Proceedings of Machine Learning Research 58 (2016) 51-60

Experimental Performance of Deliberation-Aware Responder in Multi-Proposer Ultimatum Game

Tatiana V. Guy GUY@UTIA.CAS.CZ
Marko Ruman MARKO.RUMAN@GMAIL.COM
František Hůla HULA.FRANTISEK@GMAIL.COM
Miroslav Kárný SCHOOL@UTIA.CAS.CZ
Department of Adaptive Systems
Institute of Information Theory and Automation
Czech Academy of Sciences
Prague, P.O.Box 18, 182 08, CZ

Editor: Tatiana V. Guy, Miroslav Kárný, David Rios-Insua, David H. Wolpert

Abstract

The ultimatum game serves for studying various aspects of decision making (DM). Recently, its multi-proposer version has been modified to study the influence of deliberation costs. An optimising policy of the responder, switching between several proposers at non-negligible deliberation costs, was designed and successfully tested in a simulated environment. The policy design was done within the framework of Markov Decision Processes with rewards also allowing to model the responder's feeling for fairness. It relies on simple Markov models of proposers, which are recursively learnt in a Bayesian way during the game course. This paper verifies, whether the gained theoretically plausible policy, suits to real-life DM. It describes experiments in which this policy was applied against human proposers. The results – with eleven groups of three independently acting proposers – confirm the soundness of this policy. It increases the responder's economic profit due to switching between proposers, in spite of the deliberation costs and the used approximate modelling of proposers. Methodologically, it opens the possibility to learn systematically willingness of humans to spent their deliberation resources on specific DM tasks.

Keywords: decision making; deliberation effort; Markov decision process; ultimatum game

1. Introduction

Maximizing of expected utility is perceived as "rational" within traditional economic models (von Neumann and Morgenstern (1944); Thaler (2000)). The observed discrepancies between theoretically optimal DM and real DM, e.g. (Gong et al. (2013), Jones (1999), Regenwetter et al. (2011)), can be diminished by changing the behaviour of DM subjects (by educating them) and by modifying decision rewards and models used in prescriptive theories. Our paper deals with the latter case and focuses on the influence of *deliberation effort* in DM.

A proper theory respecting deliberation effort should take into account that any decision made, either by humans or machines, costs time (Ortega and Stocker (2016)) energy and possibly other, limited, resources (Ortega et al. (2016)); a sample of different application

domains and related references are in Ruman et al. (2016). In this paper, we rely on, the design of DM policies respecting deliberation effort treated as an application of standard Markov Decision Process (MDP, Puterman (1994)) with the reward explicitly influenced by deliberation costs and with the environment model learnt in a Bayesian way (Peterka (1981)). The solution was developed for designing policy of the responder in multi-proposer ultimatum games (UG, Rubinstein (1994)). The simplicity of the UG enables extensive tests confronting prescriptive and human DM. In the UG, the proposer offers the split of a given amount of money and the responder either accepts this split, and the money are split, or refuses it, and none of players gets anything. In multi-proposer versions, the responder has the right to select the proposer among several of them. Any change of the proposer between rounds is penalized. In this way, the influence of deliberation effort is respected. Simulations documented in Ruman et al. (2016) confirmed the expected behaviour of the proposed responder's policy. However, a key question remained open: Will this policy be successful in real-life? It is a specific case of the generally inspected question: Does a prescriptive, theoretically justified, solution suit real life DM? Our experiments with human proposers, which form the core of this paper, provide answers to the posed questions.

The paper layout is as follows. Section 2 recalls formalization and the optimal design of the responder's policy in the multi-proposer UG. Section 3 describes the performed experiments and their results. Section 4 contains discussion. Section 5 provides concluding remarks.

2. Tested Decision-Making Policy

The section formalizes the multi-proposer UG and recalls the essence of the tested policy proposed in Ruman et al. (2016) as an application of MDP (Puterman (1994)).

