On Decentralized Implicit Negotiation in Modified Ultimatum Game

Jitka Homolová, Eliška Zugarová, Miroslav Kárný, and Tatiana V Guy

Department of Adaptive Systems
The Czech Academy of Sciences,
Institute of Information Theory and Automation
P.O. Box 18, 182 08 Prague 8

{jkratoch,school,guy}@utia.cas.cz, zugareli@gmail.com

Abstract. Cooperation and negotiation are important elements of human interaction within extensive, flatly organized, mixed human-machine societies. Any sophisticated artificial intelligence cannot be complete without them. Multi-agent system with dynamic locally independent agents, that interact in a distributed way is inevitable in majority of modern applications. Here we consider a *modified* Ultimatum game (UG) for studying negotiation and cooperation aspects of decision making. The manuscript proposes agent's optimizing policy using Markov decision process (MDP) framework, which covers *implicit* negotiation (in contrast with explicit schemes as in [5]). The proposed solution replaces the classical game-theoretical design of agents' policies by an adaptive MDP that is: i) more realistic with respect to the knowledge available to individual players; ii) provides a first step towards solving negotiation essential in conflict situations.

Keywords: cooperation, negotiation, economic game, ultimatum game, Markov decision process

1 Introduction

Human intelligence develops in a context of different social interactions (within family, school, etc.) and their importance can hardly be overestimated. Artificial intelligence cannot be complete without supporting cooperation and negotiation as the basics of any dynamic (repetitive, history-dependent). Many applications in the fields of Artificial Intelligence, Computer Science and Economics (cooperative control, flocking, distributed sensor networks, communication networks, transportation) represent multi-agent systems with dynamic locally independent agents that need to coordinate to reach certain quantities of interest for mutual satisfaction. Dynamics, i.e. significant dependence of the actual decision-making task on the past history of agents' interactions is one of the important characteristics of these systems. Studying cooperation without a facilitator can bring us a better understanding of human behavior [9, 3] as well as an improving of overall performance of distributed computer systems [12, 21].

Cooperative aspects can be well studied using economic games [9], [19]. Here we consider a modification of the Ultimatum game (UG) [10], which is an economic game with simple rules. Two players (hereinafter agents) are to decide how to divide some goods between themselves. In the standard UG roles of agents are asymmetric: one agent proposes a division and the second one either accepts it or decides that none of agents gets anything.

It has been reported [6] that human-players do not behave rationally and their decisions do not follow the normative strategy implied by game theory [18]. The irrationality of a human-player in the Ultimatum Game has been already discussed. The paper [7] suggests that the human

is also driven by the sense for fairness. The paper further introduces three models of player's distinguishing of an affection degree with which the player is under influence of his individual tendency to consider a social profit (fairness) or an economic profit. The experimental results [7] have shown people do behave rationally with the reward that respected fairness.

The considered modification of UG balances the roles of both agents. The good division then can be interpreted as cake splitting. Both agents make a request for a certain number of pieces of the cake. In case of the demands compatibility, the cake is divided accordingly and they both get what they asked for. Oppositely, in case of the number of pieces of the cake exceeding, they both get nothing. This modification is close to the Nash's bargaining game [13].

There are some examples illustrating the real-life use of bargaining in human decision making: negotiating over a price by a seller and a buyer, bargaining over a trade agreement by two companies, finding a balance between the needs of employees and the interests of the employer by a company and a labor union or immigration quotas debating by European countries. In all of these examples, the common objective of all parties is reaching a consensus.

There is a number of studies focused on cooperation in decision making [11], [20], [14]. However, a reliable solution, applicable to distributed facilitator-free scenario and especially to the problem of human decision-making, still does not exist. *The main goal of this paper is to make a step towards distributed negotiation allowing cooperation within a flatly organized, dynamic human-machine society.* This goal can be reached if the policy of each individual agent is designed from the agent's perspective only. This justifies a formulation when the co-players are taken as a part of the agent's environment modeled in a feasible way. The theory of Markov decision processes (MDP) provides such a feasible approach. The intention to reach unlimited scalability excludes the (otherwise relevant) framework of Bayesian games [8]. It should be emphasized that the article does not aim to challenge existing solutions of UG, but uses modified version of UG for studying negotiation aspects.

Section 2 introduces necessary notations, recalls Markov Decision Process (MDP) and outlines a solution concept. Section 3 introduces modified UG and formulates it in terms of MDP. The key step is defining reward function that *implicitly* motivates negotiation. Section 4 describes two different models of a non-optimizing agent. Section 5 describes simulation experiments and Section 6 summarizes the results and adds some remarks on future research directions.

2 Preliminaries

This section introduces notations and recalls necessary notions.

$\mathbb{N}$, $\mathbb{R}$	set of natural numbers, set of real numbers
$x \in \mathbf{X}$	value x from the set of values $\mathbf{X}$
x_t	value of x at discrete time t
$p(x\|y)$	conditional probability of random variable x conditioned on random variable y
$E[x]$	expectation of random variable x
$E[x\|y]$	conditional expectation of random variable x conditioned on random variable y

2.1 Markov Decision Process

The addressed problem is formulated as a discrete-time discrete-valued *Markov Decision Process* (MDP). Let us remind a single-agent MDP (a detailed description can be found in [17]).

