Neural Information Processing Systems Foundation

The 2nd International Workshop on

DECISION MAKING WITH MULTIPLE IMPERFECT DECISION MAKERS

held in conjunction with the 25th Annual Conference on Neural Information Processing Systems

December 16, 2011, Sierra Nevada, Spain

Printed in the Czech Republic

ISBN 978-80-903834-6-3

Neural Information
Processing Systems
Foundation

The 2nd International Workshop on

DECISION MAKING WITH MULTIPLE IMPERFECT DECISION MAKERS

held in conjunction with the 25th Annual Conference on Neural Information Processing Systems (NIPS 2011)

December 16, 2011

Sierra Nevada, Spain

http://www.utia.cz/NIPSHome

Organising and Programme Committee

Tatiana V Guy, Institute of Information Theory and Automation, Czech Republic

Miroslav Kárný, Institute of Information Theory and Automation, Czech Republic

David Rios Insua, Royal Academy of Sciences, Spain

Alessandro E.P. Villa, University of Lausanne, Switzerland

David Wolpert, Intelligent Systems Division, NASA Ames Research Center, USA

The partial support of Czech Science Foundation GA102/08/0567 is acknowledged.

Scope of the Workshop

Prescriptive Bayesian decision making supported by the efficient theoretically well-founded algorithms is known to be a powerful tool. However, its application within multiple-participants' settings needs an efficient support of *imperfect* participant (decision maker, agent), which is characterised by limited cognitive, acting and evaluative resources.

The interacting and multiple-task-solving participants prevail in the natural (societal, biological) systems and become more and more important in the artificial (engineering) systems. Knowledge of conditions and mechanisms influencing the participant's individual behaviour is a prerequisite to better understanding and rational improving of these systems. The diverse research communities permanently address these topics focusing either on theoretical aspects of the problem or (more often) on practical solution within a particular application. However, different terminology and methodologies used significantly impede further exploitation of any advances occurred. The workshop will bring the experts from different scientific communities to complement and generalise the knowledge gained relying on the multi-disciplinary wisdom. It extends the list of problems of the preceding NIPS workshop:

How should we formalise rational decision making of a single imperfect decision maker? Does the answer change for interacting imperfect decision makers? How can we create a feasible prescriptive theory for systems of imperfect decision makers?

The workshop especially welcomes contributions addressing the following questions:

What can we learn from natural, engineered, and social systems? How emotions influence decision making? How to present complex prescriptive outcomes to the human? Do common algorithms really support imperfect decision makers? What is the impact of imperfect designers of decision-making systems?

The workshop aims to brainstorm on promising research directions, present relevant case studies and theoretical results, and to encourage collaboration among researchers with complementary ideas and expertise. The workshop will be based on invited talks, contributed talks and posters. Extensive moderated and informal discussions ensure targeted exchange.

List of Invited Talks

- *Automated Preferences Elicitation*
  Miroslav Kárný, Tatiana V.Guy
- *Automated Explanations for MDP Policies*
  Omar Zia Khan, Pascal Poupart, James P. Black
- *Modeling Humans as Reinforcement Learners: How to Predict Human Behavior in Multi- Stage Games*
  Ritchie Lee, David H. Wolpert, Scott Backhaus, Russell Bent, James Bono, Brendan Tracey
- *Random Belief Learning*
  David Leslie
- *An Adversarial Risk Analysis Model for an Emotional Based Decision Agent*
  Javier G. Rázuri, Pablo G. Esteban, David Rios Insua
- *Bayesian Combination of Multiple, Imperfect Classifiers*
  Edwin Simpson, Stephen Roberts, Ioannis Psorakis, Arfon Smith, Chris Lintott
- *Emergence of Reverse Hierarchies in Sensing and Planning by Optimizing Predictive Information*
  Naftali Tishby
- *Effect of Emotion on the Imperfectness of Decision Making*
  Alessandro E.P. Villa, Marina Fiori, Sarah Mesrobian, Alessandra Lintas,
  Vladyslav Shaposhnyk, Pascal Missonnier

Time	**Title**	**Authors**
7:30—7:50	Opening session	Organisers
7:50—8:20	Emergence of Reverse Hierarchies in Sensing and Planning by Optimizing Predictive Information	Naftali Tishby
8:20—8:50	Modeling Humans as Reinforcement Learners: How to Predict Human Behavior in Multi-Stage Games	Ritchie Lee, David H. Wolpert, Scott Backhaus, Russell Bent, James Bono, Brendan Tracey
8:50—9:20	**Coffee break**	
9:20—9:50	Automated Explanations for MDP Policies	Omar Zia Khan, Pascal Poupart, James P. Black
9:50—10:20	Automated Preference Elicitation	Miroslav Kárný, Tatiana V.Guy
10:20—10:40	**Poster spotlights**	
10:40—11:40	**Posters and Demonstrations**	
	Artificial Intelligence Design for Real-time Strategy Games	Firas Safadi, Raphael Fonteneau, Damien Ernst
	Distributed Decision Making by Categorically-Thinking Agents	Joong Bum Rhim, Lav R. Varshney, Vivek K. Goyal
	Bayesian Combination of Multiple, Imperfect Classifiers	Edwin Simpson, Stephen Roberts, Arfon Smith, Chris Lintott
	Decision Making and Working Memory in Adolescents with ADHD after Cognitive Remediation	Michel Bader, Sarah Leopizzi, Eleonora Fornari, Olivier Halfon, Nouchine Hadjikhani
	Towards Distributed Bayesian Estimation: A Short Note on Selected Aspects	Kamil Dedecius, Vladimíra Sečkárová
	Variational Bayes in Fully Probabilistic Control for Multi-Agent Systems	Václav Šmídl, Ondřej Tichý
	Towards a Supra-Bayesian Approach to Merging of Information	Vladimíra Sečkárová
	Ideal and Non-Ideal Predictors in Estimation of Bellman Function	Jan Zeman
	DEMO: Interactive Two-Actors Game	Ritchie Lee
	DEMO: AIsoy Robots	David Rios Insua
11:45—4:00	**Break**	
4:00—4:30	Effect of Emotion on the Imperfectness of Decision Making	Alessandro E. P. Villa, Marina Fiori, Sarah Mesrobian, Alessandra Lintas, Vladyslav Shaposhnyk, Pascal Missonnier
4:30—5:00	Decision Support for a Social Emotional Robot	Javier G. Rázuri, Pablo G. Esteban, David Rios Insua,
5:00—6:00	**Posters & Demonstrations and Coffee Break**	
6:00—6:30	Random Belief Learning	David Leslie
6:30—7:00	Bayesian Combination of Multiple, Imperfect Classifiers	Edwin Simpson, Stephen Roberts, Ioannis Psorakis, Arfon Smith, Chris Lintott
7:00—8:00	Panel Discussion & Closing Remarks	Organisers

Emergence of reverse hierarchies in sensing and planning by optimizing predictive information

Naftali Tishby
Interdisciplinary Center for Neural Computation
The Ruth & Stan Flinkman family Chair in Brain Research
The Edmond and Lilly Safra Center for Brain Sciences, and
School of Computer Science and Engineering
The Hebrew University of Jerusalem
tishby@cs.huji.ac.il

Abstract

Efficient planning requires prediction of the future. Valuable predictions are based on information about the future that can only come from observations of past events. Complexity of planning thus depends on the information the past of an environment contains about its future, or on the "predictive information" of the environment. This quantity, introduced by Bilaek et. al., was shown to be subextensive in the past and future time windows, i.e.; to grow sub-linearly with the time intervals, unlike the full complexity (entropy) of events which grow linearly with time in stationary stochastic processes. This striking observation poses interesting bounds on the complexity of future plans, as well as on the required memories of past events. I will discuss some of the implications of this subextesivity of predictive information for decision making and perception in the context of pure information gathering (like gambling) and more general MDP and POMDP settings. Furthermore, I will argue that optimizing future value in stationary stochastic environments must lead to hierarchical structure of both perception and actions and to a possibly new and tractable way of formulating the POMDP problem.

Modeling Humans as Reinforcement Learners: How to Predict Human Behavior in Multi-Stage Games

Ritchie Lee
Carnegie Mellon University Silicon Valley
NASA Ames Research Park MS23-11
Moffett Field, CA 94035
ritchie.lee@sv.cmu.edu

David H. Wolpert
Intelligent Systems Division
NASA Ames Research Center MS269-1
Moffett Field, CA 94035
david.h.wolpert@nasa.gov

Scott Backhaus
Los Alamos National Laboratory
MS K764, Los Alamos, NM 87545
backhaus@lanl.gov

Russell Bent
Los Alamos National Laboratory
MS C933, Los Alamos, NM 87545
rbent@lanl.gov

James Bono
Department of Economics
American University
4400 Massachusetts Ave. NW
Washington DC 20016
bono@american.edu

Brendan Tracey
Department of Aeronautics and Astronautics
Stanford University
496 Lomita Mall, Stanford, CA 94305
btracey@stanford.edu

Abstract

This paper introduces a novel framework for modeling interacting humans in a multi-stage game environment by combining concepts from game theory and reinforcement learning. The proposed model has the following desirable characteristics: (1) Bounded rational players, (2) strategic (i.e., players account for one another's reward functions), and (3) is computationally feasible even on moderately large real-world systems. To do this we extend level-K reasoning to policy space to, for the first time, be able to handle multiple time steps. This allows us to decompose the problem into a series of smaller ones where we can apply standard reinforcement learning algorithms. We investigate these ideas in a cyber-battle scenario over a smart power grid and discuss the relationship between the behavior predicted by our model and what one might expect of real human defenders and attackers.

1 Introduction

We present a model of interacting human beings that advances the literature by combining concepts from game theory and computer science in a novel way. In particular, we introduce the first time-extended level-K game theory model [1, 2]. This allows us to use reinforcement learning (RL) algorithms to learn each player's optimal policy against the level $K-1$ policies of the other players. However, rather than formulating policies as mappings from belief states to actions, as in partially observable Markov decision processes (POMDPs), we formulate policies more generally as mappings from a player's observations and memory to actions. Here, memory refers to all of a player's past observations.

This model is the first to combine all of the following characteristics. First, players are strategic in the sense that their policy choices depend on the reward functions of the other players. This is in

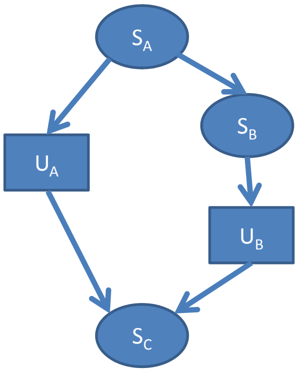


Figure 1: An example semi Bayes net.

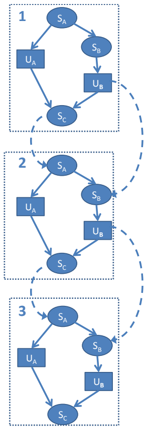


Figure 2: An example iterated semi Bayes net.

contrast to learning-in-games models whereby players do not use their opponents' reward information to predict their opponents' decisions and to choose their own actions. Second, this approach is computationally feasible even on real-world problems. This is in contrast to equilibrium models such as subgame perfect equilibrium and quantal response equilibrium [3]. This is also in contrast to POMDP models (e.g. I-POMDP) in which players are required to maintain a belief state over spaces that quickly explode. Third, with this general formulation of the policy mapping, it is straightforward to introduce experimentally motivated behavioral features such as noisy, sampled or bounded memory. Another source of realism is that, with the level-K model instead of an equilibrium model, we avoid the awkward assumption that players' predictions about each other are always correct.

