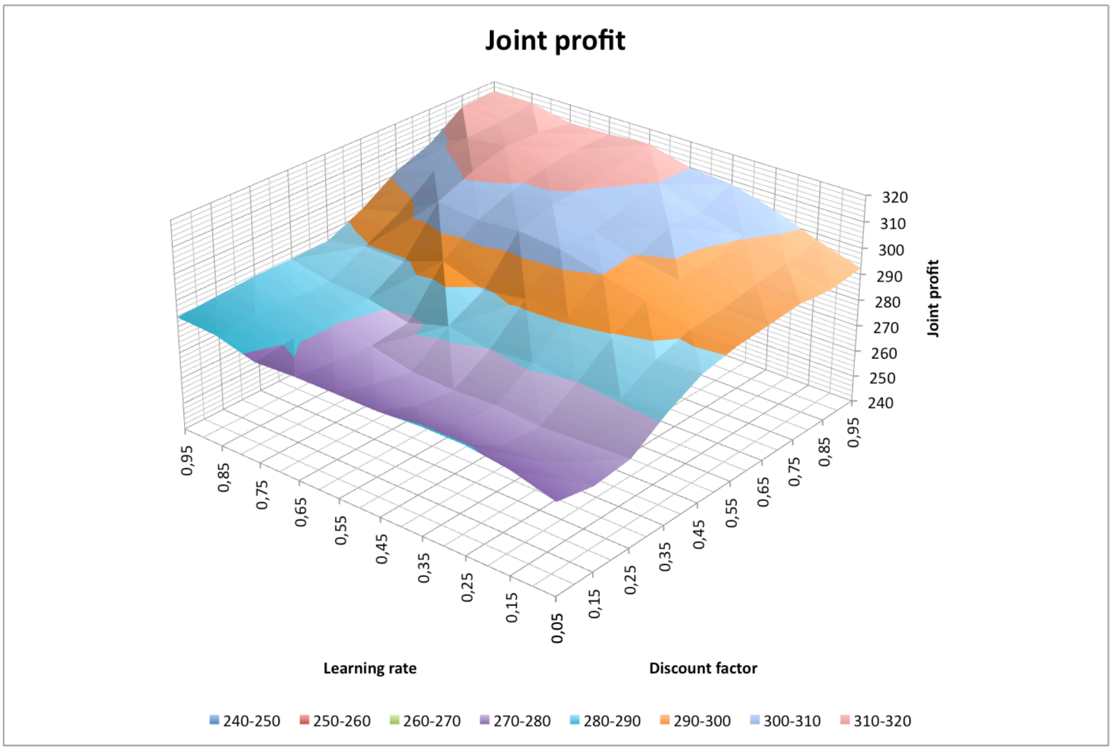

Chart 3 Joint profit of agents without a memory in the market with two turns.

It is obvious, that in case of agents without a memory, which operate on the market with two turns is seen significant joint profit increase with increase of both parameters. The higher learning rate we have, the higher average joint profit we obtain and this holds for the discount factor as well.

What is more interesting in this case than joint profit is agents' selection of its first period. Therefore in case of the market with two periods, we must provide results of agents' selection of period in which they want to produce. We decided to provide the map of agents' selection as the reader can see on Chart 4 and Chart 5.

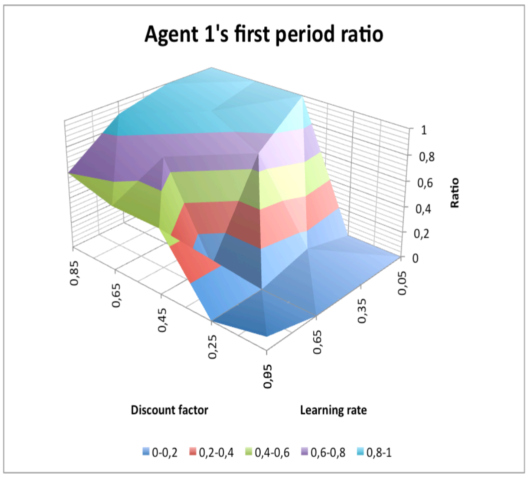

Chart 4 Agent 1's ratio of selection the first period, with inverse order of the learning rate.

