Deliberation-aware Responder in Multi-Proposer Ultimatum Game

Marko Ruman, František Hůla, Miroslav Kárný, and Tatiana V. Guy*

Department of Adaptive Systems
Institute of Information Theory and Automation
Czech Academy of Sciences
P.O.Box 18, 182 08 Prague
{marko.ruman,hula.frantisek}@gmail.com
{school,guy}@utia.cas.cz

Abstract The article studies deliberation aspects by modelling a responder in multi-proposers ultimatum game (UG). Compared to the classical UG, deliberative multi-proposers UG suggests that at each round the responder selects the proposer to play with. Any change of the proposer (compared to the previous round) is penalised. The simulation results show that though switching of proposers incurred non-negligible deliberation costs, the economic profit of the deliberation-aware responder was significantly higher in multi-proposer UG compared to the classical UG.

Keywords: deliberation effort; Markov decision process; ultimatum game

1 Introduction

The role of deliberation in decision making (DM) has been addressed in many ways. Examples can be found elsewhere, see, for instance, political sciences [4], economy [6], behavioral science [3]. The reason is simple: any decision made either by human or machine costs time, energy and possibly other resources, which are always limited. Importance of the proper balance between deliberation and quality of the resulting decision is repeatedly confirmed by a considerable effort devoted within different communities: computation costs in computer sciences [8],[9]; transaction costs in financial sciences [7]; cooperation effort in social sciences [11], negotiation in multi-agent systems [10] and many others. Despite many promising results, see for instance recent work [5], the well-justified theoretical framework of deliberation is still missing.

The present article contributes to this problem by modelling a responder's DM in multi-proposers ultimatum game (UG) [2], introduces deliberation effort into reward function and optimises it. The simplicity of UG makes it a powerful test case providing a general insight into human DM, which can further serve

* This work was partially supported by the Grant Agency of the Czech Republic under the project GA16-09848S.

to other fields. The basic model of UG consists of two players (*proposer* and *responder*) having different roles. The proposer's task is to share some known amount of money between him and the responder. The responder's role is to accept or reject the proposal. Acceptance leads to splitting the money according to the proposal, whereas rejection means none player gets anything.

Compare to [2], deliberative multi-proposers UG scenario suggests that at each round the responder selects the proposer to play with. However any change of the proposer (compare to the previous round) is penalised. The responder has no or little information about the proposers, thus the responder's optimal strategy should maximise economic profit while minimising deliberation cost under incomplete knowledge. It should be stressed that such modification of UG scenario makes a sense only for the studying either deliberation aspects or cooperative aspects of human DM. In the last case, repetitive selecting/non-selecting serves as a kind of the responder's feedback to a particular proposer and may influence future decision policy of the proposer.

Markov decision process (MDP) framework [1] has proven to be very useful for DM in stochastic environments. The paper considers modelling the responder's DM in the deliberation-aware multi-proposer multi-round UG experiment by MPD formalism and describes how *the responder's deliberation effort* can be respected and optimised.

The paper layout is as follows. Section 2 introduces necessary notations and formulates the problem. Section 3 introduces an optimal solution. Section 4 specialises the reward function of economic responder playing in multi-proposer UG. The experimental setup is described in Section 5. Section 6 summarises the main results and discusses open problems and possible solution ways.

2 Problem Formulation

The section introduces notations and a basic concept of Markov Decision Process (MDP) necessary to solve our problem. For more background on MDP, see [1].

2.1 Preliminaries

Throughout the paper, we use x_t to denote value of x at discrete time labelled by $t = 1, \ldots, t \in \mathbb{N}$. Bold capitals $\mathbf{X}$ denote a set of x-values; an abbreviation *pd* means probability density function, $p_t(x|y)$ is a conditional pd. $\chi(x, y)$ is a function defined on $\mathbb{R} \times \mathbb{R}$ as $\chi(x, y) = \begin{cases} 1 & x \neq y, \\ 0 & x = y. \end{cases}$

MDP provides us a mathematical framework for describing an *agent* (decision maker), which interacts with a *stochastic system* by taking appropriate actions to achieve her goal. The decisions about actions are made in the points of time referred as *decision epochs*. In each decision epoch, the agent's decisions are influenced only by a state of the stochastic system in a particular decision epoch, not by history of the system.