We investigate all this for modeling a cyber-battle over a smart power grid. We discuss the relationship between the behavior predicted by our model and what one might expect of real human defenders and attackers.

2 Game Representation and Solution Concept

In this paper, the players will be interacting in an iterated semi net-form game. To explain an iterated semi net-form game, we will begin by describing a semi Bayes net. A semi Bayes net is a Bayes net with the conditional distributions of some nodes left unspecified. A pictoral example of a semi Bayes net is given in Figure 1. Like a standard Bayes net, a semi Bayes net consist of a set of nodes and directed edges. The ovular nodes labeled "S" have specified conditional distributions with the directed edges showing the dependencies among the nodes. Unlike a standard Bayes net, there are also rectangular nodes labeled "U" that have unspecified conditional dependencies. In this paper, the unspecified distributions will be set by the interacting human players. A semi net-form game, as described in [4], consists of a semi Bayes net plus a reward function mapping the outcome of the semi Bayes net to rewards for the players.

An iterated semi Bayes net is a Bayes net which has been time extended. It comprises of a semi Bayes net (such as the one in Figure 1), which is replicated T times. Figure 2 shows the semi Bayes net replicated three times. A set of directed edges L sets the dependencies between two successive iterations of the semi Bayes net. Each edge in L connects a node in stage $t-1$ with a node in stage t as is shown by the dashed edges in Figure 2. This set of L nodes is the same between every two successive stages. An iterated semi net-form game comprises of two parts: an iterated semi Bayes net and a set of reward functions which map the results of each step of the semi Bayes net into an incremental reward for each player. In Figure 2, the unspecified nodes have been labeled "U_A" and "U_B" to specify which player sets which nodes.

Having described above our model of the strategic scenario in the language of iterated semi net-form games, we now describe our solution concept. Our solution concept is a combination of the level-K model, described below, and reinforcement learning (RL). The level-K model is a game theoretic solution concept used to predict the outcome of human-human interactions. A number of studies [1, 2] have shown promising results predicting experimental data in games using this method. The solution to the level-K model is defined recursively as follows. A level K player plays as though all other players are playing at level $K-1$, who, in turn, play as though all other players are playing at level $K-2$, etc. The process continues until level 0 is reached, where the level 0 player plays according to a prespecified prior distribution. Notice that running this process for a player at $K \geq 2$ results in ricocheting between players. For example, if player A is a level 2 player, he plays as though player B is a level 1 player, who in turn plays as though player A is a level 0 player playing according to the prior distribution. Note that player B in this example may not actually be a level 1 player in reality – only that player A assumes him to be during his reasoning process.

This work extends the standard level-K model to time-extended strategic scenarios, such as iterated semi net-form games. In particular, each Undetermined node associated with player i in the iterated semi net-form game represents an action choice by player i at some time t. We model player i's action choices using the policy function, ρ_i, which takes an element of the Cartesian product of the spaces given by the parent nodes of i's Undetermined node to an action for player i. Note that this definition requires a special type of iterated semi-Bayes net in which the spaces of the parents of each of i's action nodes must be identical. This requirement ensures that the policy function is always well-defined and acts on the same domain at every step in the iterated semi net-form game. We calculate policies using reinforcement learning (RL) algorithms. That is, we first define a level 0 policy for each player, ρ_i^0. We then use RL to find player i's level 1 policy, ρ_i^1, given the level 0 policies of the other players, ρ_{-1}^0, and the iterated semi net-form game. We do this for each player i and each level K.[1]

3 Application: Cybersecurity of a Smart Power Network

In order to test our iterated semi net-form game modeling concept, we adopt a model for analyzing the behavior of intruders into cyber-physical systems. In particular, we consider Supervisory Control and Data Acquisition (SCADA) systems [5], which are used to monitor and control many types of critical infrastructure. A SCADA system consists of cyber-communication infrastructure that transmits data from and sends control commands to physical devices, e.g. circuit breakers in the electrical grid. SCADA systems are partly automated and partly human-operated. Increasing connection to other cyber systems creating vulnerabilities to SCADA cyber attackers [6].

Figure 3 shows a single, radial distribution circuit [7] from the transformer at a substation (node 1) serving two load nodes. Node 2 is an aggregate of small consumer loads distributed along the circuit, and node 3 is a relatively large distributed generator located near the end of the circuit. In this figure V_i, p_i, and q_i are the voltage, real power, and reactive power at node i. P_i, Q_i, r_i, and x_i are the real power, reactive power, resistance and reactance of circuit segment i. Together, these values represent the following physical system [7], where all terms are normalized by the nominal system voltage.

$$P_2 = -p_3, \;\; Q_2 = -q_3, \;\; P_1 = P_2 + p_2, \;\; Q_1 = Q_2 + q_2 \tag{1}$$

$$V_2 = V_1 - (r_1 P_1 + x_1 Q_1), \;\; V_3 = V_2 - (r_2 P_2 + x_2 Q_2) \tag{2}$$

In this model, r, x, and p_3 are static parameters, q_2 and p_2 are drawn from a random distribution at each step of the game, V_1 is the decision variable of the defender, q_3 is the decision variable of the attacker, and V_2 and V_3 are determined by the equations above. The injection of real power p_3 and reactive power q_3 can modify the P_i and Q_i causing the voltage V_2 to deviate from 1.0. Excessive deviation of V_2 or V_3 can damage customer equipment or even initiate a cascading failure beyond the circuit in question. In this example, the SCADA operator's (defender's) control over q_3 is compromised by an attacker who seeks to create deviations of V_2 causing damage to the system.

In this model, the defender has direct control over V_1 via a variable-tap transformer. The hardware of the transformer limits the defenders actions at time t to the following domain

$$D_{\mathcal{D}}(t) = \langle \min(v_{max}, V_{1,t-1} + v), V_{1,t-1}, \max(v_{min}, V_{1,t-1} - v) \rangle$$

[1] Although this work uses level-K and RL exclusively, we are by no means wedded to this solution concept. Previous work on semi net-form games used a method known as Level-K Best-of-M/M' instead of RL to determine actions. This was not used in this paper because the possible action space is so large.

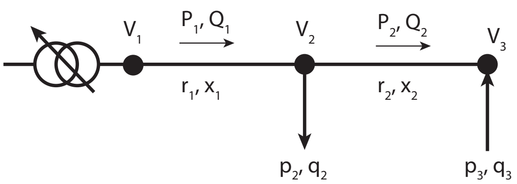


Figure 3: Schematic drawing of the three-node distribution circuit.

where v is the voltage step size for the transformer, and v_{min} and v_{max} represent the absolute min and max voltage the transformer can produce. Similarly, the attacker has taken control of q_3 and its actions are limited by its capacity to produce real power, $p_{3,max}$ as represented by the following domain.

$$D_{\mathcal{A}}(t) = \langle -p_{3,max}, \ldots, 0, \ldots, p_{3,max} \rangle$$

Via the SCADA system and the attacker's control of node 3, the observation spaces of the two players includes

$$\Omega_{\mathcal{D}} = \{V_1, V_2, V_3, P_1, Q_1, M_{\mathcal{D}}\}, \;\; \Omega_{\mathcal{A}} = \{V_2, V_3, p_3, q_3, M_{\mathcal{A}}\}$$

where $M_{\mathcal{D}}$ and $M_{\mathcal{A}}$ are used to denote each two real numbers that represent the respective player's memory of the past events in the game. Both the defender and attacker manipulate their controls in a way to increase their own rewards. The defender desires to maintain a high quality of service by maintaining the voltages V_2 and V_3 near the desired normalized voltage of one while the attacher wishes to damage equipment at node 2 by forcing V_2 beyond operating limits, i.e.

$$R_{\mathcal{D}} = -\left(\frac{V_2 - 1}{\epsilon}\right)^2 - \left(\frac{V_3 - 1}{\epsilon}\right)^2, \;\; R_{\mathcal{A}} = \Theta[V_2 - (1+\epsilon)] + \Theta[(1-\epsilon) - V_2]$$

Here, $\epsilon \sim 0.05$ for most distribution system under consideration, Θ is a Heaviside step function.

Level 1 defender policy The level 0 defender is modeled myopically and seeks to maximize his reward by following a policy that adjusts V_1 to move the average of V_2 and V_3 closer to one, i.e.

$$\pi_{\mathcal{D}}(V_2, V_3) = \arg\min_{V_1 \in D_{\mathcal{D}}(t)} \frac{(V_2 + V_3)}{2} - 1$$

Level 1 attacker policy The level 0 attacker adopts a *drift and strike* policy based on intimate knowledge of the system. If $V_2 < 1$, we propose that the attacker would decrease q_3 by lowering it by one step. This would cause Q_1 to increase and V_2 to fall even farther. This policy achieves success if the defender raises V_1 in order to keep V_2 and V_3 in the acceptable range. The attacker continues this strategy, pushing the defender towards v_{max} until he can quickly raise q_3 to push V_2 above $1 + \epsilon$. If the defender has neared v_{max}, then a number of time steps will be required to for the defender to bring V_2 back in range. More formally this policy can be expressed as

```
LEVEL0ATTACKER()
$V^* = \max_{q \in D_{\mathcal{A}}(t)} |V_2 - 1|$;
if $V^* > \theta_{\mathcal{A}}$
    then return $\arg\max_{q \in D_{\mathcal{A}}(t)} |V_2 - 1|$;
if $V_2 < 1$
    then return $q_{3,t-1} - 1$;
return $q_{3,t-1} + 1$;
```

where $\theta_{\mathcal{A}}$ is a threshold parameter.

3.1 Reinforcement Learning Implementation

Using defined level 0 policies as the starting point, we now bootstrap up to higher levels by training each level K policy against an opponent playing level $K-1$ policy. To find policies that maximize reward, we can apply any algorithm from the reinforcement learning literature. In this paper, we use an ϵ-greedy policy parameterization (with $\epsilon = 0.1$) and SARSA on-policy learning [8]. Training updates are performed epoch-wise to improve stability. Since the players' input spaces contain continuous variables, we use a neural-network to approximate the Q-function [9]. We improve performance by scheduling the exploration parameter ϵ in 3 segments during training: An ϵ of near unity, followed by a linearly decreasing segment, then finally the desired ϵ.

3.2 Results and Discussion

We present results of the defender and attacker's behavior at various level K. We note that our scenario always had an attacker present, so the defender is trained to combat the attacker and has no training concerning how to detect an attack or how to behave if no attacker is present. Notionally, this is also true for the attacker's training. However in real-life the attacker will likely know that there is someone trying to thwart this attack.