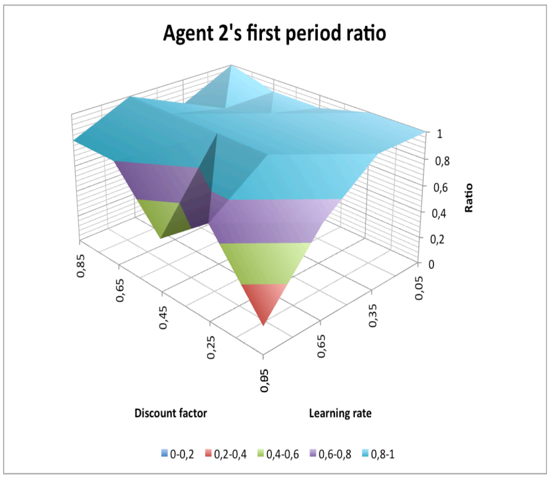

Chart 5 Agent 2's ratio of selection the first period, with inverse order of the learning rate.

We used inverse order of the learning rate for better charts' readability. We can see on both charts tendency to choose period one with increase of the discount factor and with decrease of the learning rate.

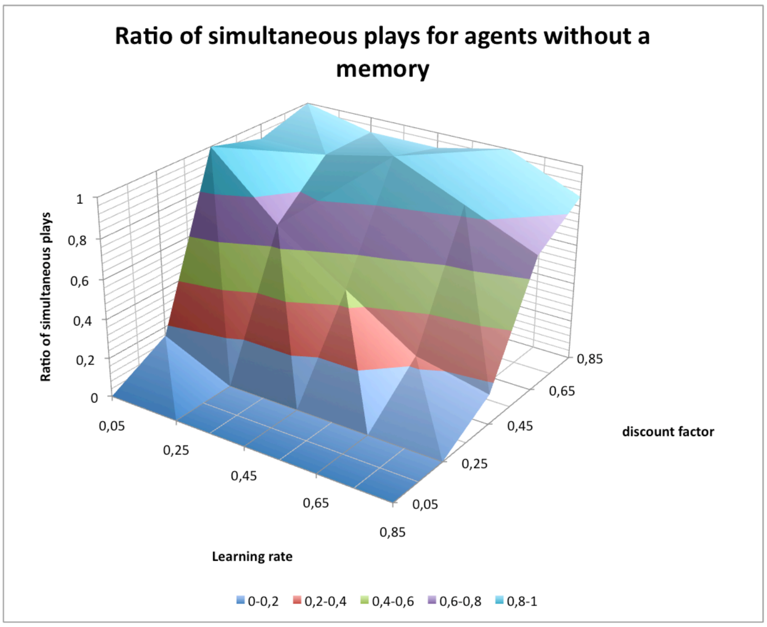

Chart 6 Ratio of simultaneous plays for agents without a memory dependent on the learning rate and the discount factor

Another interesting feature we can look at is how many times agents chose to produce simultaneously and how many times they decided to produce sequentially. For this purpose we have Chart 6.

If we count a ratio of values that are smaller than 0,5 we obtain ratio of choosing sequential play rather than simultaneous. The value higher than 0,5 means that more than 50 percent of simulations for a fixed pair *(learning rate, discount factor)* finished in convergence to a simultaneous play. Whereas, the value smaller than 0,5 means that more than 50% of simulations for a fixed pair *(learning rate, discount factor)* finished in convergence to a sequential play, what is exactly what we are looking for. For this case, the ratio of values smaller then 0,5 is 52%.

As a conclusion we can confirm collusive behaviour of agents without a memory in the market with two rounds. But result the 52% of all games converge into simultaneous plays implies that 48% of all games converge into sequential plays. As a result we have that 52% of games finish as a Cournot duopoly and of games 48% finish as a Stackelberg duopoly, what means that neither of this two duopolies is preferred.

4.2 The agent with a memory

The agent with a memory differs from the agent without a memory in memorizing the state of the last round. It means that state *s* of the agent is four *(my period, my quantity, opponent's period, opponent's quantity)* where all information refer to the last round. We know a pair *(my period, my quantity),* and we have three different general possibilities for a pair *(opponent's period, opponent's quantity)* as well. Denoting this four as *s*, then the Q-value is defined as $Q_t(s,o,a)$, where *o* is the opponent's choice of the pair *(opponent's period, opponent's quantity)* for the current round.