2.2 Markov decision process

Definition 1 (Markov Decision Process). Markov Decision Process *over the discrete finite set of* decision epochs $\boldsymbol{T} = \{1, 2, ..., N\}$, $N \in \mathbb{N}$ *is defined by a tuple* $\{\boldsymbol{T}, \boldsymbol{S}, \boldsymbol{A}, p, r\}$, *where:*

- $\boldsymbol{S}$ *is a discrete, finite* state space $s \in \boldsymbol{S}$; $\boldsymbol{S} = \underset{t \in \boldsymbol{T}}{\cup} \boldsymbol{S}_t$, *where* $\boldsymbol{S}_t$ *is a set of possible states of the system at the decision epoch* $t \in \boldsymbol{T}$ *and* $s_t \in \boldsymbol{S}_t$ *is a state of the system at the decision epoch* $t \in \boldsymbol{T}$,
- $\boldsymbol{A}$ *stands for a discrete, finite* action set; $\boldsymbol{A} = \underset{t \in \boldsymbol{T}}{\cup} \boldsymbol{A}_t$, *where* $\boldsymbol{A}_t$ *is a set of admissible actions in the decision epoch* $t \in \boldsymbol{T}$ *and* $a_t \in \boldsymbol{A}_t$ *denotes chosen action in the decision epoch* $t \in \boldsymbol{T}$,
- p *represents a* transition probability function $p = p_t(s_t|s_{t-1}, a_t)$, *which is a non-negative function describing the probability that system reaches the state* s_t *after the action* a_t *is taken at the state* s_{t-1}; $\sum\limits_{s_t \in \boldsymbol{S}} p_t(s_t|s_{t-1}, a_t) = 1$, $\forall t \in \boldsymbol{T}, \forall a_t \in \boldsymbol{A}, \forall s_{t-1} \in \boldsymbol{S}_{t-1}$,
- r *stands for a* reward function $r = r_t(s_t, s_{t-1}, a_t)$, *which is used to quantify reaching of the agent's aim. The reward function* $r_t(s_t, s_{t-1}, a_t)$ *depends on the state* s_t *that the system occupies after action* a_t *is made.*

At the decision epoch t, an agent chooses an action a_t to be executed. As a result the system transits to a new state $s_t \in \mathbf{S}$ stochastically determined by $p_t(s_t|s_{t-1}, a_t)$. The agent gets a reward, which equals the value of reward function $r_t(s_t, s_{t-1}, a_t)$. The agent's goal is to find the optimal *DM policy*, which maximises the average reward received over time.

To avoid explicit dependence of the reward on the future state $s_t \in \mathbf{S}_t$ the *expected reward* is introduced as follows:

$$E_t[r_t(s_t, s_{t-1}, a_t)] = \sum_{a_t \in \mathbf{A}} \sum_{\substack{s_t \in \mathbf{S} \\ s_{t-1} \in \mathbf{S}}} r_t(s_t, s_{t-1}, a_t) p_t(s_t|s_{t-1}, a_t) p_t(a_t|s_{t-1}) p_t(s_{t-1}) \tag{1}$$

In (1), $p_t(a_t|s_{t-1})$ is a *randomised decision rule* satisfying the condition $\sum\limits_{a_t \in \mathbf{A}} p_t(a_t|s_{t-1}) = 1, \forall s_{t-1} \in \mathbf{S}_{t-1}, \forall t \in \mathbf{T}$.

Definition 2 (Stochastic policy). *A sequence of randomised decision rules* $\left\{ p_t(a_t|s_{t-1}) \Big| \sum\limits_{a_t \in \boldsymbol{A}} p_t(a_t|s_{t-1}) = 1, \forall s_{t-1} \in \boldsymbol{S}_{t-1}, \forall t \in \boldsymbol{T} \right\}$ *forms the stochastic policy* $\pi_t \in \boldsymbol{\pi}$, *where* $p_t(a_t|s_{t-1})$ *is the probability of action* a_t *at the state* s_{t-1}.

To solve MDP (Definition 1) we need to find an optimal policy maximising the sum of expected rewards (1).