2.1 Preliminaries

Throughout, bold capital $\mathbf{X}$ is a set of x-values; x_t is the value of x at the *decision epoch* $t \in \mathbf{T} = \{1, \ldots, N\}$ bounded by a horizon $N \in \mathbb{N}$ (here, the number of game rounds); $p(x|y)$ is a conditional probability. MDP provides a general framework for describing an *agent* (here, responder), which interacts with an *environment* (here, available proposers) by taking appropriate actions to achieve her goal. The decisions about *actions* $a_t \in \mathbf{A}$ (here, select the proposer to play with and accept or reject her offer) made are only influenced by the observed environment *state* $s_{t-1} \in \mathbf{S} \subset \mathbb{N}$, not by the whole environment history. The state s_{t-1} evolves to s_t according to the *transition probabilities*, $p = (p(s_t|a_t, s_{t-1}))_{t\in\mathbf{T}}$, influenced by actions $(a_t)_{t\in\mathbf{T}}$, and the agent receives rewards $r = (r(s_t, a_t, s_{t-1}))_{t\in\mathbf{T}}$. Given the initial state $s_0 = 0$, the tuple $(\mathbf{T}, \mathbf{S}, \mathbf{A}, r, p)$ describes MDP. The agent evaluates *randomized DM policies* $\pi = \left((p(a_t|s_{t-1}))_{a_t\in\mathbf{A}, s_{t-1}\in\mathbf{S}} \right)_{t\in\mathbf{T}}$ — formed by randomized *decision rules* $p(a_t|s_{t-1})$, $a_t \in \mathbf{A}$, $s_{t-1} \in \mathbf{S}$, $t \in \mathbf{T}$ — based on the *expected reward*

$$E_\pi[r(s_t, a_t, s_{t-1})] = \sum_{a_t\in\mathbf{A}} \sum_{s_t\in\mathbf{S}} \sum_{s_{t-1}\in\mathbf{S}} r(s_t, a_t, s_{t-1}) p(s_t|a_t, s_{t-1}) p(a_t|s_{t-1}) p(s_{t-1}), \quad (1)$$

where state probabilities $p(s_t)$ evolve according to

$$p(s_t) = \sum_{a_t\in\mathbf{A}} \sum_{s_{t-1}\in\mathbf{S}} p(s_t|a_t, s_{t-1}) p(a_t|s_{t-1}) p(s_{t-1})$$

with $p(s_0) = \delta(s_0, 0)$. The used Kronecker symbol $\delta(x, y) = 1$ if $x = y$, $\delta(x, y) = 0$ if $x \neq y$.

The agent seeks for the *optimal policy* π^{opt} maximizing the sum of expected rewards (1) up to the horizon N

$$\pi^{opt} \in \operatorname*{argmax}_{\pi} \sum_{t \in \mathbf{T}} E_{\pi}[r(s_t, a_t, s_{t-1})]. \tag{2}$$

2.2 Deliberation-Aware Multi-Proposer Ultimatum Game

The considered *multi-proposer N-round UG* assumes a *group* of $K \in \mathbb{N}$ proposers $\mathcal{P} \in \boldsymbol{\mathcal{P}} = \{\mathcal{P}^1, \ldots, \mathcal{P}^K\}$ and one responder $\mathcal{R}$. The responder's goal is the same as in the traditional UG, i.e. to influence her accumulated profit R_t, see (4), while accepting or rejecting the offers. The main difference is that at the beginning of each round $t \in \mathbf{T}$ the responder chooses a proposer $\mathcal{P}_t \in \boldsymbol{\mathcal{P}} = \{\mathcal{P}^1, \ldots, \mathcal{P}^K\}$ to play with. The choice of a proposer $\mathcal{P}_t$ at round $t \in \mathbf{T}$ is the first responder's action $a_{1t} = \mathcal{P}_t \in \boldsymbol{\mathcal{P}} = \mathbf{A}_1$. To model deliberation costs, the choice of a proposer different from that in the previous round is penalized by a *deliberation penalty* $d \in \mathbb{N}$. This leads to the *accumulated deliberation cost* of the responder

$$\mathsf{D}_t = d \sum_{\tau=1}^{t} (1 - \delta(\mathcal{P}_\tau, \mathcal{P}_{\tau-1})). \tag{3}$$