Definition 1 (MDP). *The fully observable MDP is characterized by* $\{\mathbf{T}, \mathbf{S}, \mathbf{A}, p, R\}$*, where* $\mathbf{T} = \{1, 2, ..., N\}, N \in \mathbb{N}$*, is a set of decision epochs;* $\mathbf{S}$ *is a set of all possible system states and*

$\mathbf{A}$ *denotes a set of all actions available to the agent. Function* $p : \mathbf{S} \times \mathbf{A} \times \mathbf{S} \mapsto [0, 1]$ *expresses the transition probability* $p_t(s_{t+1}|s_t, a_t)$ *that moves the system from state* $s_t \in \mathbf{S}$ *to state* $s_{t+1} \in \mathbf{S}$ *after an agent chooses action* $a_t \in \mathbf{A}$*;* $R : \mathbf{S} \times \mathbf{A} \times \mathbf{S} \mapsto \mathbb{R}$ *is a real-valued function representing the agent's reward* $r_t(s_{t+1}, s_t, a_t)$ *after taking action* $a_t \in \mathbf{A}$ *in state* $s_t \in \mathbf{S}$ *stimulating the transition to state* s_{t+1}*.*

The full observability means that the agent observes *precise* state s_{t+1} reached *after* action a_t is taken. This does not imply however that the agent is able to precisely predict future states.

In each decision epoch $t \in \mathbf{T}$, the agent chooses the action $a_t \in \mathbf{A}$ based on the *randomized DM rule* $p_t(a_t|s_t)$, which is a non-negative function representing a probability of the action a_t in the given state $s_t \in \mathbf{S}$. The agent's goal is to find an optimal *DM policy* π_t, that is a sequence of DM rules mapping states to actions to maximize expected reward over some horizon $h \in \{1, 2, ..., N - t\}$:

$$\pi_{t,h} = \left\{ p_\tau(a_\tau|s_\tau) \middle| s_\tau \in \mathbf{S}_\tau, a_\tau \in \mathbf{A}_\tau, \sum_{a_\tau \in \mathbf{A}_\tau} p_\tau(a_\tau|s_\tau) = 1, \forall s_\tau \in \mathbf{S}_\tau \right\}_{\tau=t}^{t+h}. \tag{1}$$

MDP with finite horizon h evaluates quality of the DM policy by an *expected total reward* $E\left[\sum_{\tau=t}^{t+h} r_\tau(s_{\tau+1}, s_\tau, a_\tau)|s_t\right]$.

In game round $t \in \mathbf{T}$ and state $s_t \in \mathbf{S}$, the *expected reward function* is defined as:

$$E_t\big[r_t(s_{t+1}, s_t, a_t)|s_t\big] = \sum_{s_{t+1} \in \mathbf{S}_{t+1} a_t \in \mathbf{A}_t} r_t(s_{t+1}, s_t, a_t) p_t(s_{t+1}, a_t|s_t), \tag{2}$$

where $p_t(s_{t+1}, a_t|s_t) = p_t(s_{t+1}|a_t, s_t)p_t(a_t|s_t)$.

The solution to MDP [17] is a sequence of functions $\left\{p_\tau^{opt}(a_\tau|s_\tau)\right\}_{\tau=t}^{t+h}$ that maximizes the expected reward and forms the optimal decision policy:

$$\pi_{t,h}^{opt} = \arg\max_{\{\pi_{t,h}\}} E\left[\sum_{\tau=t}^{t+h} r_\tau(s_{\tau+1}, s_\tau, a_\tau)|s_t\right]. \tag{3}$$

3 Modified Ultimatum Game

We use a modified version of the UG, a so-called cooperative UG scenario considering two agents $\mathcal{A}$ and $\mathcal{B}$ and an available amount of money (goods) q to split. Unlike the classical UG [18] roles of both players are the same. In the considered modification of the UG each round is treated as a round of an N-round repeated game. In the round, there is an action stage and a reward stage. During the *action* stage, each agent decides on own demand without knowing that of the co-player. Note that their interests are competitive. At the *reward* stage both agents get their rewards depending on whether amount q can cover the sum of their demands. In addition to plausibility for modeling cooperation and negotiation, the modified UG allows to adjust the amount of information the agents obtain (the degree of uncertainty), which is obviously an important aspect of the policy choice.