Definition 3. *The optimal solution to MDP is a policy* π_t^{opt} *that maximises the expected accumulated reward (1),* $\pi_t^{opt} = \{p_\tau^{opt}(a_\tau|s_{\tau-1})\}_{\tau=1}^{t} \subset \boldsymbol{\pi}$.

$$\max_{\{p_t(a_t|s_{t-1})\}_{t=1}^{N}} \sum_{t\in\boldsymbol{T}} E_t[r_t(s_t, s_{t-1}, a_t)|s_{t-1}] = \sum_{t\in\boldsymbol{T}} \sum_{a_t\in\boldsymbol{A}} \sum_{s_t,s_{t-1}\in\boldsymbol{S}} r_t(s_t, s_{t-1}, a_t) p_t(s_t|s_{t-1}, a_t) p_t^{opt}(a_t|s_{t-1}) p_t(s_{t-1}) \quad (2)$$

2.3 Deliberation-Aware Multi-Proposer Ultimatum Game

Compare to general formulation of UG [2], the considered *multi-proposer N-round UG* scenario assumes $n_P \in \mathbb{N}$ proposers and one responder. The goal is the same as in traditional UG, i.e. to maximise a total profit while sharing a fixed amount of money q. The main difference is that at the beginning of each round the responder chooses a proposer to play with. For choosing different proposer than that in the previous round, the responder is penalised by a so-called *deliberation penalty* $d \in \mathbb{N}$. Then, similarly to [2] the selected proposer offers a split $o_t \in \{1, 2, \ldots, q-1\}$ for the responder and $(q - o_t)$ for herself. If the responder accepts the offer, money split according to the proposal, otherwise none of the players get anything. Proposers not selected in this round play passive role.

Let us define a multi-proposer N-round UG via MDP (see Section 2.2) with proposers representing *stochastic environment* and the responder acting as *agent*. All proposers are part of the environment and have their policies fixed.

Definition 4. *Multi-proposer UG in MDP framework over a set of decision epochs (*game rounds*)* $\boldsymbol{T}$ *is defined as in Definition 1 and*

- $s_t = (o_t, P_t, D_t, Z_{R,t}, Z_{P,t}^1, Z_{P,t}^2, \ldots, Z_{P,t}^{n_P})$ *is environment state at* $t \in \boldsymbol{T}$*, where*
  $o_t \in \mathbf{O}$ *is an offer*
  $P_t \in \{P^1, \ldots, P^{n_P}\}$ *is the proposer chosen in the round* $(t-1)$
  $D_t \in \mathbf{D}$ *is the deliberation accumulated up to round* t*,* $D_t = \sum_{\tau=1}^{t} d\chi(a_{1,\tau}, P_\tau)$
  $Z_{R,t}$ *and* $Z_{P,t}^i$ *is an accumulated economic profit of the responder and proposer* P^i*, respectively*
- $a_t = (a_{1,t}, a_{2,t})$ *is a two-dimensional action, where* $a_{1,t} \in \boldsymbol{A}_1 = \{1, 2, ..., n_P\}$ *denotes the selection of a proposer to play with;* $a_{2,t} \in \boldsymbol{A}_2 = \{1, 2\}$ *stands for the acceptance* $(a_{2,t} = 2)$ *or the rejection* $(a_{2,t} = 1)$ *of the offer* o_t*,* $\boldsymbol{A} = \boldsymbol{A}_1 \times \boldsymbol{A}_2$.
- *The transition probabilities* $p = p_t(s_t|s_{t-1}, a_1)$ *and the reward function* $r = r_t(s_t, s_{t-1}, a_t)$ *are assumed to be known.*

The responder's accumulated *economic profit*, $Z_{R,t} \in \mathbf{Z}_R$, at the round t is:

$$Z_{R,t} = \sum_{\tau=1}^{t} o_\tau(a_{2,\tau} - 1), \quad (3)$$

and accumulated *economic profit of the ith proposer*, $Z^i_{P,t} \in \mathbf{Z}^i_P$, equals

$$Z^i_{P,t} = \sum_{\tau=1}^{t} (q - o_\tau)(a_{2,\tau} - 1)\chi(a_{1,\tau}, i),\ \forall i = 1, 2, ..., n_P. \tag{4}$$

The action $a_{2,t}$, see Definition 4, considers dependence on offer $o_t \in \mathbf{O}$. However action $a_{1,t}$ is made without this knowledge, thus

$$p_t(a_t|o_t, s_{t-1}) = p_t(a_{1,t}, a_{2,t}|o_t, s_{t-1}) = p_t(a_{1,t}|s_{t-1})p_t(a_{2,t}|o_t, a_{1,t}, s_{t-1}). \tag{5}$$

Thus, the optimal policy for MDP, given by Definition 4, is searched among sequences of functions $\big(p_t(a_{1,t}|s_{t-1}), p_t(a_{2,t}|o_t, a_{1,t}, s_{t-1})\big)_{t=1}^{N}$.