Then, as in the original UG (Rubinstein (1994)) the chosen proposer $\mathcal{P}_t$ offers a split $o_t \in \mathbf{O} = \{1, 2, \ldots, q-1\}$, $q \in \mathbb{N}$, for the responder and $(q - o_t)$ for herself. Money split according to the proposal if the responder *accepts the offer*, if she chooses the action $a_{2t} = 2$. None of the players get anything if the responder *rejects the offer*, if she chooses the action $a_{2t} = 1$. The accumulated responder's (economic) *profit*, R_t, at round $t \in \mathbf{T}$, is

$$\mathsf{R}_t = \sum_{\mathcal{P} \in \boldsymbol{\mathcal{P}}} \mathsf{P}_{\mathcal{P}t}, \qquad \mathsf{P}_{\mathcal{P}t} = \sum_{\tau=1}^{t} o_\tau (a_{2\tau} - 1) \delta(a_{1\tau}, \mathcal{P}), \qquad \mathcal{P} \in \boldsymbol{\mathcal{P}}. \tag{4}$$

Proposers play a passive role whenever they are not selected in the round. The accumulated proposers' (economic) profits are

$$Z_{\mathcal{P}t} = \sum_{\tau=1}^{t} (q - o_\tau)(a_{2\tau} - 1) \delta(a_{1\tau}, \mathcal{P}), \qquad \forall \mathcal{P} \in \boldsymbol{\mathcal{P}}. \tag{5}$$

In the multi-proposer UG, the responder R can be modelled as an agent in MDP, which tries to maximize her accumulated profit while minimising her accumulated deliberation cost.

Definition 1 *The multi-proposer UG in the MDP framework, with epochs $t \in \mathbf{T}$ identified with* game rounds*, is described through*

- *the environment state at $t \in \mathbf{T}$*

$$s_t = (o_t, \sigma_t) \;\; \textit{with } \sigma_t = (\mathcal{P}_t, \mathsf{D}_t, \mathsf{R}_t, Z_{\mathcal{P}^1 t}, Z_{\mathcal{P}^2 t}, ..., , Z_{\mathcal{P}^K t}), \quad \textit{where} \tag{6}$$

  $o_t \in \mathbf{O}$ is an offer made by the proposer $\mathcal{P}_t$. The accumulated deliberation cost D_t, the accumulated responder's (economic) profit R_t and the accumulated (economic) profits of proposers $Z_{\mathcal{P}t}$, $\mathcal{P} \in \boldsymbol{\mathcal{P}}$, are defined by (3), (4) and (5), respectively.

- *the two-dimensional action* $a = (a_1, a_2) \in \mathbf{A}_1 \times \mathbf{A}_2$ *consists of the selection* $a_1 \in \mathbf{A}_1 = \boldsymbol{\mathcal{P}}$ *of the proposer to play with and of* $a_2 \in \mathbf{A}_2 = \{1, 2\} = \{reject, accept\}$ *the offered split made by the selected proposer.*

  The selection of the proposer $a_{1t} = \mathcal{P}_t \in \boldsymbol{\mathcal{P}}$ *is based on the state* s_{t-1} *(6), while the action* $a_{2t} \in \mathbf{A}_2$ *also depends on the offer* $o_t \in \mathbf{O}$ *of the selected proposer* $\mathcal{P}_t \in \boldsymbol{\mathcal{P}}$*. Thus,*

$$p(a_t|o_t, s_{t-1}) = p(a_{1t}, a_{2t}|o_t, s_{t-1}) = p(a_{1t}|s_{t-1})p(a_{2t}|o_t, a_{1t}, s_{t-1}). \tag{7}$$

  Consequently, the optimal policy π^{opt} *is searched among sequences of functions*

$$\pi = \big(p(a_{1t}|s_{t-1}), p(a_{2t}|o_t, a_{1t}, s_{t-1})\big)_{t\in\mathbf{T}}. \tag{8}$$

- *The reward function with the penalty for the deliberation costs and respecting also self-fairness (Guy et al. (2015)) is considered*

$$r(s_t, a_t, s_{t-1}) = w(\mathsf{R}_t - \mathsf{R}_{t-1}) - (1-w)(Z_{\mathcal{P}_t t} - Z_{\mathcal{P}_t(t-1)}) - (\mathsf{D}_t - \mathsf{D}_{t-1}),\ w \in [0, 1]. \tag{9}$$

- *The transition probabilities* $p = p(s_t|a_{1t}, s_{t-1})$ *are assumed to be known, possibly as point estimates resulting from recursive estimation (Hůla et al. (2016)).*

For the inspection of the influence of deliberation costs, the risk neutral *economic responder*, caring about pure economic profit balanced with deliberation costs, is of interest. It is a special case of (9) with the weight $w = 1$. The results in Hůla et al. (2016), were adaptive proposer was studied, indicate that the economic player generates insufficiently exciting actions causing non-convergence of parameter estimates. In the cited case, the economic proposer generated very narrow range of offers and could not learn reactions of the responder to values out of this range. To avoid it, the self-fair modification of the responder's policy with weight $w \neq 1$ in (9), was used, but the results were judged according to the responder's profit.