The overall scenario is as follows. At the beginning of the round $t \in \mathbf{T}$, each agent $k \in \{\mathcal{A}, \mathcal{B}\}$ chooses action $a_t^k \in \mathbf{A}^k$ that is a portion of q he wants to receive in this round. In case that demands sum up to less than q, both agents receive what they had asked for, otherwise, neither of the agents gets anything. Thus, the pure economic profit of agent $\mathcal{A}$ in round $t \in \mathbf{T}$ equals:

$$z_t^{\mathcal{A}} = a_t^{\mathcal{A}} \chi(a_t^{\mathcal{A}}, a_t^{\mathcal{B}}), \quad z_t^{\mathcal{A}} \in \mathbf{Z}_{\mathcal{A}} \tag{4}$$

where $\mathbf{Z}_{\mathcal{A}} = \{0, 1, 2, ..., q-1\}$ is a set of all possible profits of agent $\mathcal{A}$ in one game round, and

$$\chi(a_t^{\mathcal{A}}, a_t^{\mathcal{B}}) = \begin{cases} 1 & \text{if } a_t^{\mathcal{A}} + a_t^{\mathcal{B}} \leq q, \\ 0 & \text{if } a_t^{\mathcal{A}} + a_t^{\mathcal{B}} > q. \end{cases} \tag{5}$$

3.1 Modified UG as MDP

Interactive nature of the game implies that effect of the agent's individual actions depends on actions of the opponent. They are perceived and also modified in dependence on the game history, here, for simplicity restricted to Markovian case. We are interested in a *distributed* solution of the cooperation and consider the game from a point of view of a single agent $\mathcal{A}$. Generally the whole approach can be applied to agent $\mathcal{B}$ too as the process is fully observable to each of them and they may work with the structurally same reward function.

Definition 2 (Modified UG as MDP of the Agent $\mathcal{A}$). *The modified UG is modeled by* $\{\mathbf{T}, \mathbf{S}, \mathbf{A}, p, R\}$*, Definition 1, where*

- $\mathbf{A} = \mathbf{A}^{\mathcal{A}} = \{1, 2, ..., q-1\}$ *is a set of all possible actions;* $a_t^{\mathcal{A}} \in \mathbf{A}^{\mathcal{A}}$
- $s_t = (a_{t-1}^{\mathcal{A}}, a_{t-1}^{\mathcal{B}}) \in \mathbf{S}$*, where* $a_t^{\mathcal{B}} \in \mathbf{A}^{\mathcal{B}}$ *is a portion of* q *demanded by agent* $\mathcal{B}$ *at* $t \in \mathbf{T}$*,* $\mathbf{A}^{\mathcal{B}} = \{1, 2, ..., q-1\}$
- *initial state* $s_1 = (a_0^{\mathcal{A}}, a_0^{\mathcal{B}})$ *is preset to a fair offer corresponding to half of* q*, i.e.* $a_0^{\mathcal{A}} = a_0^{\mathcal{B}} = \frac{q}{2}$
- *reward of agent* $\mathcal{A}$ *is defined by* $r_t = a_t^{\mathcal{A}} \chi(a_t^{\mathcal{A}}, a_t^{\mathcal{B}}) - \omega^{\mathcal{A}} \mid q - (a_t^{\mathcal{A}} + a_t^{\mathcal{B}}) \mid$*, where* $\omega^{\mathcal{A}} \in [0, 1]$ *is a weight reflecting the degree of cooperation of the agent* $\mathcal{A}$*.*

In the definition of the reward function, the first term represents pure economical profit, cf. (4). The second term expresses "unused potential" if some monetary amount left after division; and "overshoot" in case when sum of demands being greater than q. The reward thus equals $a_t^{\mathcal{A}} - \omega^{\mathcal{A}} \cdot \mid q - (a_t^{\mathcal{A}} + a_t^{\mathcal{B}}) \mid$ if $\chi(a_t^{\mathcal{A}}, a_t^{\mathcal{B}}) = 1$ and $-\omega^{\mathcal{A}} \cdot \mid q - (a_t^{\mathcal{A}} + a_t^{\mathcal{B}}) \mid$ when no consensus has been reached.

The cooperativeness weight $\omega^{\mathcal{A}}$ depends on the personality traits of an agent and reflects importance of the second term in the agent's reward definition.Thus, $\omega^{\mathcal{A}}$ is specific and private for each agent working with it. Value $\omega^{\mathcal{A}} = 0$ implies reward depending on pure economical profit, while $\omega^{\mathcal{A}} > 0$ makes the designed strategy to be respecting the degree of influence of "overshoot" as well as "unused potential" on the agent. This weight reflects the human style of playing and therefore the optimal policy is implicitly forced by the discussed summand to take into account actions of the co-player, i.e. to cooperate.

The optimal policy, see Definition 2, can be computed by solving classical MDP using a dynamic programming algorithm [2], [16]. It needs the transition probability $p(s_{t+1}|a_t^{\mathcal{A}}, s_t)$. Conditional independence of agents' actions and the definition of the state imply

$$p(s_{t+1}|a_t^{\mathcal{A}}, s_t) = p(a_{t+1}^{\mathcal{A}}|a_t^{\mathcal{A}}, a_t^{\mathcal{B}}) p(a_{t+1}^{\mathcal{B}}|a_t^{\mathcal{A}}, a_t^{\mathcal{B}}). \tag{6}$$

The first factor is a part of the optimized policy. The second factor models agent $\mathcal{B}$ and can be recursively estimated using Bayesian paradigm [15]. To simplify explanation, it is assumed to be given. Its considered forms are discussed below.

4 Models of the Agent $\mathcal{B}$

To bring better understanding of the role and the effect of the introduced reward (see Definition 2), two different models of the non-optimizing agent $\mathcal{B}$ are presented and further studied. Studies of learning and optimizing agents with different prior knowledge and different cooperativeness degrees are straightforward natural extensions of the current work and are supposed to be reported in the future. The goal is to verify whether considering or conforming the second agent's actions leads to a higher profit and a higher number of successful rounds comparing to a non-cooperative strategy.