Similarly as in the previous case, we divide the definition of Q-values consequently:

$$Q_t(s,o,a) = \begin{cases} Q_{t,1}(s,a) & \text{for first period,} \\ Q_{t,2}(s,p_o,q_o,a) & \text{for second period.} \end{cases}$$

where *o* is the opponent's choice for the current period and it is pair *(period, quantity)*. For choosing the first period, we will obtain pair *(null, null)* for the opponent's choice. Therefore Q-value in round *t* is $Q_t(s,o,a) = Q_t(s,(null,null),a)$, which can be reduced to $Q_{t,1}(s,a), a \in A$, where *A* is a finite set of all possible actions of the agent. As far as the second period is concerned, the Q-value in round *t* is $Q_t(s,o,a) = Q_{t,2}(s,p_o,q_o,a),\ p_o \in P$. The symbol *P* stands for a finite set of all possible period, $q_o \in Q \cup \{null\}$ where *Q* is finite set of all possible quantities which the opponent can choose and also a *null* value in case we do not know the opponent's period.

Similarly, instead of $Q_t(p)$ for the Q-value for period *p* in round *t*, we have $Q_t(s,p)$. Therefore, the probabilities for the Q-values are:

$$\Pr(p) = \frac{\exp\left(\left(Q_t(s,p) - \max_{p''\in P} Q_t(s,p'')\right)/\beta\right)}{\sum_{p'\in P} \exp\left(\left(Q_t(s,p') - \max_{p''\in P} Q_t(s,p'')\right)/\beta\right)}$$

for period and

$$\Pr(Q_t(s,o,a)) = \frac{\exp\left(\left(Q_t(s,o,a) - \max_{a''\in A} Q_t(s,o,a'')\right)/\beta\right)}{\sum_{a'\in A} \exp\left(\left(Q_t(s,o,a') - \max_{a''\in A} Q_t(s,o,a'')\right)/\beta\right)}$$

for quantities.

Having covered this, we can proceed to the explanation of the updating process. As before, the updating process for the Q-values is different for quantities and periods. Therefore, the updating rule for the Q-values for quantities is:

$$Q_{t+1}(s,o,a) = \begin{cases} (1-\alpha)Q_t(s,o,a) + \alpha\left(\pi_t + \gamma_1 \max_{a'\in A} Q_t(s,o,a')\right) & \text{if } a = a_t, \\ Q_t(s,o,a) & \text{otherwise.} \end{cases}$$

and the updating rule for Q-values for periods is given by:

$$Q_{t+1}(s,o,a) = \begin{cases} (1-\alpha)Q_t(s,o,a) + \alpha\gamma_2 \max\limits_{a' \in A} Q_t(s,o,a') & \text{if } a = a_t, \\ Q_t(s,o,a) & \text{otherwise}. \end{cases}$$

Please note an important distinction: we have a different discount rate for updating quantities and for updating periods. The discount rate for updating quantities determines how much the agent takes into consideration the result of current game to the next one. Discount rate of value zero means, that the agent does not take into consideration the result of current game to the next one, whereas the discount rate close to value one means that the agent does. We should note that in our simulations we used same setting for the agent, which takes the result of current game to the next one into consideration, as in Waltman and Kaymak [9] for a case of simulating not considering agents. It means that for the not considering agent, which takes the result of current game to the next one into consideration, are the discount factors equal. Denote the agent, which takes the result of current game to the next one into consideration, as the considering agent and the agent, which does not take the result of current game to the next one into consideration, as not considering agent.

4.2.1 Result of the computer simulations

We tried to achieve accurate results, so every simulation of the market and therefore convergence to some joint profit was the result of one million turns in one market simulation. It is a sufficient number as far as last 100 000 turns the joint profit of firms changes does not change too much. Finally, we made a final joint profit as an average of the last 10 000 turns of joint profit for each simulation, what we consider to be sufficient number.

We have two different parameters to examine, thus we had to make a market simulation for each combination of pair *(learning rate, discount factor)*. We wanted to provide credible and significant results, so for every pair combination we ran 100 market simulations.

Having pair *(learning rate, discount factor)* and increment for each value is 0,1, so as we started at value 0,05 and finished at value 0,95, it is 10 values for each parameter, what give us together 100 combinations of pair *(learning rate, discount factor)*. As above, we decided to display all these values in a graphical way. However, we provided all generated data for every combination of pair *(learning rate, discount factor)* and you can find this values on attached CD.

Considering and not considering agent

Firstly, lets have a look at the chart of joint profit for not considering and considering agents. This join profit is displayed on Chart 7. As we can see, we obtained results that are more interesting. To explain it we need to know how is decision, either to produce simultaneously or sequentially, dependent on the pair *(learning rate, discount factor)*.