Level 0 defender vs. level 0 attacker The level 0 defender (see Figure 4(a)) tries to keep both V_2 and V_3 close to 1.0 to maximize his immediate reward. Because the defender makes steps in V_1 of 0.02, he does nothing for $30 < t < 60$ because any such move would not increase his reward. For $30 < t < 60$, the p_2, q_2 noise causes V_2 to fluctuate, and the attacker seems to randomly drift back and forth in response. At $t = 60$, the noise plus the attacker and defender actions breaks this "symmetry", and the attacker increases his q_3 output causing V_2 and V_3 to rise. The defender responds by decreasing V_1, indicated by the abrupt drops in V_2 and V_3 that break up the relatively smooth upward ramp. Near $t = 75$, the accumulated drift of the level 0 attacker plus the response of the level 0 defender pushes the system to the edge. The attacker sees that a strike would be successful (i.e., post-strike $V_2 < 1 - \theta_A$), and the level 0 defender policy fails badly. The resulting V_2 and V_3 are quite low, and the defender ramps V_1 back up to compensate. Post strike ($t > 75$), the attackers threshold criterion tells him that an immediate second strike would would not be successful, however, this shortcoming will be resolved via level 1 reinforcement learning. Overall, this is the behavior we have built into the level 0 players.

Level 1 defender vs. level 0 attacker During the level 1 training, the defender likely experiences the type of attack shown in Figure 4(a) and learns that keeping V_1 a step or two above 1.0 is a good way to keep the attacker from putting the system into a vulnerable state. As seen in Figure 4(b), the defender is never tricked into performing a sustained drift because the defender is willing to take a reduction to his reward by letting V_3 stay up near 1.05. For the most part, the level 1 defender's reinforcement learning effectively counters the level 0 attacker drift-and-strike policy.

Level 0 defender vs. level 1 attacker The level 1 attacker learning sessions correct a shortcoming in the level 0 attacker. After a strike ($V_2 < 0.95$ in Figure 4(a)), the level 0 attacker drifts up from his largest negative q_3 output. In Figure 4(c), the level 1 attacker anticipates that the increase in V_2 when he moves from $m = -5$ to $m = 5$ will cause the level 0 defender to drop V_1 on the next move. After this drop, the level 1 attacker also drops from $m = +5$ to -5. In essence, the level 1 attacker is leveraging the anticipated moves of the level 0 defender to create oscillatory strikes that push V_2 below $1 - \epsilon$ nearly every cycle.

Acknowledgments

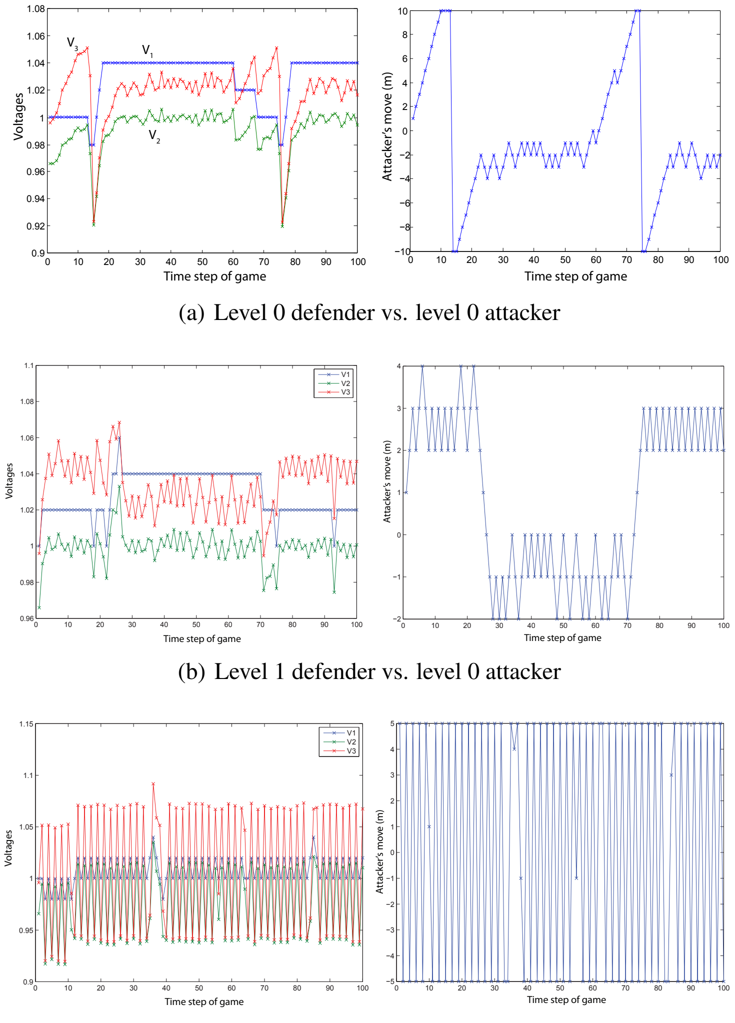
This research was supported by the NASA Aviation Safety Program SSAT project, and the Los Alamos National Laboratory LDRD project Optimization and Control Theory for Smart Grid.

References

[1] Miguel Costa-Gomes and Vincent Crawford. Cognition and behavior in two-person guessing games: An experimental study. *American Economic Review*, 96(5):1737–1768, December 2006.

[2] Dale O. Stahl and Paul W. Wilson. On players' models of other players: Theory and experimental evidence. *Games and Economic Behavior*, 10(1):218 – 254, 1995.

[3] Richard Mckelvey and Thomas Palfrey. Quantal response equilibria for extensive form games. *Experimental Economics*, 1:9–41, 1998. 10.1023/A:1009905800005.



(a) Level 0 defender vs. level 0 attacker

(b) Level 1 defender vs. level 0 attacker

(c) Level 0 defender vs. level 1 attacker

Figure 4: Voltages and attacker moves of various games.

[4] Ritchie Lee and David H. Wolpert. *Decision Making with Multiple Imperfect Decision Makers*, chapter Game Theoretic Modeling of Pilot Behavior during Mid-Air Encounters. Intelligent Systems Reference Library Series. Springer, 2011.

[5] K. Tomsovic, D.E. Bakken, V. Venkatasubramanian, and A. Bose. Designing the next generation of real-time control, communication, and computations for large power systems. *Proceedings of the IEEE*, 93(5):965 –979, may 2005.

[6] Alvaro A. Cárdenas, Saurabh Amin, and Shankar Sastry. Research challenges for the security of control systems. In *Proceedings of the 3rd conference on Hot topics in security*, pages 6:1–6:6, Berkeley, CA, USA, 2008. USENIX Association.

[7] K. Turitsyn, P. Sulc, S. Backhaus, and M. Chertkov. Options for control of reactive power by distributed photovoltaic generators. *Proceedings of the IEEE*, 99(6):1063 –1073, june 2011.

[8] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

[9] Lucian Busoniu, Robert Babuska, Bart De Schutter, and Ernst Damien. *Reinforcement Learning and Dynamic Programming Using Function Approximators*. CRC Press, 2010.

Automated Explanations for MDP Policies

Omar Zia Khan, Pascal Poupart and James P. Black
David R. Cheriton School of Computer Science
University of Waterloo
200 University Avenue West, Waterloo, ON, N2L 3G1, Canada
{ozkhan, ppoupart, jpblack}@cs.uwaterloo.ca

Abstract

Explaining policies of Markov Decision Processes (MDPs) is complicated due to their probabilistic and sequential nature. We present a technique to explain policies for factored MDP by populating a set of domain-independent templates. We also present a mechanism to determine a minimal set of templates that, viewed together, completely justify the policy. We demonstrate our technique using the problems of advising undergraduate students in their course selection and evaluate it through a user study.

1 Introduction

Sequential decision making is a notoriously difficult problem especially when there is uncertainty in the effects of the actions and the objectives are complex. MDPs [10] provide a principled approach for automated planning under uncertainty. State-of-the-art techniques provide scalable algorithms for MDPs [9], but the bottleneck is gaining user acceptance as it is harder to understand why certain actions are recommended. Explanations can enhance the user's understanding of these plans (when the policy is to be used by humans like in recommender systems) and help MDP designers to debug them (even when the policy is to be used by machines, like in robotics). Our explanations highlight key factors through a set of explanation templates. The set of templates are sufficient, such that they justify the recommended action, yet also minimal, such that the size of the set cannot be smaller. We demonstrate our technique through a course-advising MDP and evaluate our explanations through a user study. A more detailed description of our work can be found in [6].

2 Background

A Markov decision process (MDP) is defined by a set S of states s, a set A of actions a, a transition model (the probability $Pr\left(s'|s,a\right)$ of an action a in state s leading to state s'), a reward model (the utility/reward $R\left(s,a\right)$ for executing action a in state s), and a discount factor $\gamma \in [0,1)$. Factored MDPs [1] are typically used for MDPs with large state space where states are determined by values of some variables. A scenario sc is defined as the set of states obtained by assigning values to a subset of state variables. A policy $\pi : S \rightarrow A$ is a mapping from states to actions. The value $V^{\pi}\left(s\right)$ of a policy π when starting in state s is the sum of the expected discounted rewards earned by executing policy π. A policy can be evaluated by using Bellman's equation $V^{\pi}\left(s\right) = R\left(s,\pi\left(s\right)\right) + \gamma \sum_{s' \in S} Pr\left(s'|s,\pi\left(s\right)\right) \cdot V^{\pi}\left(s'\right)$. We shall use an alternative method to evaluate a policy based on occupancy frequencies. The discounted occupancy frequency (hereafter referred as occupancy frequency) $\lambda_{s_0}^{\pi}\left(s'\right)$ is the expected (discounted) number of times we reach state s' from starting state s_0 by executing policy π. Occupancy frequencies can be computed by solving Eq. 1.

$$\lambda_{s_0}^{\pi}\left(s'\right) = \delta\left(s', s_0\right) + \gamma \sum_{s \in S} Pr\left(s'|s,\pi\left(s\right)\right) \cdot \lambda_{s_0}^{\pi}\left(s\right) \quad \forall s' \tag{1}$$

where $\delta\left(s', s_0\right)$ is a Kroenecker delta which assigns 1 when $s' = s_0$ and 0 otherwise. The occupancy frequencies for a scenario (or a set of scenarios), $\lambda_{s_0}^{\pi}(sc)$, is the expected number of times we reach a scenario sc, from starting state s_0, by executing policy π *i.e.*, $\lambda_{s_0}^{\pi}(sc) = \sum_{s \in sc} \lambda_{s_0}^{\pi}(s)$. Let sc_r be a set of scenarios with reward value r. The dot product of occupancy frequencies and rewards gives the value of a policy, as shown in Eq. 2.

$$V^{\pi}\left(s_0\right) = \sum_r \lambda_{s_0}^{\pi}(sc_r) \cdot r \qquad (2)$$

An optimal policy π^* earns the highest value for all states (i.e., $V^{\pi *}(s) \geq V^{\pi}(s)\ \forall \pi, s$).

3 Explanations for MDPs

3.1 Templates for Explanations

Our explanation answers the question, "*Why has this action been recommended?*" by populating generic templates, at run-time, with domain-specific information from the MDP *i.e.*, occupancy frequency of a scenario. The reward function implicitly partitions the state space in regions with equal reward value. These regions can be defined as partial variable assignments corresponding to scenarios or sets of scenarios. An explanation then could be the frequency of reaching a scenario is highest (or lowest). This is especially useful when this scenario also has a relatively high (or low) reward. Below we describe templates in which the underlined phrases (scenarios and their frequencies) are populated at run-time.