In each round, *the non-optimizing agent* $\mathcal{B}$ heuristically adjusts demands according to the results of the previous round. Generally, if the sum of both agents' demands is greater than available amount q, the agent $\mathcal{B}$ lowers the demand by a portion of the excess to avoid repeating the failure. Oppositely, if q is not used up and some amount left, the agent $\mathcal{B}$ raises the demand by a portion of the amount left to increase the chance of distributing of the whole amount q in the current round.

The proposed models of the non-optimizing agent $\mathcal{B}$ differ in the degree of an aggression of the agent or, in other words, in the extent to which the demand changes. In both models, we assume that the transition probabilities do not change during the whole game.

4.1 Type B1

The non-optimizing agent $\mathcal{B}$ of type B1 represents the basic concept of the agent $\mathcal{B}$. This type of the agent is less aggressive and more fair. The DM aim of this agent's type is to reduce the difference between q and the sum of demands from the previous round by roughly one half. This means that in the round t, the agent $\mathcal{B}$ increases or decreases next demand by $\frac{1}{2}[q-(a^{\mathcal{A}}_{t-1}+a^{\mathcal{B}}_{t-1})]$ accordingly to the "unused potential" or the "overshoot" of the available amount q.

The transition probability used for the model of the agent $\mathcal{B}$ of the type B1 can be modeled by the discretized Gaussian probability distribution with the corresponding expected mode $\frac{1}{2}(q-a^{\mathcal{A}}_{t-1}+a^{\mathcal{B}}_{t-1})$:

$$p_t(a^{\mathcal{B}}_t|a^{\mathcal{A}}_{t-1},a^{\mathcal{B}}_{t-1}) \propto \exp\left(-\frac{\left(a^{\mathcal{B}}_t-\frac{1}{2}(q-a^{\mathcal{A}}_{t-1}+a^{\mathcal{B}}_{t-1})\right)^2}{2\sigma^2}\right), \tag{7}$$

where $a^{\mathcal{A}}_{t-1}, a^{\mathcal{B}}_{t-1} \in \mathbf{A}$ are the demands of the agents at round $(t-1) \in \mathbf{T}$, and σ^2 is the variance of the discrete Gaussian distribution.

Such behavior can be explained by an effort to act fairly in terms of utilizing the whole amount q. The type B1 agent anticipates that the agent $\mathcal{A}$ step back by a half of an excess or raise their demands by a half of the remained part of the available amount q. However, the agent $\mathcal{B}$ does not take into consideration the difference between the individual demands. In case of the large difference, the split then becomes unfair.

Even though the model (7) serves the creation of an active opponent of the agent $\mathcal{A}$, the agent's strategy does not correspond well to human thinking [3].

4.2 Type B2

Proposed type B2 is more complex and efficient variant of the non-optimizing agent $\mathcal{B}$. It reflects human style of playing. As described in the previous subsection, the type B1 agent adapts the actions in dependence on the difference between the sum of demands and the available amount q. For the type B2 agent, change of demand depends on the agent $\mathcal{A}$ action from the previous

round. The action range of the agent $\mathcal{A}$ is divided into three regions for which the offset of the new demand of the agent $\mathcal{B}$ is set differently.

Let $q = 10$, $t \in \mathbf{T}$ be the current game round and $a^{\mathcal{A}}_{t-1}$, $a^{\mathcal{B}}_{t-1} \in \mathbf{A}$ be the demands from the previous game round. Then demand $a^{\mathcal{A}}_{t-1}$ can belong to the following sets: $\{1, 2, 3\}$, $\{3, 4, 5\}$ and $\{6, 7, 9\}$. Transition probabilities related to the type B2 are set for each set as follows.

Case 1: The demand of the agent $\mathcal{A}$ when $a^{\mathcal{A}}_{t-1} \in \{1, 2, 3\}$.

1. If $q - (a^{\mathcal{A}}_{t-1} + a^{\mathcal{B}}_{t-1}) \geq 0$, i.e. some amount of q is left in the previous round, then the agent of the type B2 increases the demand by one third of the part of q lost in round $(t-1)$:
$$p_t(a^{\mathcal{B}}_t | a^{\mathcal{A}}_{t-1}, a^{\mathcal{B}}_{t-1}) \propto \exp\left(-\frac{\left(a^{\mathcal{B}}_t - \frac{1}{3}(q - a^{\mathcal{A}}_{t-1} + 2a^{\mathcal{B}}_{t-1})\right)^2}{2\sigma^2}\right).$$
2. If $q - (a^{\mathcal{A}}_{t-1} + a^{\mathcal{B}}_{t-1}) < 0$, i.e. the agents' demands exceed amount q in the previous round, the agent of the type B2 decreases the demand by two thirds of the amount that exceeded q:
$$p_t(a^{\mathcal{B}}_t | a^{\mathcal{A}}_{t-1}, a^{\mathcal{B}}_{t-1}) \propto \exp\left(-\frac{\left(a^{\mathcal{B}}_t - \frac{2}{3}(q - a^{\mathcal{A}}_{t-1} + \frac{1}{2}a^{\mathcal{B}}_{t-1})\right)^2}{2\sigma^2}\right).$$

Case 2: The demand of the agent $\mathcal{A}$ when $a^{\mathcal{A}}_{t-1} \in \{4, 5, 6\}$.