We should say that agents do not decide whether to play simultaneously or sequentially, but they make decision in which period they want to produce. If their decision is the same period, simultaneous play occurs, but they cannot choose whether they want to play simultaneously or sequentially by any mean.

However, we can analyze in what ratio from all market simulations from fixed pair *(learning rate, discount factor)* their decisions led to a simultaneous play. According to the results we obtained the considering agent always converges to the decision to choose the first period. Thus, we do not attach chart or any other table decision making of the considering agent. Chart 8 displays ratio of simultaneous plays for the not considering agent and as we can see, it is much more interesting.

It is useful to realize that if the considering agent converges to choose first period for every combination of pair *(learning rate, discount factor)*, the ratio of simultaneous plays also means the ratio of choosing the first period by the not considering agent.

As we can observe, low learning rate of 0,05 with combination of the agent's goal to maximize his profit and increasing discount factor lead to a choice to choose the first period in increasing rate with increasing discount factor. Therefore, we obtain simultaneous play in higher rate that in an average.

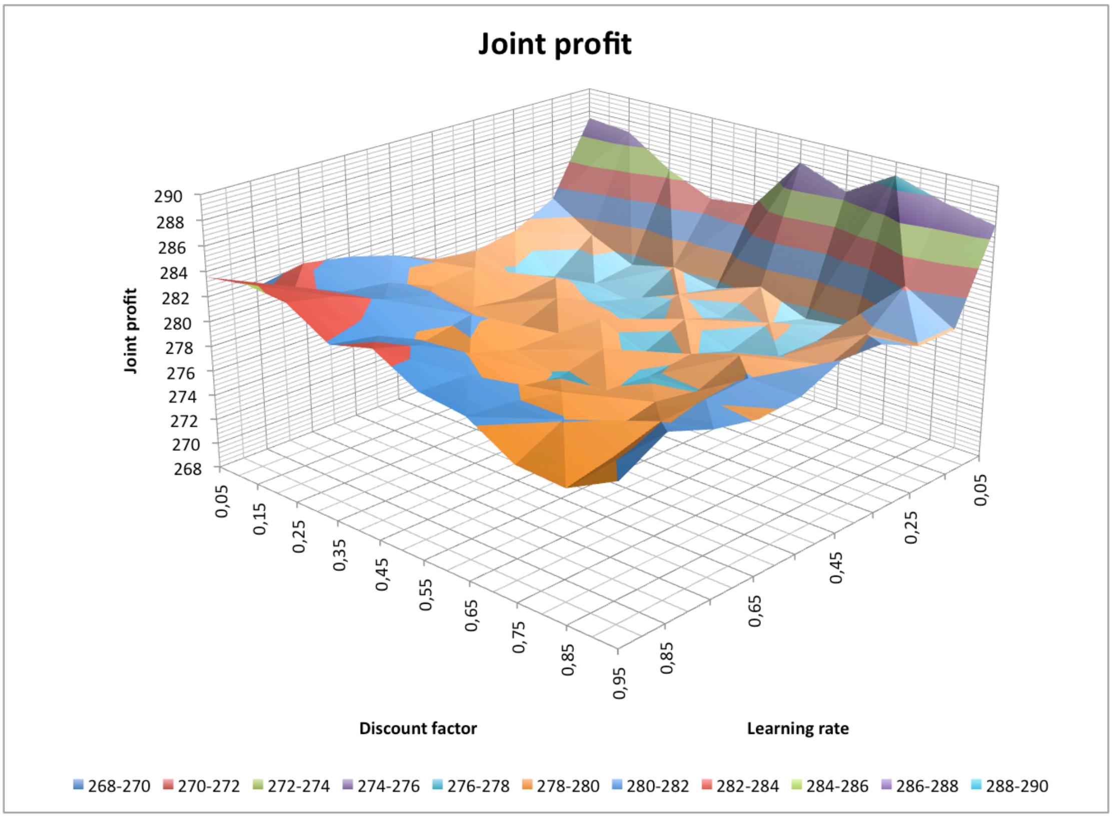

Chart 7 Joint profit of considering and not considering agent and their dependency on the learning rate and the discount factor.

Even though we have the ratio of value 1 for simultaneous plays with fixed combination *(learning rate, discount factor)* = (0.05, 0.95), result of joint profit displayed on Chart 7 for this pair is around the joint profit for a Nash equilibrium, which is 288 in our model.