3 Optimal Solution

Let the state be decomposed as follows

$$s_t = (o_t, \bar{s_t}) \text{ where } \bar{s_t} = (P_t, D_t, Z_{R,t}, Z^1_{P,t}, Z^2_{P,t}, ..., , Z^{n_P}_{P,t})\ \ \bar{s_t} \in \bar{\mathbf{S}}. \tag{6}$$

Using (6) and (5), the conditional expected reward can be expressed as:

$$E_t[r_t(\bar{s_t}, o_t, s_{t-1}, a_{1,t}, a_{2,t})|s_{t-1}] =$$
$$\sum_{a_{1,t}\in\mathbf{A}_1} \sum_{a_{2,t}\in\mathbf{A}_2} \sum_{o_t\in\mathbf{O}} \Big[\Big(\sum_{\bar{s_t}\in\bar{\mathbf{S}}} r_t(\bar{s_t}, o_t, s_{t-1}, a_{1,t}, a_{2,t}) p_t(\bar{s_t}|o_t, a_{1,t}, a_{2,t}, s_{t-1})\Big)$$
$$p_t(a_{2,t}|o_t, a_{1,t}, s_{t-1}) p_t(o_t|a_{1,t}, s_{t-1}) p_t(a_{1,t}|s_{t-1})\Big]. \tag{7}$$

Denoting the expression in round brackets in (7) by $\overline{r_t}(a_{2,t}, a_{1,t}, o_t, s_{t-1})$, the optimal decision rule $p^{opt}_t(a_{2,t}|o_t, a_{1,t}, s_{t-1})$ maximising (7) is given by

$$p^{opt}_t(a_{2,t}|o_t, a_{1,t}, s_{t-1}) = \chi(a_{2,t}, a^*_{2,t}(o_t, a_{1,t}, s_{t-1})), \text{ where} \tag{8}$$
$$a^*_{2,t}(o_t, a_{1,t}, s_{t-1}) \in \operatorname*{argmax}_{a_{2,t}\in\mathbf{A}_2} \overline{r_t}(a_{2,t}, a_{1,t}, o_t, s_{t-1})\ \forall (o_t, a_{1,t}) \in \mathbf{O} \times \mathbf{A}_1.$$

Now we have to maximize the remaining part of the expected reward (7):

$$\max_{p_t(a_{1,t}|s_{t-1})} \sum_{a_{1,t}\in\mathbf{A}_1} \Big[\Big(\sum_{a_{2,t}\in\mathbf{A}_2} \sum_{o_t\in\mathbf{O}} \overline{r_t}(a_{2,t}, a_{1,t}, o_t, s_{t-1}) p^{opt}_t(a_{2,t}|o_t, a_{1,t}, s_{t-1})$$
$$p_t(o_t|a_{1,t}, s_{t-1})\Big) p_t(a_{1,t}|s_{t-1})\Big] \tag{9}$$

Similarly to the above let us denote:

$$\overline{\overline{r_t}}(a_{1,t}, s_{t-1}) = \sum_{a_{2,t}\in\mathbf{A}_2} \sum_{o_t\in\mathbf{O}} \overline{r_t}(a_{2,t}, a_{1,t}, o_t, s_{t-1}) p^{opt}_t(a_{2,t}|o_t, a_{1,t}, s_{t-1}) p_t(o_t|a_{1,t}, s_{t-1}). \tag{10}$$

Then the optimal decision rule $p^{opt}_t(a_{1,t}|s_{t-1})$ is

$$p^{opt}_t(a_{1,t}|s_{t-1}) = \chi(a_{1,t}, a^*_{1,t}(s_{t-1})), \text{ where} \tag{11}$$
$$a^*_{1,t}(s_{t-1}) \in \operatorname*{argmax}_{a_{1,t}\in\mathbf{A}_1} \overline{\overline{r_t}}(a_{1,t}, s_{t-1}).$$