1. The agent of the type B2 behaves identically to the type B1 (7):
$$p_t(a^{\mathcal{B}}_t | a^{\mathcal{A}}_{t-1}, a^{\mathcal{B}}_{t-1}) \propto \exp\left(-\frac{\left(a^{\mathcal{B}}_t - \frac{1}{2}(q - a^{\mathcal{A}}_{t-1} + a^{\mathcal{B}}_{t-1})\right)^2}{2\sigma^2}\right).$$

Case 3: The demand of the agent $\mathcal{A}$ when $a^{\mathcal{A}}_{t-1} \in \{7, 8, 9\}$.

1. If $q - (a^{\mathcal{A}}_{t-1} + a^{\mathcal{B}}_{t-1}) \geq 0$, the agent of the type B2 raises the demand by two thirds of the amount not distributed in round $(t-1)$:
$$p_t(a^{\mathcal{B}}_t | a^{\mathcal{A}}_{t-1}, a^{\mathcal{B}}_{t-1}) \propto \exp\left(-\frac{\left(a^{\mathcal{B}}_t - \frac{2}{3}(q - a^{\mathcal{A}}_{t-1} + \frac{1}{2}a^{\mathcal{B}}_{t-1})\right)^2}{2\sigma^2}\right).$$
2. If $q - a^{\mathcal{A}}_{t-1} - a^{\mathcal{B}}_{t-1} < 0$, then the agent of the type B2 lowers the demand by one third of the exceeded amount:
$$p_t(a^{\mathcal{B}}_t | a^{\mathcal{A}}_{t-1}, a^{\mathcal{B}}_{t-1}) \propto \exp\left(-\frac{\left(a^{\mathcal{B}}_t - \frac{1}{3}(q - a^{\mathcal{A}}_{t-1} + 2a^{\mathcal{B}}_{t-1})\right)^2}{2\sigma^2}\right).$$

Case 1 can be interpreted as the type B2 agent either decreases the demand in case of an excess to prevent it happening again, or the agent increases the demand by a smaller amount in case of not spending the full amount q because of the tolerance assumption of the agent $\mathcal{A}$.

In Case 2, the type B2 agent expects the fairness of the agent $\mathcal{A}$, which means that such agent anticipates from the agent $\mathcal{A}$ to ask for only a half of the amount left or to drop a half of the excess.

In Case 3, the type B2 agent tries to stop the agent $\mathcal{A}$ from continuing to take actions of the high value (the actions from the region $\{7, 8, 9\}$), because these actions lead to an unfair splitting of the amount q. By this reason, the agent asks for the larger portion of the amount lost and is willing to retreat less in case of the excess.

Such model of the agent $\mathcal{B}$ better corresponds to the anticipated behavior of a human [7, 9].

5 Illustrative Experiments

The proposed approach is illustrated on several simulated examples ran in Matlab environment. Each game had $N = 30$ rounds. The available amount to split was $q = 10$CZK[1] in each round. The simulation ran for five different values of the cooperativeness weight $\omega^{\mathcal{A}} \in \{0, 0.25, 0.5, 0.75, 1\}$. The reward (see Definition 2) used the design of the optimal policy of the agent $\mathcal{A}$ that was found by the dynamic programming [2]. The weight $\omega^{\mathcal{A}}$ reflected individual tendency of the agent $\mathcal{A}$ to consider negotiations ($0 < \omega^{\mathcal{A}} \leq 1$) or not ($\omega^{\mathcal{A}} = 0$).

In case of $\omega^{\mathcal{A}} = 0$, the agent $\mathcal{A}$ is interested only in the monetary profit and does not cooperate. In case of $0 < \omega^{\mathcal{A}} \leq 1$, the agent is also, to some degree, committed to use the potential of each round. It means that the agent tries do not exceed the available amount q and to split it fully. This weight $\omega^{\mathcal{A}}$ does *not* express a balance between the importance of the economic profit and the cooperation.

The first example is a game of two non-optimising agents (so-called type $\mathcal{B}$ agents, see Section 4). We need it for the later comparison with a game when one of the players cooperates and optimizes.

In the current example, one agent was of the type B1 (see Section 4.1) while another of the type B2 (see Section 4.2), so the transition probabilities of both models were proportional to the Gaussian distribution. Let's recall the agent of the type B2 is a more complex version of the agent of the type B1 whose demand depends on the previous action of the opponent. The standard deviation of $\sigma = 3$ was chosen and the seed parameter for the reproducibility was set to the value of 90 during the simulation. The numerical results are summarized in the Table 1. Plots of the cumulative profits of the agents and their corresponding actions are shown in Fig. 1 and Fig. 2.