Another significant point of a simultaneous play is around pair *(learning rate, discount factor)* = (0.75, 0.45) does not achieve any significant level of joint profit. Whereas pair *(learning rate, discount factor)* = (0.95, 0.05) is a point with value 0, what means that for this pair Stackelberg duopoly is always chosen. As we can see on Chart 8, there are significant tendency to be a Stackelberg follower for the not considering agent. Surprisingly, we can observe on Chart 7, that the joint profit around point *(learning rate, discount factor)* = (0.95, 0.05) is increasing and we know that the joint profit in a Stackelberg duopoly for our model is 243. It is an evidence of collusive behaviour in a Stackelberg duopoly, because we can observe the joint profit of a Stackelberg leader and a Stackelberg follower to be significantly more than the joint profit in equilibrium of Stackelberg duopoly. Thus, it holds definition of collusive behaviour we wrote above.

If we count, what is a ratio of values that are smaller than 0,5 we obtain ratio of choosing sequential play rather than simultaneous. The value higher than 0,5 means that more than 50 percent of simulations for a fixed pair *(learning rate, discount factor)* finished in convergence to a simultaneous play. Whereas, the value smaller than 0,5 means that more than 50% of simulations for a fixed pair *(learning rate, discount factor)* finished in convergence to a sequential play, what is exactly what we are looking for. For this case, the ratio of values smaller then 0,5 is 80%

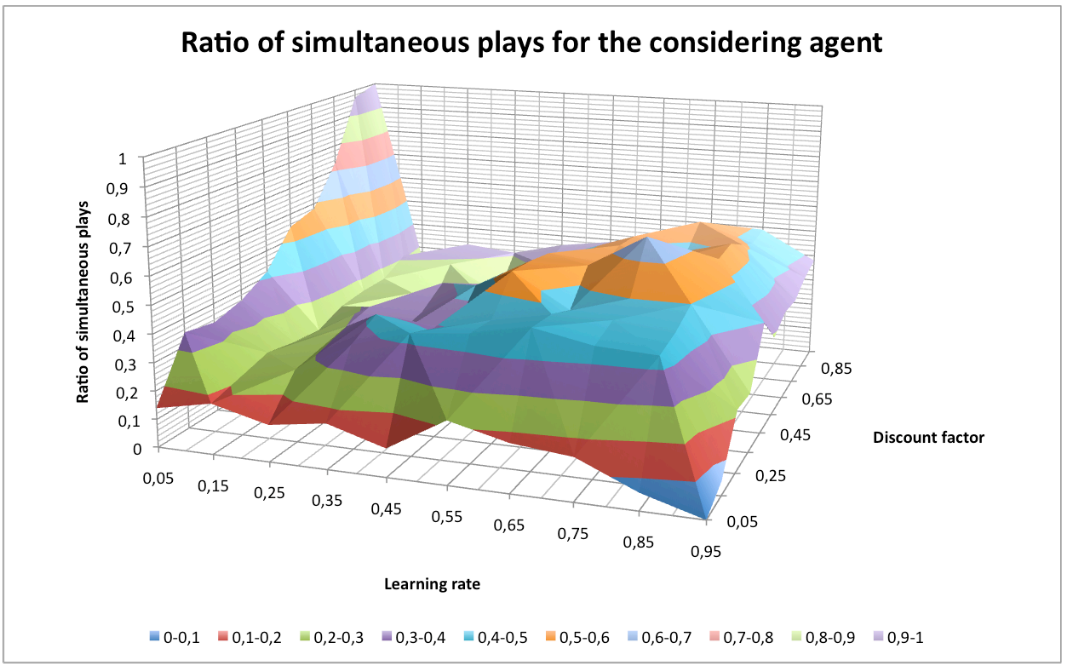

Chart 8 Ratio of simultaneous play for not considering agent dependent on the learning rate and the discount factor.

.

In general, we can say that for considering and not considering agents the not considering agent tends to choose be a Stackelberg follower. The considering agent always chooses the first period. However, we can conclude that collusion behaviour emerges in case when there is high ratio of a Stackelberg duopoly. When there is no clear choice of choosing either simultaneous or sequential play, the joint profit suffers by that situation.

Two considering agents

Chart 9 displays joint profit for two considering agents in the market with two rounds. Contrary to case of considering and not considering agents almost all values of joint profit are higher than joint profit of agents in Nash equilibrium. This is a big difference if we realize that in previous case just the highest values was similar to Nash equilibrium.