Theorem 1 (Optimal policy of the deliberation-aware responder). *A sequence of decision rules* $\left\{(p_t^{opt}(a_{1,t}|s_{t-1}), p_t^{opt}(a_{2,t}|o_t, a_{1,t}, s_{t-1}))\right\}_{t=1}^{N}$ *maximising the reward (1) forms an optimal policy and is computed via modification of dynamic programming [12] starting with* $\varphi_N(s_N) = 0$*, where*

$$\begin{aligned}
\varphi_{t-1}(s_{t-1}) &= E_t[(r_t(\bar{s_t}, o_t, s_{t-1}, a_{1,t}^*, a_{2,t}^*) + \varphi_t(s_t))|s_{t-1}, a_{1,t}^*, a_{2,t}^*] \\
a_{1,t}^*(s_{t-1}) &\in \underset{a_{1,t}\in \boldsymbol{A}_1}{\operatorname{argmax}}\, E_t\left[\overline{\overline{r_t}}(a_{1,t}, s_{t-1}) + \varphi_t(s_t) \mid s_{t-1}\right] \qquad (12) \\
a_{2,t}^*(o_t, a_{1,t}, s_{t-1}) &\in \underset{a_{2,t}\in \boldsymbol{A}_2}{\operatorname{argmax}}\, E_t\left[\overline{r_t}(a_{2,t}, a_{1,t}, o_t, s_{t-1}) + \varphi_t(s_t) \mid s_{t-1}, a_{1,t}^*\right]
\end{aligned}$$

Remark 1. Note that: i) the actions $a_{1,t}$, $a_{2,t}$ and the offer o_t do not depend on the previous offer o_{t-1} explicitly; ii) the action $a_{2,t}$ and the offer o_t do not depend on deliberation cost D_{t-1}; iii) the action $a_{2,t}$ does not depend on the economic gains of proposers.

4 Decision Making of Economic Responder

This paper considers purely self-interested type of responder (so called *economic responder*), which behaves in accordance with Game Theory and accepts all offers as anything is better than nothing. The motivation of the *economic responder* is pure economic profit, thus her reward function in the round t equals:

$$r_t(s_t, s_{t-1}, a_t) = (Z_{R,t} - Z_{R,t-1}) - (D_t - D_{t-1}). \qquad (13)$$

For simplicity of presentation let us assume that the transition probability functions of the proposers $p_t(o_t|Z_{R,t-1}, Z_{P,t-1}^{a_{1,t}}, a_{1,t})$, $\forall t \in \mathbf{T}$ are given.

The desired optimal strategy should maximize the expected reward (1) while respecting deliberation. Using (13) and Remark 1, the conditional expected reward of the economic responder reads:

$$\begin{aligned}
E_t[r_t(s_t, s_{t-1}, a_{1,t}, a_{2,t})|s_{t-1}] = \sum_{\substack{a_{1,t}\in\mathbf{A}_1 \\ a_{2,t}\in\mathbf{A}_2}} \sum_{o_t\in\mathbf{O}} \Big[&[o_t(a_{2,t} - 1) - d_t\chi(a_{1,t}, a_{1,t-1})] \\
&\times p_t(a_{2,t}|o_t, a_{1,t}, Z_{R,t-1}) p_t(o_t|Z_{R,t-1}, Z_{P,t-1}^{a_{1,t}}, a_{1,t}) \\
&\times p_t(a_{1,t}|Z_{R,t-1}, D_{t-1}, Z_{P,t-1}^1, ..., Z_{P,t-1}^{n_P})\Big]. \qquad (14)
\end{aligned}$$

With it, the optimal policy is given by Theorem 1.

5 Illustrative example

The example considered a N-round UG as described in Section 2, with $N = 30$, $q = 30$, deliberation penalty $d = 5$ and number of proposers $n_P = 3$. The transition probabilities of respective proposers were considered independent of the

economic profit. Before the simulation, the offers for all proposers were generated. The probabilities of the offers are drawn from Gaussian distribution with $\sigma = 2$ and mean equal to the pre-generated offer. Then four games were played. Classical UG with each proposer and deliberative multi-proposer N-round UG. In the 4th game the responder played according to the optimal strategy found in Section 3. We analysed the result of the simulation by comparing the Responder gain in each game. The results are summarised in Table 1 (Z_R - Responder's profit, D_R - Deliberation cost, Z_P^i - Economic gain of the i-th proposer, $\sum Z_P^i$ - The total gain of all proposers) and Figure 1.