Standard deviation $\sigma = 3$	
Percentage of the successful rounds (%)	63.3
Total profit of the agent $\mathcal{B}$ of the type B1 (CZK)	71
Total profit of the agent $\mathcal{B}$ of the type B2 (CZK)	85

Table 1. Game of type B1 agent and type B2 agent

As it can be seen from the results, the higher flexibility to the actual situation of the agent $\mathcal{B}$ of the type B2 causes the better choice of actions and so the reaching the higher total profit of such agent type.

The second example represents a set of the games with the optimizing agent $\mathcal{A}$ and the non-optimizing agent $\mathcal{B}$ of the type B1 (see Section 4.1). The simulation ran with five values of the cooperativeness weight $\omega^{\mathcal{A}}$ and three different values of the standard deviation $\sigma \in \{3, 4, 5\}$ in the model of the agent $\mathcal{B}$. The seed parameter was set to 13. The numerical results can be seen in Table 2. The progress of the cumulative profits of the agents for $\omega^{\mathcal{A}} = 0.25$ and $\sigma = 3$ is plotted in Fig. 3.

The last example focuses on a set of the games with the optimizing agent $\mathcal{A}$ and the non-optimizing agent $\mathcal{B}$ of the type B2 (see Section 4.2). The simulation also ran for five values of the cooperativeness weight $\omega^{\mathcal{A}}$ and three values of the standard deviation $\sigma \in \{3, 4, 5\}$. The seed parameter for the reproducibility was set to the value of 20. The numerical results are presented

[1] Czech crowns

Standard deviation $\sigma = 3$					
Weight $\omega^{\mathcal{A}}$	0.00	0.25	0.50	0.75	1.00
Percentage of the successful rounds (%)	70.0	76.7	70.0	70.0	66.7
Total profit of the agent $\mathcal{A}$ (CZK)	126	137	129	131	121
Total profit of the agent $\mathcal{B}$ (CZK)	61	68	64	62	56
Standard deviation $\sigma = 4$					
Weight $\omega^{\mathcal{A}}$	0.00	0.25	0.50	0.75	1.00
Percentage of the successful rounds (%)	70.0	76.7	70.0	76.7	76.7
Total profit of the agent $\mathcal{A}$ (CZK)	119	133	126	126	128
Total profit of the agent $\mathcal{B}$ (CZK)	66	75	75	75	75
Standard deviation $\sigma = 5$					
Weight $\omega^{\mathcal{A}}$	0.00	0.25	0.50	0.75	1.00
Percentage of the successful rounds (%)	76.7	73.3	73.3	73.3	73.3
Total profit of the agent $\mathcal{A}$ (CZK)	122	116	121	123	119
Total profit of the agent $\mathcal{B}$ (CZK)	78	73	70	70	71

Table 2. Game with the optimizing agent $\mathcal{A}$ and the agent $\mathcal{B}$ of the type B1.

in Table 3. The course of the cumulative profits of the agents, using $\omega^{\mathcal{A}} = 0.25$ and $\sigma = 3$, are plotted in Fig. 5. The actions of the agents can be seen in Fig. 6.

Standard deviation $\sigma = 3$					
Weight $\omega^{\mathcal{A}}$	0.00	0.25	0.50	0.75	1.00
Percentage of the successful rounds (%)	53.3	60.0	60.0	50.0	50.0
Total profit of the agent $\mathcal{A}$ (CZK)	88	102	100	87	86
Total profit of the agent $\mathcal{B}$ (CZK)	61	67	67	53	53
Standard deviation $\sigma = 4$					
Weight $\omega^{\mathcal{A}}$	0.00	0.25	0.50	0.75	1.00
Percentage of the successful rounds (%)	63.3	60.0	56.7	56.7	53.3
Total profit of the agent $\mathcal{A}$ (CZK)	99	94	92	91	88
Total profit of the agent $\mathcal{B}$ (CZK)	76	71	62	64	58
Standard deviation $\sigma = 5$					
Weight $\omega^{\mathcal{A}}$	0.00	0.25	0.50	0.75	1.00
Percentage of the successful rounds (%)	73.3	56.7	56.7	56.7	56.7
Total profit of the agent $\mathcal{A}$ (CZK)	111	88	89	91	91
Total profit of the agent $\mathcal{B}$ (CZK)	91	62	62	62	61

Table 3. Game with the optimizing agent $\mathcal{A}$ and the agent $\mathcal{B}$ of the type B2.

The results show that the optimization positively leads to the higher profit independently of the cooperativeness weight. Once again, the higher flexibility to the actual situation of the agent $\mathcal{B}$ of the type B2 brings the possibility to assert at the expense of the agent $\mathcal{A}$. Another interesting point is that too much effort to maximize the potential of each game round leads to a worsening of overall results that can be seen mainly from the percentage of the successful rounds.

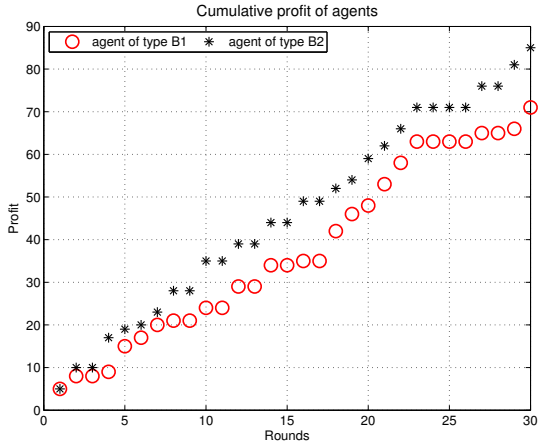

Fig. 1. Cumulative profits: non-optimizing agents, player 1 is of the type B1 and player 2 is of the type B2.