In general we can say, that the learning rate is more important than the discount factor, because we witch changing of the discount rate we cannot see any significant shifts in joint profit values. Whereas, there is a significant dependency as far as the learning rate is concerned and we can see, that the highest level of joint profit is connected with the highest rate of the learning rate.

From Chart 10 we can see, that whether two considering agents plays simultaneously or not is mainly dependent on the discount factor. The higher the discount factor is, the higher is ratio of simultaneous plays. We can see, that the learning rate also influences this ration, but not that significantly as the discount factor.

If we count, what is a ratio of values that are smaller than 0,5 we obtain ratio of choosing sequential play rather than simultaneous. The value higher than 0,5 means that more than 50 percent of simulations for a fixed pair *(learning rate, discount factor)* finished in convergence to a simultaneous play. Whereas, the value smaller than 0,5 means that more than 50 percent of simulations for a fixed pair *(learning rate, discount factor)* finished in convergence to a sequential play, what is exactly what we are looking for. For this case, the ratio of values smaller then 0,5 is 69%.

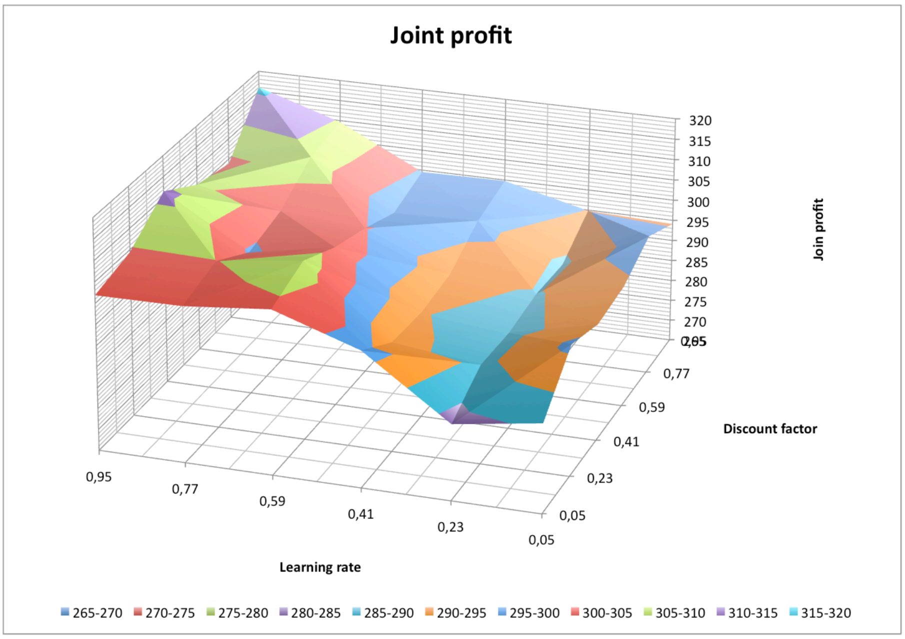

Chart 9 Joint profit of two considering agents and its dependency on the learning rate and the discount factor with.

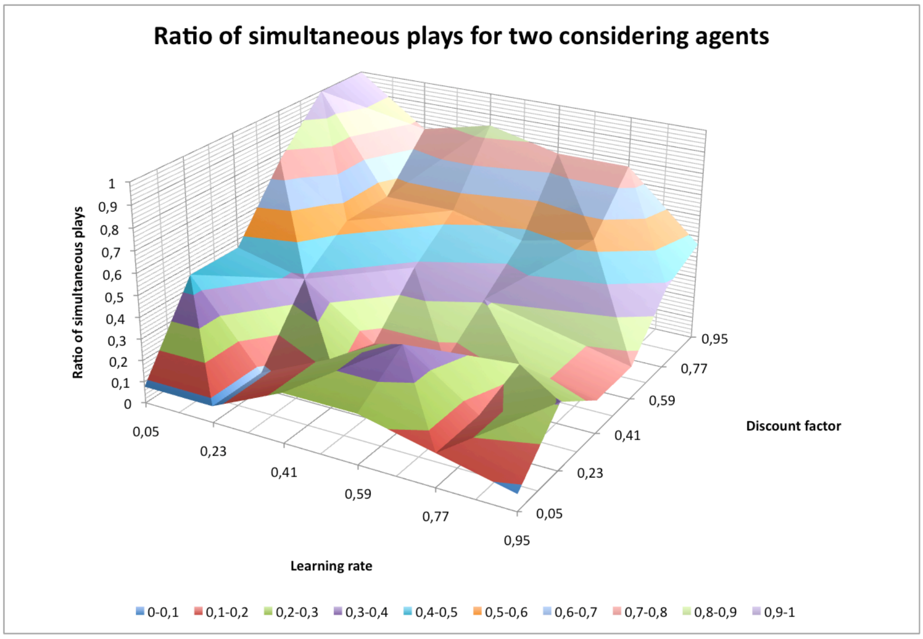

Chart 10 Ratio of simultaneous plays for two considering agents and its dependency on the learning rate and the discount factor