- **Template 1:** "*ActionName* is the only action that is likely to take you to $\underline{Var_1 = Val_1, Var_2 = Val_2, ...}$ about $\underline{\lambda}$ times, which is higher (or lower) than any other action"
- **Template 2:** "*ActionName* is likely to take you to $\underline{Var_1 = Val_1, Var_2 = Val_2, ...}$ about $\underline{\lambda}$ times, which is as high (or low) as any other action"
- **Template 3:** "*ActionName* is likely to take you to $\underline{Var_1 = Val_1, Var_2 = Val_2, ...}$ about $\underline{\lambda}$ times"

While these templates provide a method to present explanations, multiple templates can be populated even for non-optimal actions; a non-optimal action can have the highest frequency of reaching a scenario without having the maximum expected utility. Thus, we need to identify a set of templates that justify the optimal action.

3.2 Minimal Sufficient Explanations

We define an explanation as sufficient if it can prove that the recommendation is optimal, *i.e.*, the selected templates show the action is optimal without needing additional templates. A sufficient explanation cannot be generated for a non-optimal action since an explanation for another action (*e.g.*, the optimal action) will have a higher utility. A sufficient explanation is also minimal if it includes the minimum number of templates needed to ensure it is sufficient. The minimality constraint is useful for users and sufficiency constraint is useful for designers.

Let s_0 be the state where we need to explain why $\pi^*\left(s_0\right)$ is an optimal action. We can compute the value of the optimal policy $V^{\pi^*}\left(s_0\right)$ or the Q-function[1] $Q^{\pi^*}\left(s_0, a\right)$ using Eq. 2. Since a template is populated by a frequency and a scenario, the utility of this pair in a template is $\lambda_{s_0}^{\pi^*}\left(sc_r\right) \cdot r$. Let E be the set of frequency-scenario pairs that appear in an explanation. If we exclude a pair from the explanation, the utility is $\lambda_{s_0}^{\pi^*}\left(sc_i\right) \cdot \bar{r}$, where r_{min} is the minimum value for the reward variable. This definition indicates that the worst is assumed for the scenario in this pair. The utility of an explanation V_E is

[1] In reinforcement learning, the Q-function $Q^{\pi}(s, a)$ denotes the value of executing action a in state s followed by policy π.

$$V_E = \sum_{i \in E} \lambda_{s_0}^{\pi^*}(sc_i) \cdot r_i + \sum_{j \notin E} \lambda_{s_0}^{\pi^*}(sc_j) \cdot r_{min} \tag{3}$$

where the first part includes the utility from all the pairs in the explanation and the second part considers the worst case for all other pairs. For an explanation to be sufficient, its utility has to be higher than the next best action, *i.e.*, $V^{\pi^*} \geq V_E > Q^{\pi^*}(s_0, a) \quad \forall a \neq \pi^*(s_0)$. For it to be minimal, it should use the fewest possible pairs. Let us define the gain of including a pair in an explanation as the difference between the utility of including versus excluding that pair ($\lambda_{s_0}^{\pi^*}(sc_i) \cdot r_i - \lambda_{s_0}^{\pi^*}(sc_i) \cdot r_{min}$). To find a minimal sufficient explanation, we can sort the gains of all pairs in descending order and select the first k pairs that ensure $V_E \geq Q^{\pi^*}(s_0, a)$. This provides our minimal sufficient explanation.

3.3 Workflow and Algorithm

The designer identifies the states and actions, and specifies the transition and reward functions. The optimal policy is computed, using a technique such as value iteration, and is consulted to determine the optimal action. Now an explanation can be requested. The pseudo code for the algorithm to compute a minimal sufficient explanation is shown in Algorithm 1.

Algorithm 1 Computing Minimal Sufficient Explanations

```
//Inputs:  Starting State: $s_0$, Optimal Policy: $\pi$*
//Outputs: Minimal Sufficient Explanation, MSE
ComputeMSE($s_0$,$\pi$*)
 for r in R
 |   sc[r] = ComputeScenarios(r)
 |   $\lambda$[r] = ComputeOccupancyFrequency(sc[r])
 |   utilTemplate [sc,$\pi$*,$s_0$]= $\lambda$[r]*r
 |   utilNoTemplate[sc,$\pi^a$,$s_0$]=$\lambda$[r]*$r_{min}$
 |   netUtil[sc,r] = utilTemplate[sc,$\pi$*,$s_0$] - utilNoTemplate[sc,$\pi^a$,$s_0$]
 end
 sortedNet = sortDescending(netUtil)
 $V_E$ = 0; $V_{templates}$ = 0; $V_{noTemplates}$ = sum(utilNoTemplate);
 k=1; MSE, Pairs={}
 do
 |   add sc[r], $\lambda$[r] for sortedNet[k] in Pair
 |   $V_{templates}$  +=  utilTemplate[k]
 |   $V_{noTemplates}$ -=  sortedNet[k]
 |   $V_E$ = $V_{templates}$ + $V_{noTemplates}$
 |   k++
 while ($V_E$ < $V_{next}$)
 MSE = generateTemplates(Terms)
return MSE
//end computeMSE
```

The function `ComputeScenarios` returns the set of scenarios with reward value r which is available in the encoding of the reward function. The function `ComputeOccupancyFrequency` is the most expensive step which corresponds to solging the system of linear system defined in Eq. 1, which has a worst case complexity that is cubic in the size of the state space. However, in practice, the running time can often be sublinear by using variable elimination to exploit conditional independence and algebraic decision diagrams [5] to automatically aggregate states with identical values/frequencies. The function `GenerateTemplates` chooses an applicable template, from the list of templates, in the order of the list, with the last always applicable.

4 Experiments and Evaluation

4.1 Sample Explanations

We ran experiments on course-advising and hand-washing MDPs [6]. We only discuss the course-advising domain here due to space considerations. The transition model was obtained by using historical data collected over several years at the University of Waterloo. The reward function provides rewards for completing different degree requirements. The horizon of this problem is 3 steps, each step representing one term and the policy emits a pair of courses to take in that term. The problem has 117.4 million states. We precomputed the optimal policy since it does not need to be recomputed for every explanation. We were able to compute explanations in approximately 1 second on a Pentium IV 1.66 GHz laptop with 1GB RAM using Java on Windows XP with the optimal policy and second best action precomputed. A sample explanation is shown below.

- Action *TakeCS343&CS448* is the best action because:-
  - It's likely to take you to $\underline{CoursesCompleted = 6}$, $\underline{TermNumber = Final}$ about $\underline{0.86}$ times, which is as high as any other action

4.2 User Study with Students

We conducted a user study to evaluate explanations for course advising. We recruited 37 students and showed 3 different recommendations with explanations for different states. For each explanation, they were asked to rate it on various factors such as comprehension, trust-worthiness and usefulness with partial results shown in Figure 1. 59% (65/111) of the respondents indicated that they were able to understand our explanation without any other information; the rest also wanted to know the occupancy frequencies for some other actions. We can provide this information as it is already computed. 76% (84/111) believed that the explanation provided by our system was accurate, with a few wanting to know our sample size to judge the accuracy. 69% (77/111) indicated that they would require extra information beyond that presented in the explanation. When asked what other type of information is needed, we discovered that they wanted the model to cater to preferences such as student's interest, future career plans, and level of difficulty rather than the explanation being inadequate for our existing model. An important indicator of the usefulness of these explanations is that 71% (79/111) of the students mentioned that the explanation provided them with extra information that helped them in making a decision. Also while some students, 23% (26/111), initially disagreed with the recommendation, in 35% (9/26) of these cases our explanation convinced them to change their mind and agree with the original recommendation. The rest disagreed primarily because they wanted a more elaborate model, so no explanation could have convinced them.

We also asked students if they were provided with our system, in addition to the option of discussing their choices with an undergraduate advisor, would they use it. 86% of them mentioned they would use it from home and 89% mentioned they would use it before meeting with an advisor to examine different options for themselves. These numbers indicate substantial interest in our explanations. The explanations generated by our system are generic, while those provided by the advisors are domain-specific. The user study indicates that these two types of explanations are complementary and students would like to access our explanations in addition to consulting advisors.

5 Relationship to Other Explanations Strategies

Explanations have been considered an essential component of intelligent reasoning systems and various strategies have been devised to generate them. Explanations for expert systems are generally in the form of execution traces, such as in MYCIN [2]. Execution traces indicate the rules used in arriving at a conclusion. There are no specific rules in an MDP and the optimal decision is made by maximizing the expected utility which involves considering all of the transition and reward function. Thus, in our explanation we highlight the more important parts of the transition and reward function. Xplain [12] is an example of an intelligent tutoring system that also provided justifications of its decisions. In addition to the rules used by the expert system, it also needed additional domain knowledge to generate these explanations. Our current approach does not use any additional domain knowledge, however this also means we cannot justify the correctness of the transition or reward function. We can only argue about the optimal action using the specified transition and reward functions. Explanations in single-shot recommender systems [13] and case-based reasoning systems [11] are typically based on identifying similar clusters of users or cases and then demonstrating the similarity of the current choice to a cluster or case. Since MDPs are not based on the principle of recommending actions based on similarity, such an approach to generate explanations would be infeasible. Herlocker et al. [4] presented the idea of highlighting key data leading to a recommendation for explanations in recommender systems. Our approach is also motivated by this idea with the key difference that choices in MDPs also impact future states and actions rather than

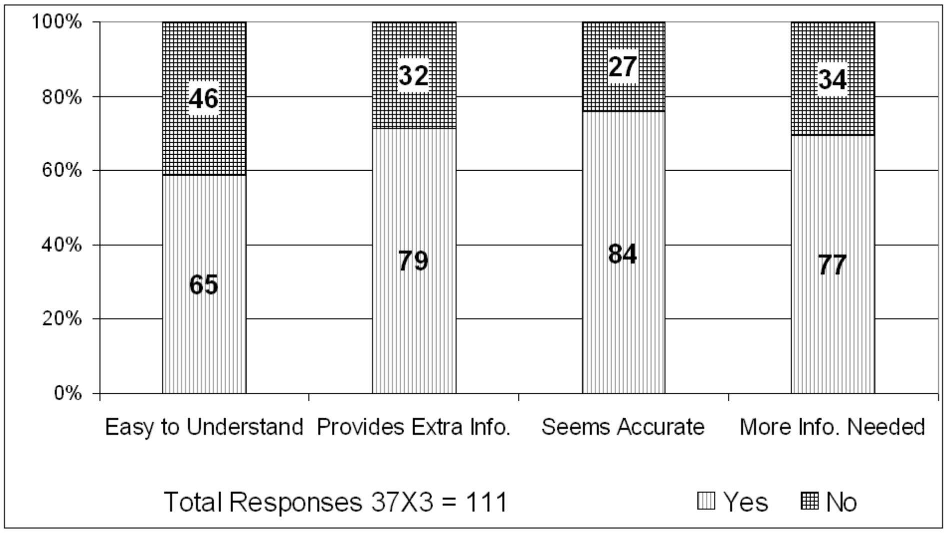


Figure 1: User Perception of MDP-Based Explanations

explaining an isolated decision. McGuinness et al. [8] identify several templates to present explanations in task processing systems based on predefined workflows. Our approach also uses templates, but we cannot use predefined workflows due to the probabilistic nature of MDPs.

Lacave et al. [7] presented several approaches to explain graphical models, including Bayesian networks and influence diagrams. Their explanations require a background in decision analysis and they present utilities of different actions graphically and numerically. We focus on users without any knowledge of utility theory. Elizalde et al. [3] present an approach to generate explanations for an MDP policy that recommends actions for an operator in training. A set of explanations is defined manually by an expert and their algorithm determines a relevant variable to be presented as explanation. Our approach does not restrict to a single relevant variable and considers the long-term effects of the optimal action (beyond one time step). We also use generic, domain-independent templates and provide a technique to determine a minimum set of templates that can completely justify an action.