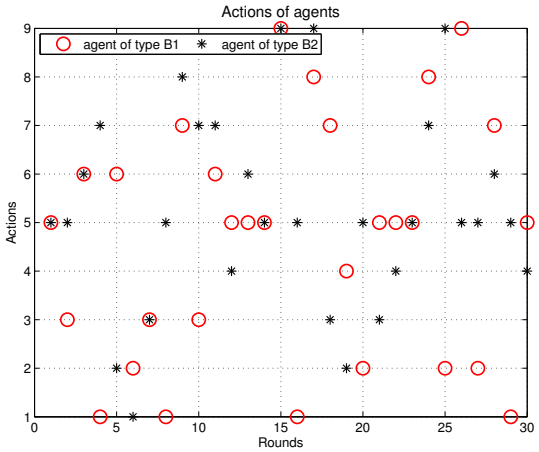

Fig. 2. Actions: non-optimizing agents, player 1 is of the type B1 and player 2 is of the type B2.

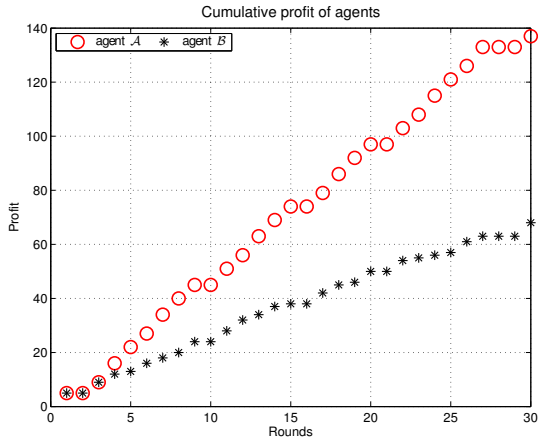

Fig. 3. Cumulative profits: player 1 is the optimizing agent $\mathcal{A}$ and player 2 is the non-optimizing agent $\mathcal{B}$ of the type B1.

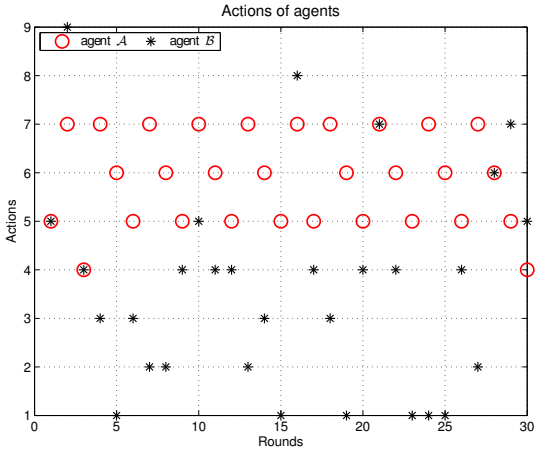

Fig. 4. Actions: player 1 is the optimizing agent $\mathcal{A}$ and player 2 is the non-optimizing agent $\mathcal{B}$ of the type B1.

6 Concluding remarks

The paper introduces an innovative approach to the implicit cooperation in the modified UG. The optimizing agent is modeled via Markov decision process. The key step of the proposed approach is the special reward function (see Definition 2) respecting not only the economic profit of the agent but also reflecting the agent's willingness for the cooperation. Such a reward function implicitly forces the agent to negotiate. By optimizing overall behavior of the agents their optimal degree of cooperation can be searched for.

The adopted approach was driven by our main concern: the search for a *feasible, fully scalable,* design of approximately optimal DM policy under the need for a cooperation. The inspected framework also allows respecting the influence of such subtle phenomena as emotions in decision

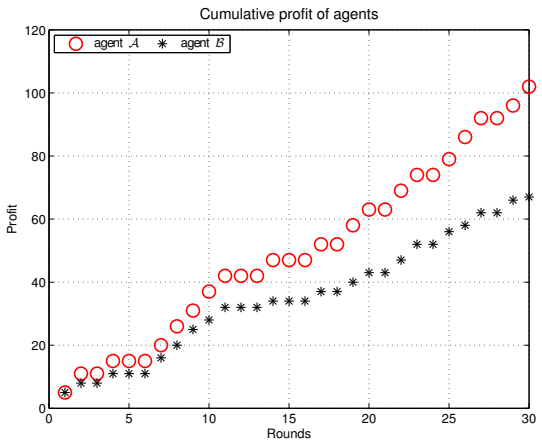

Fig. 5. Cumulative profits: player 1 is the optimizing agent $\mathcal{A}$ and player 2 is the non-optimizing agent $\mathcal{B}$ of the type B2.