2.3 Optimal Deliberation-Aware Responder Policy

Dynamic programming (Bellman (1957); Bertsekas (2001)) is used to solve (2). The special structure of policies (8) calls for a specific construction of the optimal policy. The following theorem — a tailored dynamic programming presented in Ruman et al. (2016) — provides the optimal strategy of the responder.

Theorem 2 (Optimal policy of the deliberation-aware responder) *The optimal policy* π^{opt} *constrained by (7) is formed by the sequence of decision rules*

$$\left\{(p^{opt}(a_{1t}|s_{t-1}), p^{opt}(a_{2t}|o_t, a_{1t}, s_{t-1}))\right\}_{t=1}^{N},$$

with $s_t = (o_t, \sigma_t)$ (6), which are evaluated against game course, starting with the value function $\varphi_N(s_N) = 0$, $\forall s_N \in \mathbf{S}$,

$$\begin{aligned}
\varphi_{t-1}(s_{t-1}) &= E[r(o_t, \sigma_t, a^\star_{1t}, a^\star_{2t}, s_{t-1}) + \varphi_t(s_t) | a^\star_{1t}, a^\star_{2t}, s_{t-1}] \\
a^\star_{1t}(s_{t-1}) &\in \operatorname*{argmax}_{a_{1t} \in \mathbf{A}_1} E\left[r(o_t, \sigma_t, a_{1t}, a_{2t}, s_{t-1}) + \varphi_t(s_t) \mid a_{1t}, s_{t-1}\right] \\
p^{opt}(a_{1t}|s_{t-1}) &= \delta(a_{1t}, a^\star_{1t}(s_{t-1})) \\
a^\star_{2t}(o_t, a^\star_{1t}, s_{t-1}) &\in \operatorname*{argmax}_{a_{2t} \in \mathbf{A}_2} E\left[r(o_t, \sigma_t, a^\star_{1t}, a_{2t}, s_{t-1}) + \varphi_t(s_t) \mid a^\star_{1t}, a_{2t}, o_t, s_{t-1}\right] \\
p^{opt}(a_{2t}|o_t, s_{t-1}) &= \delta(a_{2t}, a^\star_{2t}(o_t, a^*_{1t}, s_{t-1})). \qquad (10)
\end{aligned}$$

For the reward (9), the action $a^\star_{1t}$ in (10) describes the optimal, deliberation-aware, choice of the proposer and $a^\star_{2t}$ the optimal response to her offer.

3. Experiments

Successful simulation experiments of the optimal responder's policy π^{opt} (described by Theorem 2) were presented in Ruman et al. (2016). Its usefulness in real life has not been tested. This section presents results of experiments performed to fill this gap. The assumption that real proposers use time-invariant proposal policies describable by known or well-estimated probabilities $p = p(s_t|a_{1t}, s_{t-1})$ is the key potential weakness of the theoretically optimal responder's policy. Thus, the experiments mainly checked this assumption. In this experiment human-participants played roles of proposers against a virtual responder that was a computer programme implementing the deliberation-aware policy as described in Section 2.3.

3.1 Experimental Setup

Thirty three university students (mostly males, age range 19-25 years) participated in the study. The participants had no or minimal knowledge of the UG. The human-proposers could not interact/communicate during the game as the game ran through a web interface on personal computers. The keyboard was used as a response device. The participants were instructed about the UG rules and their role in the experiment. At each round a participants action was to propose an integer split of $q = 10\,CZK$. Participants were told to play trying to maximize their profit. The real money was at stake. The participants played with virtual money but they were paid their profits won in the game block (see below) at the end of experiment. The virtual responder always played with a group of three human-proposers $\mathcal{P} \in \boldsymbol{\mathcal{P}} = \{\mathcal{P}^1, \mathcal{P}^2, \mathcal{P}^3\}$. While playing, the virtual responder recursively learned the parameters $\alpha_\mathcal{P}, \beta_\mathcal{P} > 0,\ \alpha_\mathcal{P} + \beta_\mathcal{P} < 1$ of the simplified proposer's model