In general we can say, that emerge of sequential game, therefore emerge of a Stackelberg duopoly, is more probable in the case of two considering agents.

Two not considering agents

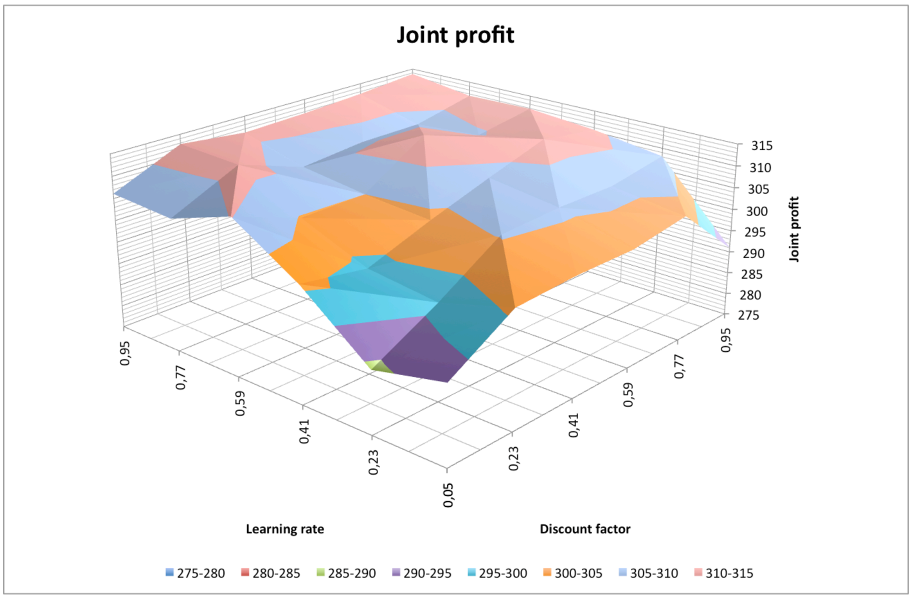

Chart 11 Joint profit of two not considering agents on the market with two periods dependent on the learning rate and discount factor.

Chart 11 displays the joint profit of two not considering agents interacting on the market with two periods. As we can see, almost all joint profits are above Nash equilibrium. This is an implication of Chart 12, when we can see that both not considering agents acts quite same as agents acting on the market with just one period, because for the majority of pairs *(learning rate, discount factor)* emerges simultaneous play in all simulations.

What is different in comparison with agents acting in the market with one period is that we have slightly different updating of the Q-values, but as we can see, we obtain collusive behaviour of agents as well.

As a conclusion we can say, that two not considering agents in the market with two turns do not converge in a Stackelberg duopoly at all.