6 Significance and Implications

While there has been a lot of work on explanations for intelligent systems, such as expert, rule-based, and case-based reasoning systems, there has not been much work for probabilistic and decision-theoretic systems. The main reason behind this discrepancy is the difference in processes through which they arrive at their conclusions. For probabilistic and decision-theoretic systems, there are well-known axioms of probability and theorems from utility theory that are applied to perform inference or compute a policy. Therefore, experts do not need to examine the reasoning trace to determine if the inference or policy computation process is correct. The trace would essentially refer to concepts such as Bayes' theorem, or the principle of maximum expected utility etc, which do not need to be verified. Instead, the input, *i.e.*, transition and reward function, need to be verified. With recent advances in scalability and the subsequent application of MDPs to real-world problems, now explanation capabilities are needed. The explanation should highlight portions of the input that lead to a particular result.

Real-world MDPs are difficult to design because they can involve millions of states. There are no existing tools for experts to examine and/or debug their models. The current design process involves successive iterations of tweaking various parameters to achieve a desirable output. At the end, the experts still cannot verify if the policy indeed accurately reflects their requirements. Our explanations provide hints to experts in debugging by indicating the components of the model that are being utilized in the decision-making process at the current step. This allows experts to verify whether the correct components are being used and focus the tweaking of the model.

Current users have to trust an MDP policy blindly, with no explanations whatsoever regarding the process of computing the recommendation or the confidence of the system in this recommendation. They cannot observe which factors have been considered by the system while making the recommendation. Our explanations can provide users with the information that the MDP is using to base its recommendation. This is especially important if user preferences are not accurately encoded.

If experts or users disagree with the optimal policy, the next step would be to automatically update the model based on interaction, *i.e.*, update the transition and reward functions if the user/expert disagree with the optimal policy despite the explanation. Any such automatic update of the model needs to be preceded by a proper understanding of the existing model, which can only be achieved through explanations, such as those provided by our system.

Just like the optimal policies for MDPs from different domains can be computed using the same underlying techniques, our technique to generate explanations is also generic and can be employed for an MDP from any domain. We have used the same approach described here to generate minimal sufficient explanations for the handwashing MDP [6]. The mechanism to present the explanation to users can then be tailored for various domains. Often a fancier graphical presentation may be more useful than a text-based template. Our focus is to produce generic explanations that can then be transformed for presentation in a user-friendly format.

7 Conclusion

We presented a mechanism to generate explanations for factored MDP in any domain without requiring any additional effort from the MDP designer. We introduced the concept of a minimal sufficient explanation through which an action can be explained using the fewest possible templates. We showed that our explanations can be generated in near-real time and conducted a user study to evaluate their effectiveness. The students appreciated the extra information provided by our generic explanations. Most of the students considered the combination of our explanation with the advisor explanation more effective than either one alone.

In the future, it would be interesting to extend this work to partially observable MDPs. Since the states are not directly observable, it is not obvious how one could generate an explanation that refers to the frequency with which some states are visited. It would also be interesting to extend this work to reinforcement learning problems where the parameters of the model (i.e., transition probabilities and reward function) are unknown or at best partially known. Finally, when an explanation is provided and the user insists that the recommended action is suboptimal, then it would be interesting to close the loop by updating the model to take into account the feedback provided by the user.

References

[1] Craig Boutilier, Richard Dearden, and Moisés Goldszmidt. Stochastic dynamic programming with factored representations. *Artificial Intelligence*, 121(1-2):49–107, 2000.

[2] W. J. Clancey. The epistemology of a rule-based expert system – a framework for explanation. *Artificial Intelligence*, 20:215–251, 1983.

[3] F. Elizalde, E. Sucar, A. Reyes, and P. deBuen. An MDP approach for explanation generation. In *AAAI Workshop on Explanation-Aware Computing*, 2007.

[4] J. Herlocker. Explanations in recommender systems. In *CHI' 99 Workshop on Interacting with Recommender Systems*, 1999.

[5] Jesse Hoey, Robert St-aubin, Alan Hu, and Craig Boutilier. SPUDD: Stochastic planning using decision diagrams. In *UAI*, pages 279–288, Stockholm, Sweden, 1999.

[6] Omar Zia Khan, Pascal Poupart, and James P. Black. Minimal sufficient explanations for factored markov decision processes. In *ICAPS*, Thessaloniki, Greece, 2009.

[7] C. Lacave, M. Luque, and F.J. Dez. Explanation of Bayesian networks and influence diagrams in Elvira. *IEEE Transactions on Systems, Man, and Cybernetics*, 37(4):952–965, 2007.

[8] D. McGuinness, A. Glass, M. Wolverton, and P. da Silva. Explaining task processing in cognitive assistants that learn. In *Proceedings of AAAI Spring Symposium on Interaction Challenges for Intelligent Assistants*, 2007.

[9] Warren B. Powell. *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. Wiley, 2nd edition, 2011.

[10] Martin Puterman. *Markov Decision Processes*. Wiley, 1994.

[11] Frode Sørmo, Jörg Cassens, and Agnar Aamodt. Explanation in case-based reasoning—perspectives and goals. *Artificial Intelligence Review*, 24(2):109–143, 2005.

[12] W. R. Swartout. Xplain: A system for creating and explaining expert consulting programs. *Artificial Intelligence*, 21:285–325, 1983.

[13] N. Tintarev and J. Masthoff. A survey of explanations in recommender systems. In *ICDE Workshop on Recommender Systems & Intelligent User Interfaces*, 2007.

Automated Preferences Elicitation

Miroslav Kárný
Department of Adaptive Systems
Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
Pod vodárenskou věží 4, 182 08 Prague 8, Czech Republic
school@utia.cas.cz

Tatiana V. Guy
Department of Adaptive Systems
Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
Pod vodárenskou věží 4, 182 08 Prague 8, Czech Republic
guy@utia.cas.cz

Abstract

Systems supporting decision making became almost inevitable in the modern complex world. Their efficiency depends on the sophisticated interfaces enabling a user take advantage of the support while respecting the increasing on-line information and incomplete, dynamically changing user's preferences. The best decision making support is useless without the proper preference elicitation. The paper proposes a methodology supporting automatic learning of quantitative description of preferences. The proposed elicitation serves to fully probabilistic design, which is an extension of Bayesian decision making.

1 Introduction

A feasible and effective solution of preference elicitation problem decides on the efficiency of any intelligent system supporting decision making. Indeed, to recommend a participant[1] an optimal sequence of optimal decisions requires knowing some information about what the participant (affected by the recommended decision, if accepted) considers as "optimal". Extracting the information about the participant's preferences or utility is known as *preference elicitation* or utility elicitation[2]. This vital problem has been repeatedly addressed within artificial intelligence, game theory, operation research and many sophisticated approaches have been proposed [7], [8], [6], [5]. A number of approaches has arisen in connection with applied sciences like economy, social science, clinical decision making, transportation, see, for instance, [18], [9]. To ensure feasibility and practical applicability, many decision support systems have been designed under various assumptions on the structure of preferences. In particular, a broadly accepted additive independence [16] of values of individual attributes is not generally valid. In many applications the preferences of attributes are mostly dependent and the assumption above significantly worsens the elicitation results[3].

To benefit from any decision support, the preferences should be known in the form allowing their processing by an intended decision support system. Unless the participant's preferences are completely provided by the participant, they should be learned from either *past data* or *domain-specific*

[1] Participant is also known as user, decision maker, agent.

[2] The term utility generally has a bit different coverage within decision-making context.

[3] The assumption can be weakened by a introducing a conditional preferential independence, [4].

information (technological knowledge, physical laws, etc.). Eliciting the needed information itself is inherently hard task, which success depends on experience and skills of an *elicitation expert*. Preferences can be elicited *from past data* directly collected on the underlying decision-making process or from indirect data learned from a number of similar situations. Despite acquiring the probabilistic information from data is well-elaborated, learning can be hard, especially when the space of possible behaviour is larger than that past data cover. Then the initial preferences for the remaining part of the behaviour should be properly assigned.

The process of eliciting of the *domain-specific information* is difficult, time-consuming and error-prone task[4]. Domain experts provide subjective opinions, typically expressed in different and incompatible forms. The elicitation expert should elaborate these opinions into a distribution describing preferences in a consistent way. Significant difficulties emerge when competitive/complementing opinions with respect to the same collection of attributes should be merged. A proper merging their individual opinions within a high-dimensional space of possible behaviour is unfeasible. Besides domain experts having domain-specific information are often unable to provide their opinion on a part of behaviour due to either limited knowledge of the phenomenon behind or the indifference towards the possible instances of behaviour. Then, similarly to the learning preferences from past data, the optimal strategy heavily depends on the initial preferences assigned to the part of behaviour not "covered" by the domain-specific information.

Process of eliciting information itself requires significant cognitive and computational effort of the elicitation expert. Even if we neglect the cost of this effort[5], the elicitation result is always very limited by the expert's *imperfection*, i.e. his inability to devote an infinite deliberation effort to eliciting. Unlike imperfection of experts providing domain-specific information, imperfection of elicitation experts can be eliminated. This motivate the search for a feasible *automated* support of preference elicitation, that does not rely on any elicitation expert.

The *dynamic* decision making strengthes the dependence on the preference elicitation. Indeed, the participant acting within a dynamically changing environment with evolving parameters may gradually change preferences. The intended change may depend on the future behaviour. The overall task is going harder when participant interacts with other dynamic imperfect participants within a common environment.

The paper concerns a construction of probabilistic description of preferences based on the information available. Dynamic decision making under uncertainty from the perspective of an imperfect participant is considered. The participant solves DM task with respect to its environment and based on a given finite set of opinions gained from providers of domain expertise or learned from the past data or both. The set indirectly represents the preferences in a non-unique way[6]. Additionally, the participant may be still uncertain about the preferences on a non-empty subset of behaviour. To design an optimal strategy, a participant employs Fully Probabilistic Design (FPD) of DM strategies, [10, 12] whose specification relies on the notion of *an ideal closed-loop model* which is essentially a probabilistic description of the preferences. In other words, an ideal closed-loop model describes the closed-loop behaviour, when the participant's DM strategy is optimal. FPD searches for the optimal strategy by minimising the divergence of the current closed-loop description on the ideal one. Adopted FPD implies availability of probabilistic description of the environment and probabilistic description of the past closed-loop data.

Section 2 specifies assumptions under which the automated preference elicitation is proposed within the considered FPD. Section 3 describes construction of the ideal closed-loop distribution based on the information provided. The proposed solution is discussed in Section 4 followed by the concluding remarks in Section 5.

[4] It should be mentioned that *practical* solutions mostly use a laborious and unreliable process of manual tuning a number of parameters of the pre-selected utility function. Sometimes the high number of parameters makes this solution unfeasible. Then there are attempts to decrease the number of parameters to reach an acceptable feasibility level.

[5] This effort is usually very high and many sophisticated approaches aim at optimising a trade-off between elicitation cost and value of information it provides (often decision quality is considered), see for instance [3].

[6] Even, when we identify instances of behaviour that cannot be distinguished from the preferences' viewpoint.