$$p(s_t|a_t, s_{t-1}, \alpha_\mathcal{P}, \beta_\mathcal{P}) = p(o_t|o_{t-1}, \alpha_\mathcal{P}, \beta_\mathcal{P}) = \begin{cases} \alpha_\mathcal{P} & \text{for offers } o_t > o_{t-1} \\ \beta_\mathcal{P} & \text{for offers } o_t < o_{t-1} \\ 1 - \alpha_\mathcal{P} - \beta_\mathcal{P} & \text{for offers } o_t = o_{t-1} \end{cases}. \qquad (11)$$

Learning essentially consists of evaluating the relative frequencies $\hat{\alpha}_\mathcal{P}, \hat{\beta}_\mathcal{P}$ corresponding to (11).

Interaction with a human group was split into *learning block* and *game block*. In the learning block, the virtual responder played $N = 20$ game rounds with each proposer $\mathcal{P} \in \boldsymbol{\mathcal{P}}$ (fixed $a_{1t} = \mathcal{P}$) while maximizing the expected accumulated reward with the reward function (9) and with the transition probabilities $p(s_t|a_t, s_{t-1}) = p(o_t|o_{t-1}, \hat{\alpha}_{\mathcal{P}_t}, \hat{\beta}_{\mathcal{P}_t})$. This made the virtual player adaptive. The learning block also helped the participants familiarize with their task.

In the game block, the virtual responder played $N = 20$ game rounds with all three human-proposers at once $\mathcal{P} \in \boldsymbol{\mathcal{P}}$ ($a_{1t} = \mathcal{P}_t$ was also optimized), while maximizing the expected accumulated reward with the reward function (9) and with the transition probabilities $p(s_t|a_t, s_{t-1}) = p(o_t|o_{t-1}, \hat{\alpha}_{\mathcal{P}_t}, \hat{\beta}_{\mathcal{P}_t})$, where the point estimates $(\hat{\alpha}_P, \hat{\beta}_P)$ of (α, β) were permanently updated. Thus, during each round the virtual responder had selected one of the human-proposers, who offered a split. Then the virtual responder decided on the acceptance/rejection of the split. The deliberation cost was set to $d = 1\,CZK$.

The results presented in Hůla et al. (2016) indicated that the economic player mostly generates insufficiently exciting actions, which result into divergence of parameter estimates. In the case of economic proposer, studied in Hůla et al. (2016), the player generated very narrow range of offers and could not learn reactions of the responder to values out of this range. To suppress this effect in our case, the self-fair modification of the virtual responder's policy with the weight $w = 0.6$ in (9) (Guy et al. (2015)) was used. The results were, however, judged according to the responder's profit.

3.2 Results

The achieved results are summarized in Table 1. It contains the *profits* $\mathsf{Pk} = \mathsf{P}_{\mathcal{P}^k N}$, (4), when playing with the proposer $\mathcal{P}^k$, $k = 1, 2, 3$, in the learning block. The average

$$\mathsf{RM} = \frac{1}{3}\sum_{k=1}^{3} \mathsf{Pk} \tag{12}$$

of these profits is comparable with the virtual *responder's profit* $\mathsf{R} = \mathsf{R}_N$ (4) in the game block. Switching between proposers is profitable if and only if the difference $\mathsf{R} - \mathsf{RM} > 0$. The results in the table are ordered according to this difference. Sample descriptive statistics are provided.

The ordered differences are also presented in Figure 1. Figure 2 presents two more detailed samples of the game results related to games with the worst and the best responder's results (1 and 11, Table 1).