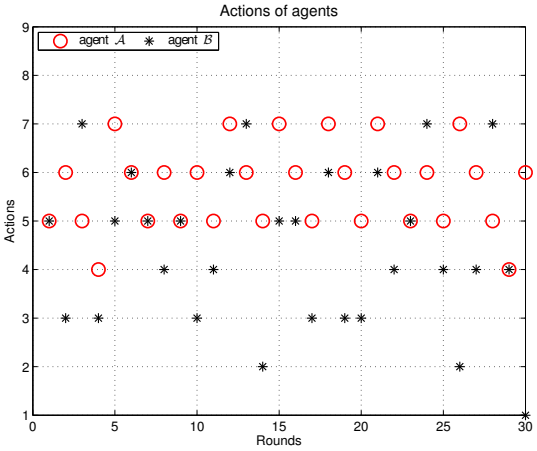

Fig. 6. Actions: player 1 is the optimizing agent $\mathcal{A}$ and player 2 is the non-optimizing agent $\mathcal{B}$ of the type B2.

making [1]. In this respect, the gained results are the first promising step in creating the applicable DM strategies allowing for the *fully scalable* distributed cooperation and negotiation in the extensive *mixed human-machine* societies.

The foreseen research will consider: i) different types of the agents (learning, optimizing ones, differing in cooperativeness degree, using more realistic non-symmetric reward [4] and others); ii) including the fairness aspects into the reward function [7]; iii) dependence of economic profit on the cooperativeness weight; iv) learning the co-player's model and its cooperativeness weights.

Acknowledgement: The research has been supported by the project GA16-09848S.

References

1. Avanesyan, G., Kárný, M., Knejflová, Z., Guy, T.V.: Demo: What lies beneath players' non-rationality in ultimatum game? In: Preprints of the 3rd Int. Workshop on Scalable Decision Making held in conjunction with ECML-PKDD 2013. ÚTIA AVČR, Prague, Czech Republic (2013)
2. Bellman, R.E.: Dynamic Programming. Princeton University Press, Princeton, NJ (1957)
3. Fehr, E., Fischbacher, U., Gächter, S.: Strong reciprocity, human cooperation, and the enforcement of social norms. Human Nature 13, 1–25 (2002)
4. Fehr, E., Schmidt, K.: A theory of fairness, competition, and cooperation. Quarterly Journal of Economics 114(3), 817–868 (1999)
5. Goranko, V., Turrini, P.: Non-cooperative games with preplay negotiations. CoRR abs/1208.1718 (2012)
6. Güth, W., Schmittberger, R., Schwarze, B.: An experimental analysis of ultimatum bargaining. Journal of Economic Behavior & Organization 3(4), 367 – 388 (1982)
7. Guy, T.V., Kárný, M., Lintas, A., Villa, A.: Theoretical models of decision-making in the ultimatum game: Fairness vs. reason. Springer Singapore pp. 185–191 (2016)
8. Harsanyi, J.: Games with incomplete information played by Bayesian players, I–III. Management Science 50(12) (2004), supplement
9. Haselhuhn, M.P., Mellers, B.A.: Emotions and cooperation in economic games. Cognitive Brain Research 23, 24–33 (2005)

10. Hula, F., M.Ruman, Kárný, M.: Adaptive proposer for ultimatum game. In: Villa, A., Masulli, P., J., A. (eds.) Artificial Neural Networks and Machine Learning ICANN 2016. Theoretical Computer Science and General Issues, vol. 9886–9887, pp. 330–338. Springer (2016)
11. Johansson, B., Speranzon, A., Johansson, M., Johansson, K.H.: On decentralized negotiation of optimal consensus. Automatica 44(4), 1175 – 1179 (2008)
12. Kraus, S.: Negotiation and cooperation in multi-agent environments. Artificial Intelligence 94(1–2), 79–97 (1997)
13. Nash, J.F.: The bargaining problem. Econometrica 18(2), 155–162 (1950)
14. Olfati-Saber, R., Fax, J., Murray, R.: Consensus and cooperation in networked multi-agent systems. Proceedings of the IEEE 95(1), 215–233 (2007)
15. Peterka, V.: Bayesian approach to system identification. In: Eykhoff, P. (ed.) In Trends and Progress in System Identification. pp. 239–304. Pergamon Press (1981)
16. Powell, W.B.: Approximate Dynamic Programming. John Wiley and Sons, 2nd edition edn. (2011)
17. Puterman, M.L.: Markov Decission Processes. John Wiley & Sons, Inc. (1994)
18. Rubinstein, A.: Perfect equilibrium in a bargaining model. Econometrica 50(1), 97–109 (1982)
19. Sanfey, A.G.: Social decision-making: Insights from game theory and neuroscience. Science 318, 598–602 (2007)
20. Semsar-Kazerooni, E., Khorasani, K.: Multi-agent team cooperation: A game theory approach. Automatica 45(10), 2205 – 2213 (2009)
21. Xuan, P., Lesser, V., Zilberstein, S.: Communication decisions in multiagent cooperation: Model and experiments. In: Proceedings of the Fifth International Conference on Autonomous Agents. pp. 616–623. Montreal, Canada (2001), http://rbr.cs.umass.edu/shlomo/papers/XLZagents01.html