2 Assumptions

The considered participant deals with a DM problem, where the reached decision quality is expressed in terms of a ℓ_a-tuple of attributes $a = (a_1, \ldots, a_{\ell_a}) \in \boldsymbol{a} = \prod_{i=1}^{\ell_a} \boldsymbol{a}_i$, $\ell_a < \infty$. $\prod$ denotes Cartesian product of sets $\boldsymbol{a}_i$ the respective attribute entries belong to. The occurrence of attributes depends on an optional ℓ_d-dimensional decision $d = (d_1, \ldots, d_{\ell_d}) \in \boldsymbol{d} = \prod_{j=1}^{\ell_d} \boldsymbol{d}_j$, $\ell_d < \infty$. In the considered preference elicitation problem, the following assumptions are adopted.

A1 The participant is able to specify its preferences on the respective entries of attributes $a_i \in \boldsymbol{a}_i$ such that the most preferred value of each attribute is uniquely defined. For convenience, let the best attribute value be zero.

A2 The participant has at disposal a probabilistic model $\mathcal{M}(a|d)$, which is the probability density (pd[7]) of the attributes a conditioned on decisions d. The support of $\mathcal{M}(a|d)$ is assumed to include $(\boldsymbol{a}, \boldsymbol{d})$.

A3 The participant has[8] a joint pd $\mathcal{P}(a, d)$, describing behaviour (a, d) of the closed loop formed by the acting participant and by its environment[9]. The support of $\mathcal{P}(a, d)$ is assumed to include $(\boldsymbol{a}, \boldsymbol{d})$.

A4 The participant uses fully probabilistic design (FPD), [12], of decision-making strategies. FPD considers a specification of the ideal pd $\mathcal{I}(a, d)$ assigning high values to desired pairs $(a, d) \in (\boldsymbol{a}, \boldsymbol{d})$ and small values to undesired ones. The optimal randomised strategy $\mathcal{S}^{opt}(d)$ is selected among strategy-describing pds $\mathcal{S} \in \boldsymbol{\mathcal{S}}$ as a minimiser of the Kullback-Leibler divergence (KLD, [17])

$$\mathcal{S}^{opt} \in \operatorname{Arg}\min_{\boldsymbol{\mathcal{S}}} \int_{(\boldsymbol{a},\boldsymbol{d})} \mathcal{M}(a|d)\mathcal{S}(d) \ln\left(\frac{\mathcal{M}(a|d)\mathcal{S}(d)}{\mathcal{I}(a,d)}\right) \mathrm{d}(a,d) = \operatorname{Arg}\min_{\boldsymbol{\mathcal{S}}} \mathcal{D}(\mathcal{M}\mathcal{S}||\mathcal{I}).$$

Note that the use of FPD represents no constraints as for a classical preference-quantifying utility $\mathcal{U}(a, d) : (\boldsymbol{a}, \boldsymbol{u}) \rightarrow [-\infty, \infty)$ it suffices to consider the ideal pd of the form

$$\mathcal{I}(a,d) = \frac{\mathcal{M}(a|d)\exp(\mathcal{U}(a,d)/\lambda)}{\int_{(\boldsymbol{a},\boldsymbol{d})} \mathcal{M}(a|d)\exp(\mathcal{U}(a,d)/\lambda)\,\mathrm{d}(a,d)},\ \lambda > 0.$$

Then, the FPD with such an ideal pd and $\lambda \rightarrow 0$ arbitrarily well approximates the standard Bayesian maximisation of the expected utility [15].

3 Preference Elicitation

Under the assumptions **A1** – **A4**, the addressed elicitation problem reduces to a justified, algorithmic (elicitation-expert independent) construction of the ideal pd $\mathcal{I}(a, d)$.

The following steps constitute the considered construction of the preference-expressing ideal.

S1 Each ideal pd $\mathcal{I}(a, d)$ determines marginal pds $\mathcal{I}_i(a_i)$ on the respective attribute entries $a_i \in \boldsymbol{a_i}$, $i = 1, \ldots, \ell_a$. The marginal ideal pd $\mathcal{I}_i(a_i)$ respects the highest preference for $a_i = 0$ if

$$\mathcal{I}_i(a_i = 0) \geq \mathcal{I}_i(a_i),\ \forall a_i \in \boldsymbol{a}_i. \tag{1}$$

Thus, the ideal pds meeting (1) for $i = 1, \ldots, \ell_a$ respect the participant's preferences.

S2 A realistic ideal pds (meeting (1)) should admit a complete fulfilment of preferences with respect to any individual attribute entry a_i whenever the design focuses solely on it. It is reasonable to restrict ourselves to such ideal pds as the ideal pd, which cannot be reached at least with respect to individual attributes is unrealistic.

[7] pd, Radon-Nikodým derivative [21] of the corresponding probabilistic measure with respect to a dominating, decision-strategy independent, measure denoted d.

[8] or can learn it

[9] The closed-loop model $\mathcal{P}(a, d)$ can alternatively describe a usual behaviour of other participants in similar DM tasks.

The complete fulfilment of preferences requires an existence of decision strategy $\mathcal{S}_i(d)$ such that the closed-loop model $\mathcal{M}(a|d)\mathcal{S}_i(d)$ has the marginal pd on a_i equal to the corresponding marginal $\mathcal{I}_i(a_i)$ of the considered ideal pd $\mathcal{I}(a,d)$.

FPD methodology is used to specify *realistic marginal pds*, $\mathcal{I}_i^r(a_i)$, $i = 1,\dots,\ell_a$.

S3 The set of ideal pds $\mathcal{I}(a,d)$ having given realistic marginal pds $\mathcal{I}_i^r(a_i)$, $i = 1,\dots,\ell_a$ is non-empty as it contains the ideal pd independently combining the expressed marginal preferences $\mathcal{I}(a,d) = \prod_{i=1}^{\ell_a} \mathcal{I}_i^r(a_i)$. Generally, the discussed set contains many pds. Without a specific additional information, the chosen pd should at least partially reflect behaviour occurred in the past. Then the adequate representant of this set is the minimiser of the KLD [22] of $\mathcal{I}(a,d)$ on the joint pd $\mathcal{P}(a,d)$. According to **A3** $\mathcal{P}(a,d)$ describes the past closed-loop behaviour and serves as the most uncertain (the least ambitious) ideal: in the worst case, the ideal pd qualifies the past behaviour as the best one. The minimiser over the set of ideal pds having marginal pds $\mathcal{I}_i^r(a_i)$, $i = 1,\dots,\ell_a$, is described below and provides the final solution of the addressed elicitation problem.

The pds $\mathcal{I}_i^r(a_i)$, discussed in Step **S2** can be obtained as follows. Let us consider the ith entry a_i. Then ℓ_a-tuple a can be split $a = (a_{-i}, a_i)$, where a_{-i} contains all attributes except a_i and the ideal pd factorises [20]

$$\mathcal{I}(a,d) = \mathcal{I}_i(a_{-i}, d|a_i)\mathcal{I}_i(a_i). \tag{2}$$

When solely caring about the ith attribute, any distribution of (a_{-i}, d) can be accepted as the ideal one. This specifies the first factor of the ideal pd (2) as ([11])

$$\mathcal{I}_i^l(a_{-i}, d|a_i) = \frac{\mathcal{M}(a|d)\mathcal{S}(d)}{\int_{(\boldsymbol{a}_{-i},\boldsymbol{d})} \mathcal{M}(a|d)\mathcal{S}(d)\,\mathrm{d}(a_{-i}, d)}. \tag{3}$$

This choice, complemented by an arbitrary choice of $\mathcal{I}_i(a_i)$ specifies an ideal pd on $(\boldsymbol{a},\boldsymbol{d})$ and the strategy ${}^i\mathcal{S}(d)$ minimising KLD of the closed-loop model $\mathcal{MS}$ on it cares about the ith attribute only. For the inspected ideal pd, the optimised KLD optimised with respect a strategy $\mathcal{S}$ reads

$$\mathcal{D}(\mathcal{MS}||\mathcal{I}) = \int_{(\boldsymbol{a},\boldsymbol{d})} \mathcal{M}(a|d)\mathcal{S}(d) \ln\left(\frac{\int_{\boldsymbol{d}} \mathcal{M}(a_i|d)S(d)\,\mathrm{d}d}{\mathcal{I}_i(a_i)}\right) \mathrm{d}(a,d). \tag{4}$$

Let us assume that there is ${}^id \in \boldsymbol{d}$ such that $\mathcal{M}(a_i = 0|\,{}^id) \geq \mathcal{M}(a_i|\,{}^id),\ \ \forall a_i \in \boldsymbol{a}_i$. Then, the ideal pd $\mathcal{I}(a,d) = \mathcal{I}_i(a_{-i}, d|a_i)\mathcal{I}_i^r(a_i)$ with

$$\mathcal{I}_i^r(a_i) = \mathcal{M}(a_i|\,{}^id) \tag{5}$$

meets (1) and is the realistic marginal pd in the sense described in **S2**. Indeed, the deterministic strategy ${}^i\mathcal{S}(d) = \delta(d-\,{}^id) =$ pd concentrated on ${}^id$ and ideal pd $\mathcal{I}_i^l\mathcal{I}_i^r$ make the KLD (4) $\mathcal{D}(\mathcal{M}\,{}^i\mathcal{S}||\mathcal{I}_i^l\mathcal{I}_i^r)$ equal to zero, which is the absolute minimum.

The constraints (5) on the marginal ideal pds exhaust all information about the preferences available, see **A1** – **A3**. It remains to select one among multitude of such ideal pds meeting (5). The minimum KLD (cross-entropy) principle [22] recommends to select the ideal pd, which minimises its KLD on a pd representing the most uncertain preference description. As discussed in **S3**, the pd describing the past history serves to this purpose. The following proposition explicitly specifies the minimiser and provides the solution of the addressed preference elicitation problem.

Proposition 1 (The recommended ideal pd) *The ideal pd $\mathcal{I}(a,d)$ describing the supplied preferences via (5) and minimising KLD $\mathcal{D}(\mathcal{I}||\mathcal{P})$, where $\mathcal{P}$ describes the past history, has the form*

$$\mathcal{I}(a,d) = \mathcal{P}(d|a)\prod_{i=1}^{\ell_a} \mathcal{I}_i^r(a_i) \tag{6}$$

$$= \quad \frac{\mathcal{P}(d,a)}{\int_{\boldsymbol{d}} \mathcal{P}(a,d)\,\mathrm{d}d} \prod_{i=1}^{\ell_a} \mathcal{M}(a_i|\,{}^id),\ \ \textit{with}\ {}^id \in \boldsymbol{d} :\ \mathcal{M}(a_i = 0|\,{}^id) \geq \mathcal{M}(a_i|\,{}^id),\ \forall i \in \{1,\dots,\ell_a\}.$$

Proof

The convex functional $\mathcal{D}(\mathcal{I}||\mathcal{P})$ on the convex set given by considered constraints (5) has the unique global minimum. Thus, it suffices consider weak variations of the corresponding Lagrangian functional. The pd (6) makes them equal to zero and thus it is the global minimiser searched for.