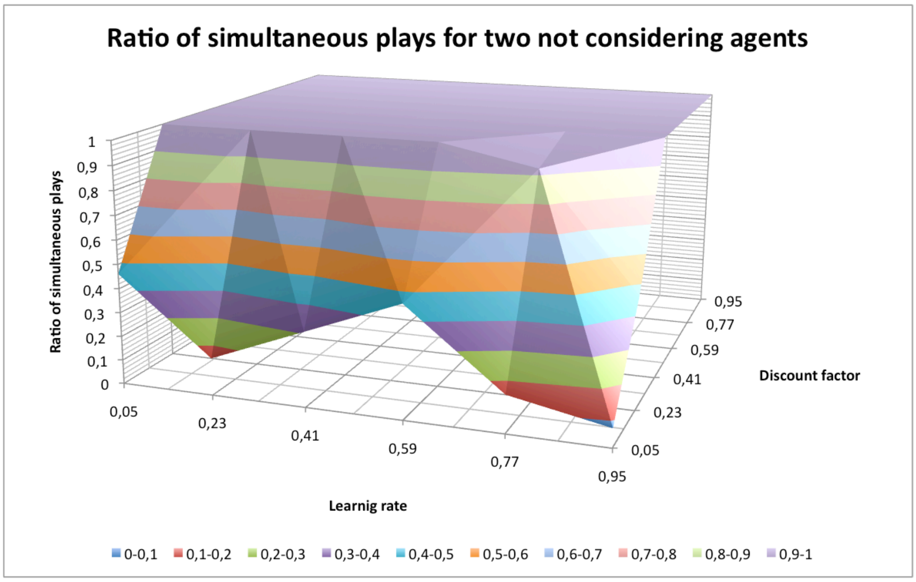

Chart 12 Ratio of simultaneous plays for two not considering agents dependent on the learning rate and the discount factor

5. Conclusion

Agent-based computational economics brings us new manners how to model economic situations. With bottom-up modelling methodology, we can take a closer look on processes, which leads to some economic situations, on the level of each interacting entity and thus earn a better understanding. We do not need to understand how process works, all the modeller need to do is correctly simulate agent's behaviour, e.g. profit maximization. Strict conditions of perfect rationality, perfect information and thus homogenity of economic entities substitute agent-based computational economics with more realistic assumptions of "bounded" rationality, imperfect information and heterogenity of agents.

One of our questions at the beginning was, whether ACE can simulate collusive behaviour in a Cournot duopoly. Following the example of Waltman and Kaymak [9] we successfully implemented model of a Cournot duopoly. The agent in the model followed his goal, what was profit maximization. This assumption corresponds with the real world indeed. We also simulated "bounded" rationality by using a logit model as a core of decision making process, through so-called Boltzman exploration strategy. Obtained profit was only information about the external world the agent was able to observe, so we simulated imperfect information. We created two different kinds of the agent – with and without a memory. Both kinds of agents were able to learn how to choose quantities and we could see the uprise of collusive behaviour, as the joint profit of agents was higher that the joint profit of agents in a Nash equilibrium.

We were also concerned in a question if between agents whose are allowed endogenously choose the period they want to produce in, arises a Cournot duopoly, a Stackelberg duopoly or mixture of both without any apparent tendency to one of them. Cournot duopoly model was extended, possibility to choose whether to produce in the first or the second period was added. Each round must agents choose period first and then quantity produced. Similarly, as in a simulations of market with one period, we modelled the agent with and without a memory. Agent with a memory can be further divided into the agent who considers the current game to the next one and the agent who does not consider the current game to the next one.

The heterogenity of economic entities is another hard nut to crack for traditional economic models, but with ACE we were able easily simulate interaction of considering and not considering agent. This possibility brought us a result that emergence of a Stackelberg duopoly was clearly present if at least one of the interacting agents was considering agent.

All in all, the main goal of this thesis, to simulate collusive behaviour in a Cournot model and look at the emergence of a Stackelberg duopoly for agents with endogenous timing of production was successfully satisfied.

Bibliography

1. Alkemade F.: *Evolutionary Agent-Based Economics*. Technische Universiteit Eindhoven, 2004.
2. Axelrod R., Tesfatsion L.: *On-Line Guide for Newcomers to Agent-Based Modeling in the Social Sciences*. Iowa State University, 2007, viewed 1st March 2009, http://www.econ.iastate.edu/tesfatsi/abmread.htm.
3. Chen S.-H.: *Evolutionary Computation in Economics and Finance*. Springer, 2002.
4. Gravelle H., Rees R.: *Microeconomics*. Pearson Education, 2004.
5. Hamilton J. H., Slutsky S. M.: *Endogenous Timing in Duopoly Games: Stackelberg or Cournot Equilibria*. Journal of Economic Literature, c.n. 026, 611, 1988.
6. Huck S., Norman H.-T., Oechssler J.: *Two are few and four are many: number effects in experimental oligopolies*. Journal of Economic Behaviour and Organization 53, pg. 435-446, 2004.
7. Tesfatsion L.: *Agent-Based Computational Economics: A Constructive Approach to Economic Theory*. Economic Department, Iowa State University, 2005.
8. Tesfatsion L.: *Agent-Based Computational Economics, Growing Economies from Bottom Up*. Material link collection, Iowa State University, 2009, viewed 14 May 2009, http://www.econ.iastate.edu/tesfatsi/ace.htm.
9. Waltman L., Kaymak U.: *Q-learning agents in a Cournot oligopoly model*. Journal of Economic Dynamics & Control 32, pg. 3275–3293, 2008.