Table 1: Data obtained from the simulation of four games

No of game	$Z_R - D_R$	D_R	Z_R	Z_P^1	Z_P^2	Z_P^3	$\sum Z_P^i$
1	515	0	515	385	0	0	385
2	458	0	458	0	442	0	442
3	494	0	494	0	0	406	406
4	628	40	668	110	38	84	232

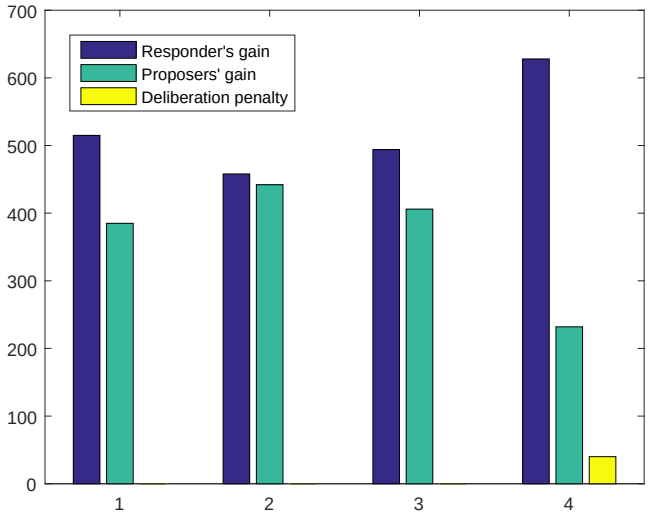

Figure 1: Responder's overall profit (lowered by the deliberation penalty), economic gain of all proposers and deliberation penalty for each of 4 games

6 Concluding Remarks

The paper examined the deliberation of the responder in multi-proposer Ultimatum Game. The responder behaviour was modelled by MDP and the deliberation cost was included into the responder's reward function and optimised so as economic profit. Comparison of the overall responder profit gained in the classical UG and in the deliberative multi-proposer UG was made. The results shown that though switching of proposers incurred non-negligible deliberation costs, the economic profit of the deliberation-aware responder was significantly higher in multi-proposer UG.

Many challenging aspects remain to be studied, in particular: i) modelling other types of responders considering not only pure economic profit, but non-economic aspects (fairness); ii) incorporating affective aspects of decision making, for example emotional state of the responder. Another direction is adding an adaptive feature, i.e. learning of the stochastic environment, i.e. learning the proposer model, see [13].

References

1. M. L. Puterman. *Markov Decision Processes*, Wiley, 1994.
2. A. Rubinstein *Perfect Equilibrum in a bargaining model, Econometrica 50(1), 97-109, 1982*, Wiley, 1994.
3. Sulkin, T. and A.F. Simon, Habermas in the lab: a study of deliberation in an experimental setting, *Political Psychology* 22: 809826, 2001.
4. M. Persson, P. Esaiasson, and M. Gilljam The effects of direct voting and deliberation on legitimacy beliefs: an experimental study of small group decision making *European Political Science Review*, 2006, pp. 1-19.
5. A. Beara, and D.G.Randa. Intuition, deliberation, and the evolution of cooperation. *PNAS,* 2016.
6. G.Loewenstein, and T. O'Donoghue Animal spirits: Affective and deliberative processes in economic behavior. Available at papers.ssrn.com, 2004.
7. O. Compte, P. Jehiel. Bargaining over Randomly Generated Offers: A new perspective on multi-party bargaining. *c.E.R.A.S.-E.N.P.C., C.N.R.S., France*, 2004.
8. D.A. Ortega, P.A. Braun Information, utility and bounded rationality, *Artificial General Intelligence*, LNCS, Volume 6830, Springer, 269-274, 2011.
9. E.J. Horvitz Reasoning about beliefs and actions under computational resource constraints, arXiv:1304.2759, 2013.
10. F. Dignum, V.Dignum, R.Prada, and Catholijn M.Jonker. A conceptual architecture for social deliberation in multi-agent organizations. *Multiagent and Grid Systems*, 11(3), 147-166, 2015
11. G. Gigerenzer. *Adaptive Thinking: Rationality in the Real World.* Oxford University Press, 2000.
12. R. Bellman. *Dynamic programming.* Princeton University Press, Princeton, 1957.
13. F. Hůla, M.Ruman, M.Kárný. Adaptive Proposer for Ultimatum Game. In *Proceedings of ICANN 2016.* 2016