□

4 Discussion

Many important features of the proposed solution (6) is implied by the fact that the constructed ideal pd reflects the relation $\mathcal{M}(a|d)$ between attributes and decisions. Specifically,

- The marginal ideal pds (5) are *not* fully concentrated on the most desirable attribute value (0), which reflects the fact that $a_i = 0$ cannot be reached with certainty.
- A specific ${}^{b}d$ is a bad decision comparing to other ${}^{o}d$ if $\mathcal{P}(a = 0|\,{}^{b}d) << \mathcal{P}(a = 0|\,{}^{o}d)$. As the closed-loop model $\mathcal{P}(a, d) = \mathcal{P}(a|d)\mathcal{P}(d)$ is a factor in (6), the decision ${}^{b}d$ is perceived as a bad one by the ideal pd (6) unless an unbalanced experience is faced, i.e. unless $\mathcal{P}(\,{}^{b}d) >> \mathcal{P}(\,{}^{o}d)$. Thus, the constructed ideal distinguishes the good and bad decisions made in past if they both occur in a balanced way.

  The danger of an unbalanced occurrence of good and bad decisions can be counteracted by modifying $\mathcal{P}(a, d)$ in order to stimulate exploration. It suffices to take it as a mixture of the closed-loop model gained from observations and of an exploration-allowing "flat" pd.
- The functional form of the ideal pd is determined by the model $\mathcal{M}(a|d)$: it is not created in an ad hoc, model independent, manner unlike utilities [16].
- It is always possible to project the constructed ideal pd into a class of feasible pds by using information criterion justified in [2, 14], if the constructed ideal pd is too complex for numerical treatment or analysis.
- The model $\mathcal{M}(a|d)$ as well as the closed-loop model of the past history $\mathcal{P}(a, d)$ can be learnt in a standard Bayesian way [1, 20]. Consequently, the preference description (6), derived from them, is learned, too.
- The involved pds can quantify the joint distributions of discrete-valued as well as continuous valued attributes. This simplifies the elicitation of preferences given jointly by categorical and numerical attributes.
- The approach can be directly extended to a dynamic DM, in which attributes and decisions evolve in time. It suffices to apply Proposition 1 to factors of involved pds.
- The construction can be formally performed even when several best (mutually ordered) attributes are admitted in a variant of Assumption **A1**. The subsequent evaluations following the same construction line are, however, harder.
- The considered preference specification is quite common but it does not cover all possibilities. For instance, an attribute $a_i \in \boldsymbol{a}_i$ may have preferences specified on a proper subset $\emptyset \neq \boldsymbol{\alpha}_i \subset \boldsymbol{a}_i$. If $a_i = 0 \in \boldsymbol{\alpha}_i$ is considered as the most desirable value of the attribute, the proposed elicitation way applies with a reduced requirement $\mathcal{M}(a_i = 0|\,{}^{i}d) \geq \mathcal{M}(a_i|\,{}^{i}d)$, $\forall a_i \in \boldsymbol{\alpha}_i$, cf. (1). Then, the proposed procedure can be used without essential changes. The real problem arises when there is no information whether the most preferred attribute is in $\prod_{i=1}^{\ell_a} \boldsymbol{\alpha}_i$ or not. Then, the participant has to provide an additional feedback by specifying a rank of the newly observed attribute with respect to the initially set values 0. The problem is tightly connected with a sequential choice of the best variant, e.g., [19].

5 Concluding Remarks

The solution proposes a methodology of automated preference elicitation of the ideal pd for a common preference specification. Covering other preference specifications is the main problems to be addressed. Also, the proposed solution is to be connected with an alternative view presented in [13], where the preference elicitation was directly treated as a learning problem and reduced to a specification of a prior pd on parameters entering environment model (and thus learnable) and parameters entering only the ideal pd (and thus learnable only via a well-specified join prior pd). The design of specific algorithmic solutions for commonly used environment models is another topic to be covered. In spite of the width of the problems hidden behind these statements, the selected methodological direction is conjectured to be adequate and practically promising.

Acknowledgments

The support of the projects GAČR 102/08/0567 and MŠMT 1M0572 is acknowledged.

References

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

[2] J. M. Bernardo. Expected information as expected utility. *The Annals of Statistics*, 7(3):686–690, 1979.

[3] C. Boutilier. A POMDP formulation of preference elicitation problems. In S. Biundo, editor, *AAAI-02 Proc. of the Fifth European Conference on Planning*, pages 239–246. AAAI Press/MIT Press, Durham, UK, 2002.

[4] C. Boutilier, R.I. Brafman, C. Geib, and D. Poole. A constraint-based approach to preference elicitation and decision making. pages 19–28. Stanford, CA, 1997.

[5] U. Chajewska and D. Koller. Utilities as random variables: Density estimation and structure discovery. In *Proceedings of UAI–00*, pages 63–71. 2000.

[6] N. Cooke. Varieties of knowledge elicitation techniques. *International Journal of Human-Computer Studies*, 41:801–849, 1994.

[7] K. Gajos and D.Weld. Preference elicitation for interface optimization. 2005.

[8] S. Zilles H. P.Viappiani and C. Boutilier. Learning complex concepts using crowdsourcing: A Bayesian approach. In *Proceedings of the Second Conference on Algorithmic Decision Theory (ADT-11)*. Piscataway, NJ.

[9] H.B. Jimison, L.M. Fagan, R.D. Shachter, and E.H. Shortliffe. Patient-specific explanation in models of chronic disease. *AI in Medicine*, 4:191 –205, 1992.

[10] M. Kárný. Towards fully probabilistic control design. *Automatica*, 32(12):1719–1722, 1996.

[11] M. Kárný, J. Böhm, T. V. Guy, L. Jirsa, I. Nagy, P. Nedoma, and L. Tesař. *Optimized Bayesian Dynamic Advising: Theory and Algorithms*. Springer, London, 2006.

[12] M. Kárný and T. V. Guy. Fully probabilistic control design. *Systems & Control Letters*, 55(4):259–265, 2006.

[13] M. Kárný and T.V. Guy. Preference elicitation in fully probabilistic design of decision strategies. In *Proc. of the 49th IEEE Conference on Decision and Control*. IEEE, 2010.

[14] M. Kárný and T.V. Guy. On support of imperfect bayesian participants. In T.V. Guy, M. Kárný, and D.H. Wolpert, editors, *Decision Making with Imperfect Decision Makers*, volume 28. Springer, Berlin, 2012. Intelligent Systems Reference Library.

[15] M. Kárný and T. Kroupa. Axiomatisation of fully probabilistic design. *Inf. Sciences*, 2011.

[16] R.L. Keeney and H. Raiffa. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. JohnWiley and Sons Inc., 1976.

[17] S. Kullback and R. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–87, 1951.

[18] G. Linden, S. Hanks, and N. Lesh. Interactive assessment of user preference models: The automated travel assistant. 1997.

[19] A.A.J. Marley and J.J. Louviere. *Journal of Mathematical Psychology*, 49(6):464–480, 2005.

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

[21] M.M. Rao. *Measure Theory and Integration*. John Wiley, New York, 1987.

[22] J. Shore and R. Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on Information Theory*, 26(1):26–37, 1980.

Effect of emotion on the imperfectness of decision making

Alessandro A.E.P. Villa[*]
NeuroHeuristic Research Group
HEC-ISI University of Lausanne
CH-1015 Lausanne, Switzerland
alessandro.villa@nhrg.org

Marina Fiori
Institute of Psychology
University de Lausanne
CH-1015 Lausanne, Switzerland
Marina.Fiori@unil.ch

Sarah Mesrobian
NeuroHeuristic Research Group
HEC-ISI University of Lausanne
CH-1015 Lausanne, Switzerland
sarah.mesrobian@nhrg.org

Alessandra Lintas
NeuroHeuristic Research Group
HEC-ISI University of Lausanne
CH-1015 Lausanne, Switzerland
alessandra.lintas@nhrg.org

Vladyslav Shaposhnyk
NeuroHeuristic Research Group
HEC-ISI University of Lausanne
CH-1015 Lausanne, Switzerland
vladyslav.shaposhnyk@nhrg.org

Pascal Missonnier
NeuroHeuristic Research Group
HEC-ISI University of Lausanne
CH-1015 Lausanne, Switzerland
pascal.missonnier@nhrg.org

Abstract

Human decision making has demonstrated imperfectness and essential deviation from rationality. Emotions are a primary driver of human actions and the current study investigates how perceived emotions may affect the behavior during the Ultimatum Game (UG), while recording event-related potentials (ERPs) from scalp electrodes. We observed a negative correlation ($p < 0.001$) between positive emotions, in particular happiness, and the amount offered by a participant acting as a Proposer in the UG. Negative emotions, in particular fear, showed a positive correlation ($p < 0.05$) with the offer. The ERPs revealed invariant components at short latency in brain activity in posterior parietal areas irrespective of the Responder or Proposer role. Conversely, significant differences appeared in the activity of central and frontal areas between the two conditions at latencies 300-500 ms.

1 Introduction

Although research has demonstrated the substantial role emotions play in decision-making and behavior [1] traditional economic models emphasize the importance of rational choices rather than their emotional implications. The concept of expected value is the idea that when a rational agent must choose between two options, it will compute the utility of outcome of both actions, estimate their probability of occurrence and finally select the one that offers the highest gain. In the field of neuroeconomics a few studies have analyzed brain and physiological activation during economical monetary exchange [2, 3] revealing that activation of the insula and higher skin conductance [4] were associated to rejecting unfair offers. The aim of the present research is to further extend the understanding of emotions in economic decision-making by investigating the role of basic emotions

[*] http://www.neuroheuristic.org

(happiness, anger, fear, disgust, surprise, and sadness) in the decision-making process. To analyze economic decision-making behavior we used the Ultimatum Game (UG) task [5] while recording EEG activity. This task has been widely used to investigate human interaction, in particular the differences between behavior expected according to the 'rational' model of game theory and observed 'irrational' behavior. One hypothesis that has been suggested to explain this divergence is that participants tend to engage in the 'tit-for-tat' type of choice establishing a sort of reciprocity rule [6]. In order to examine potential interaction effects between reciprocity rules and emotions we employed repetitive trials to analyze the evolution of participants' strategy along the game. In addition, we analyzed the role of individual differences, in particular the personality characteristic of honesty and the tendency to experience positive and negative emotions, as factors potentially affecting the monetary choice. [7].

2 Materials and methods

2.1 Behavioral paradigm

We administered participants some questionnaire to measure their personality traits (the Hexaco personality questionnaire, [8]) as well as their tendency to experience positive and negative affect (the PANAS scale [9]). The Ultimatum Game (UG) is an anonymous, single-shot two-player game, in which the "Proposer" (Player 1) has a certain sum of money at his disposal and must propose a share to the "Responder" (Player 2) [5]. The Responder can either accept or reject this offer. If the Responder accepts the proposal, the share is done accordingly. However, if the Responder refuses, both players end up with nothing. In either case the game ends after the Responder's decision. The Subjects were comfortably seated in a sound- and light-attenuated room, watched a computer-controlled monitor at a distance of 57 cm, and were instructed to maintain their gaze on a central fixation cross throughout the experiment. Subjects volunteered to participate in the study and played with virtual money. They were tested along three series, each one composed of 2 Blocks. During the first Block the participants acted as Proposers (Fig. 1a), while during the second Block the computer made the offer and the participants acted as Responders (Fig. 1b). Each Block was composed by 30 trials, which means that 90 trials were collected overall for each condition. The task was implemented on a personal computer using the E-Prime software (Psychology Software Tools, Inc., Sharpsburg, PA 15215-2821 USA).