4. Discussion

The experimental results confirm the potential applicability of the proposed responder's policy based on MDP in real life. Table 1 shows that the average and median of responder's profits increased by switching in spite of the relatively high switching cost, see boldface numbers. The increase realized in experiments with seven human groups from the performed eleven experiments, see Figure 1. Three experiments led to quite bad results and the difference of the average and median indicates that the distribution of differences $\mathsf{R} - \mathsf{RM}$ has heavy tail at negative profit differences. Thus, there is a space for improvement of the

Game	P1	P2	P3	RM	R	R − RM
1	70.00	25.00	117.00	70.67	51.00	-19.67
2	95.00	119.00	137.00	117.00	102.00	-15.00
3	163.00	130.00	121.00	138.00	131.00	-7.00
4	81.00	87.00	75.00	81.00	80.00	-1.00
5	96.00	80.00	106.00	94.00	96.00	2.00
6	140.00	94.00	60.00	98.00	101.00	3.00
7	79.00	69.00	58.00	68.67	77.00	8.33
8	76.00	88.00	91.00	85.00	94.00	9.00
9	63.00	76.00	83.00	74.00	86.00	12.00
10	75.00	93.00	76.00	81.33	95.00	13.67
11	64.00	103.00	120.00	95.67	110.00	14.33
mean	91.09	87.64	94.91	91.21	93.00	**1.79**
median	79.00	88.00	91.00	85.00	95.00	**3.00**
minimum	63.00	25.00	58.00	68.67	51.00	-19.67
maximum	163.00	130.00	137.00	138.00	131.00	14.33
standard deviation	32.10	27.47	26.80	20.96	20.32	11.52

Table 1: Pk is the virtual-responder's profit when playing with proposer $\mathcal{P}^k$ with no switching, RM is the average (12), $\mathsf{R} = \mathsf{R}_N$ (4) is the virtual responder's profit in the game block (with penalized switching).

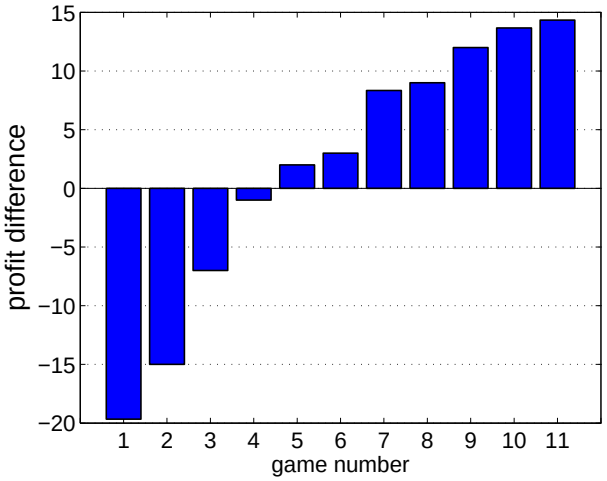

Figure 1: Profit differences in the game block and the average profit (12).

applied policy. Primarily, it has to improve the environment modelling, i.e. modelling of the proposers. Specifically:

- The assumed structure (11) should be refined by making probabilities of offers dependent on the values of differences $o_t - o_{t-1}$. A geometric change can still provide

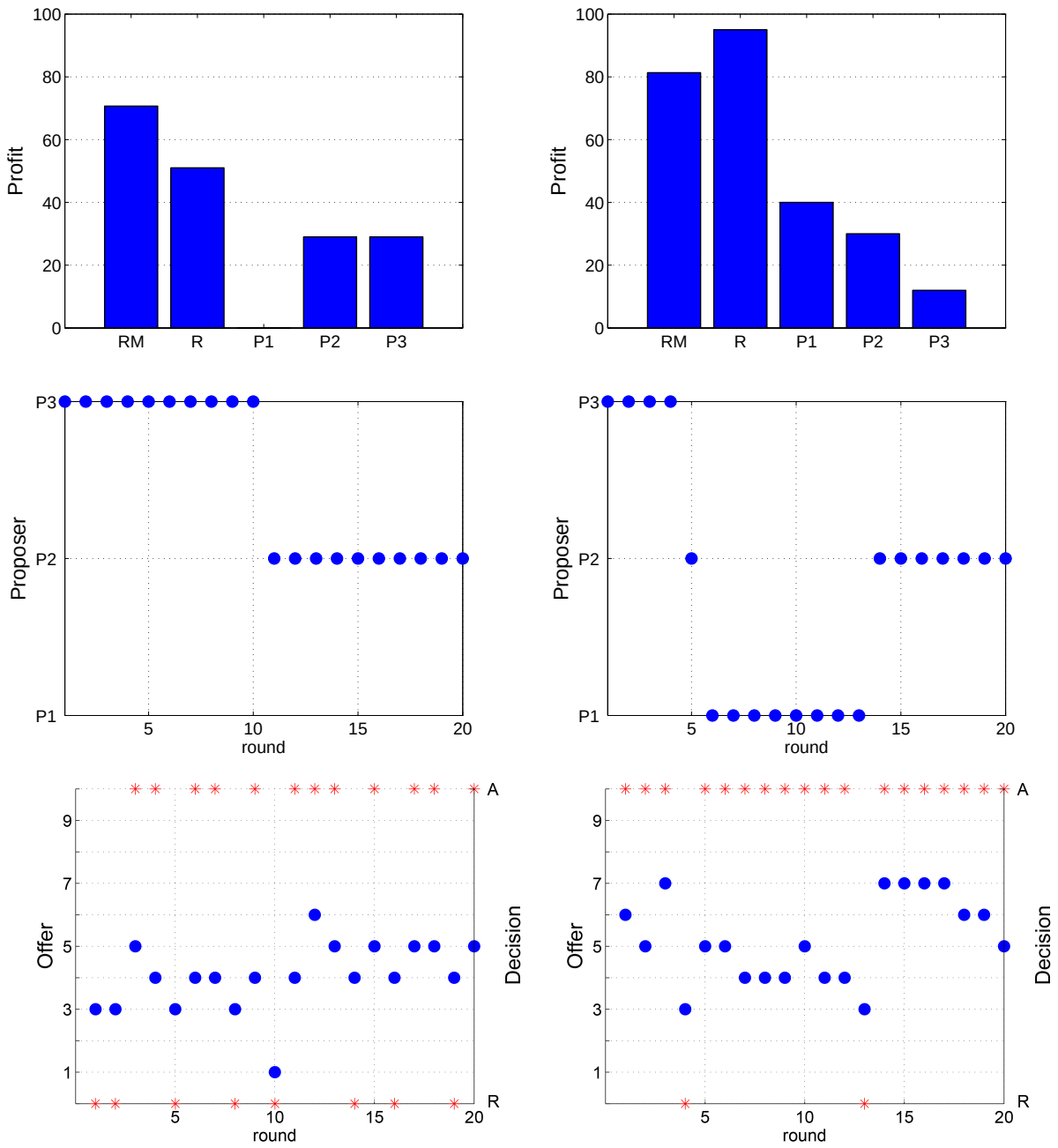

Figure 2: Left column concerns game 1, right column game 11 . First row shows the average profit RM (12), the virtual responder's profit $\mathsf{R} = \mathsf{R}_N$ (4) in the game block, and profits $\mathsf{Pk} = \mathsf{P}_{\mathcal{P}^k N}$ while playing with proposers $\mathcal{P}^k$ in the learning block. Second row shows switching between proposers $\mathcal{P}^k$ in the game block. Third row presents offers (•) and decisions Accept/Reject (∗) within the game course played in the game block.

parsimonious parametrization when considering constant (possibly asymmetric) decrease rates of probabilities $p(o_t|o_{t-1})$.

- Carefully selected prior probabilities of estimated parameters, based, for instance, on the already run games, may speed up estimation and increase the responder's profit. This is the main advantage of the adopted Bayesian estimation (Berger (1985); Garthwaite et al. (2005)), which is confirmed in the UG context in Hůla et al. (2016).
- The adopted exploitive strategy, definitely influenced the results as seen in Figure 2. The exploration-supporting "trick" based on optimization of non-economic profit reward (assuming *self-fair* virtual responder (Guy et al. (2015)) should be replaced by a more systematic treatment based on approximate dynamic programming (Si et al. (2004)) needed when dealing with learnt environment models (Feldbaum (1960)).

5. Concluding Remarks

This paper experimentally examined the influence of deliberation costs of the virtual responder in a multi-proposer Ultimatum Game with human proposers. It confirmed that the use of the MDP machinery while taking proposers as a part of environment is an efficient tool for solving game-like situations. This solution is close to the Bayesian games (Harsanyi (2004)). Our paper supports the claim that the approach caring about dynamics and incomplete knowledge of players makes the adopted theory applicable in real life. It offers extreme flexibility in modelling player's aims. In our case, it respects the influence of deliberation costs on decision making. This makes it not only a useful design tool but also an analytical tool. The analysis concerns a real decision maker, whose acting differs from "rational", purely economic, behaviour. We can analyse it as an inversion problem: assuming that she *is* rational but uses a different than economic reward, we can learn it from her actions. This idea was successfully used in learning of self-fairness in Guy et al. (2015). This paper opens the way of learning *laziness*, the *personal penalty of deliberation effort*. This is one direction of future work, which has to deal with others aspects like: i) extension of the presented experimental study to more groups of proposers with improved learning; ii) fighting with the curse of dimensionality inherent to the Bayesian games; iii) joint modelling of non-profit influences (deliberation costs, fairness, emotions, etc.).

Acknowledgments

The research reflected in this paper has been supported by GAČR GA16-09848S. Thorough feedbacks from anonymous reviewer help to improve the paper significantly.

References

R. Bellman. *Dynamic Programming.* Princeton University Press, New York, 1957.

J.O. Berger. *Statistical Decision Theory and Bayesian Analysis.* Springer, New York, 1985.

D.P. Bertsekas. *Dynamic Programming and Optimal Control.* Athena Scientific, US, 2001.

A.A. Feldbaum. Theory of dual control. *Autom. Remote Control*, 21(9), 1960.

A.A. Feldbaum. Theory of dual control. *Autom. Remote Control*, 22(2), 1961.

P.H. Garthwaite, J.B. Kadane, and A. O´Hagan. Statistical methods for eliciting probability distributions. *J. of the American Statistical Association*, 100(470):680–700, 2005.

J. Gong, Y. Zhang, Z. Yang, Y. Huang, J. Feng, and W. Zhang. The framing effect in medical decision-making: a review of the literature. *Psychology, Health and Medicine*, 18 (6):645–653, 2013.

T.V. Guy, M. Kárný, A. Lintas, and A.E.P. Villa. Theoretical models of decision-making in the ultimatum game: Fairness vs. reason. In R. Wang and Xiaochuan X. Pan, editors, *Advances in Cognitive Neurodynamics (V), Proceedings of the Fifth International Conference on Cognitive Neurodynamics*. Springer, 2015.

F. Hůla, M. Ruman, and M. Kárný. Adaptive proposer for ultimatum game. In *Artificial Neural Networks and Machine Learning – Proceedings ICANN 2016*, pages 330–338. Barcelona, 2016.

J.C. Harsanyi. Games with incomplete information played by Bayesian players, I–III. *Management Science*, 50(12), 2004. ISSN 0025-1909. Supplement.

B.D. Jones. Bounded rationality. *Annual Rev. Polit. Sci.*, 2:297–321, 1999.

P.A. Ortega and A.A. Stocker. Human decision-making under limited time. *arXiv preprint arXiv:1610.01698*, 2016.

P.A. Ortega, D.A. Braun, J. Dyer, K.E. Kim, and N. Tishby. Information-theoretic bounded rationality. *arXiv preprint arXiv:1512.06789v1*, 2016.

V. Peterka. Bayesian system identification. In P. Eykhoff, editor, *Trends and Progress in System Identification*, pages 239–304. Pergamon Press, Oxford, 1981.

M.L. Puterman. *Markov Decision Processes*. Wiley, 1994.

M. Regenwetter, J. Dana, and C.P. Davis-Stober. Transitivity of preferences. *Psychological Review*, 118(1):42–56, 2011.

A. Rubinstein. Perfect equilibrum in a bargaining model. *Econometrica*, 50:97–109, 1994. 1982.

M. Ruman, F. Hůla, M. Kárný, and T.V. Guy. Deliberation-aware responder in multi-proposer ultimatum game. In *Artificial Neural Networks and Machine Learning – Proceedings ICANN 2016*, pages 230–237. Barcelona, 2016.

J. Si, A.G. Barto, W.B. Powell, and D. Wunsch, editors. *Handbook of Learning and Approximate Dynamic Programming*, Danvers, May 2004. Wiley-IEEE Press. ISBN 0-471-66054-X.

R.H. Thaler. From homo economicus to homo sapiens. *Journal of Economic Perspectives*, 14:133–141, 2000.

J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, New York, London, 1944.