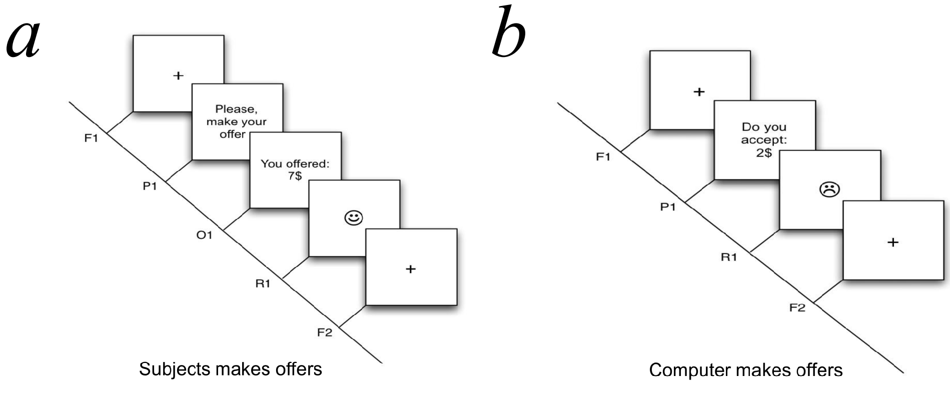


Figure 1: Illustration of Ultimatum Game task along series composed of 2 Blocks. During the first Block the participants acted as Proposers (a), while during the second Block the computer made the offer and the humans acted as Responders (b).

Participants were subtly primed with emotional figures while making the decision to share money with, or accept the offer of, an hypothetical partner. Becoming aware of an emotional state may hinder its effect on subsequent behavior. Thus, we instructed participants to make their economic decision while keeping in the background emotional images, which were meant to induce different

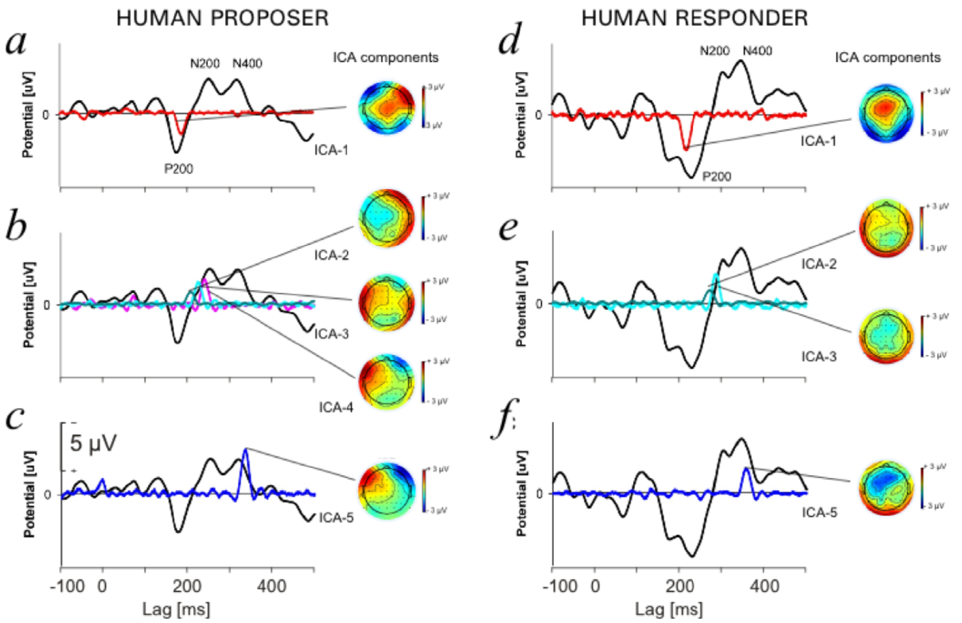


Figure 2: Active Indipendent Component Analysis-components (ICA-1 to ICA-5) between 200 ms and 350 ms accounting for ERP variance in each of the condition tasks. The topographic distributions of ICA-components are presented on the left side of the grand average waveforms.

emotional states. The images were selected among a broad set of pictures painted by the artist Rose Coleman. In contrast to most of the pictures used in databases of the kind of the International Affective Picture System (IAPS) [10, 11] our data set is based on non-figurative abstract pictures. Participants were asked to rate the emotional content of the image only at the end of the experiment.

2.2 Electrophysiological recording

Continuous EEG was recorded using 64 surface electrodes (ActiveTwo MARK II Biosemi EEG System, BioSemi B.V., Amsterdam). Electrophysiological signals were sampled at 2048 Hz with lower cutoff at 0.05 Hz and upper cutoff at 200 Hz (DC ampliers and software by BioSemi B.V.). The electro-oculogram was recorded using two pairs of bipolar electrodes in both vertical and horizontal directions. Time stamps of visual stimuli presentations and keyboard press gestures were recorded with markers in the continuous EEG data le. The start of a trial was initiated by pressing the spacebar at the beginning. The EEG recordings were analyzed with NeuroScan software (NeuroScan Inc, Herndon, VA, USA) and open source softwares. ERPs were averaged with a 200 ms baseline epoch prior to trigger onset and band-pass filtered from 0.3 Hz to 30 Hz.

3 Results

We started by analyzing data in one of the 2 experimental conditions: when the participants offered an amount of money to share to an hypothetical partner. Overall offers were balanced around average (average= 5.03, SD= 1.34 on a scale from 1 to 9 CHF). A first analysis conducted across subjects (IBM-SPSS version 19, Chicago, IL, USA) revealed a negative correlation between the emotional content of the figure in the background, in particular related to happiness, and the amount that was offered, $r(630) = .14$, $p < 0.001$. Positive correlations were found with emotions such as fear, $r(630) = .12$, $p < 0.05$ and sadness, $r(630) = .09$, $p < 0.05$. To further explore this relationship we created two clusters of emotional contents of the figures employed in the background, one indicating high content of positive emotions, such as surprise and happiness, and the other with high content of negative emotions, such as disgust, fear and sadness. Then we conducted a 1-way ANOVA in which emotional content (positive and negative) was used to predict the amount of money offered. Preliminary results revealed that indeed positive content predicted a lower offer than negative content, $F(2, 629) = 10.42$, $p < 0.001$.

In the electrophysiological recordings we observed significant differences in the latencies of the P200 peak, with human Responders' peak occurring 30-50 ms later than the peak elicited in Proposers. The N400 latency was similar for central and frontal areas but it was about 100 ms longer than the N400 latency observed in Pz. This wave is likely to be associated with information processing and response preparation, especially in the central and frontal areas. Figure 2 shows that Independent Component Analysis (ICA, [12]) revealed ICA-components ICA-1 and ICA-5 associated to the previously described P200 and N400 components of each experimental condition. The ICA-1 component accounted for the same percentage of variance in Proposer and Responder conditions (17% and 11%, respectively; Fig. 2a,d). Differences between the conditions appeared for the ICA-5 component, which explained 22% of the variance for Proposer and only 7% for the Responder condition (Fig. 2c,f). ICA-2, ICA-3 and ICA-4 explained 15%, 17% and 34%, respectively, of the variance in Proposer condition. On the opposite only two components ICA-2 and ICA-3 explained 9% and 18%, respectively, of the variance in the Responders.

4 Discussion

We presented here only preliminary results on the effect of emotions on Proposer's decision making. Data from the other experimental condition, namely when the participant decided to accept or refuse an offer, will allow us to investigate whether the same emotional content may produce similar or different effects on economic decision-making. In particular, the analysis we plan to conduct will focus on the relationship among monetary decision, indirect measures of emotions with EEG, and subjective evaluation of the emotional images shown during decision making. Finally we plan to analyze whether emotional state will mediate the relationship between certain personality characteristics, such as greed avoidance, and certain decision outcomes, such as higher sharing offers. The combined psychological and neurophysiological approach will allow us to produce a model of the most likely neurological circuit suitable to be influenced by affective choices. Our results will provide additional evidence to the role emotions and individual differences play in economic decision-making.

Acknowledgments

The Authors acknowledge the support of the Swiss National Science Foundation grant CR13I1-138032.

References

[1] A Bechara and A Damasio. The somatic marker hypothesis: A neural theory of economic decision. *Games Econ Behav*, 52(2):336–372, 2005.

[2] A Bechara, H Damasio, D Tranel, and A R Damasio. Deciding advantageously before knowing the advantageous strategy. *Science*, 275(5304):1293–1295, Feb 1997.

[3] A G Sanfey, J K Rilling, J A Aronson, L E Nystrom, and J D Cohen. The neural basis of economic decision-making in the Ultimatum Game. *Science*, 300(5626):1755–1758, Jun 2003.

[4] M van't Wout, L J Chang, and A G Sanfey. The influence of emotion regulation on social interactive decision-making. *Emotion*, 10(6):815–821, Dec 2010.

[5] Werner Güth. The Generosity Game and Calibration of Inequity Aversion. *Journal of Socio-Economics*, 39:155–157, 2010.

[6] Yu-Han Chang, Tomer Levinboim, and Rajiv Maheswaran. The social ultimatum game. In Tatiana Valentine Guy, Miroslav Kárný, and David H. Wolpert, editors, *Decision Making with Imperfect Decision Makers*, volume 28 of *Intelligent Systems Reference Library*, chapter 6, pages 135–158. Springer Verlag, Berlin Heidelberg, Germany, 2011.

[7] Alessandro E. P. Villa, Pascal Missonnier, and Alessandra Lintas. Neuroheuristics of Decision Making: from neuronal activity to EEG. In Tatiana Valentine Guy, Miroslav Kárný, and

David H Wolpert, editors, *Decision Making with Imperfect Decision Makers*, volume 28 of *Intelligent Systems Reference Library*, chapter 7, pages 159–194. Springer-Verlag, Berlin Heidelberg, Germany, 2011.

[8] M C Ashton and K Lee. The HEXACO-60: a short measure of the major dimensions of personality. *J Pers Assess*, 91(4):340–345, Jul 2009.

[9] D Watson, L A Clark, and A Tellegen. Development and validation of brief measures of positive and negative affect: the PANAS scales. *J Pers Soc Psychol*, 54(6):1063–1070, Jun 1988.

[10] M. M. Bradley and P. J. Lang. The International Affective Picture System (IAPS) in the study of emotion and attention. In J. A. Coan and J. J. B. Allen, editors, *Handbook of Emotion Elicitation and Assessment*, pages 29–46. Cambridge University Press, New York, USA, 2007.

[11] E S Dan-Glauser and K R Scherer. The Geneva affective picture database (GAPED): a new 730-picture database focusing on valence and normative significance. *Behav Res Methods*, 43(2):468–477, Jun 2011.

[12] A Delorme and S Makeig. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J Neurosci Methods*, 134(1):9–21, Mar 2004.

An Adversarial Risk Analysis Model for an Emotional Based Decision Agent

Javier G. Rázuri
Department of Statistics and Operations Research & AISoy Robotics
Universidad Rey Juan Carlos
Madrid
javierfrancisco.guerrero.razuri@urjc.es

Pablo G. Esteban
Department of Statistics and Operations Research & AISoy Robotics
Universidad Rey Juan Carlos
Madrid
pablo.gomez.esteban@urjc.com

David Ríos Insua
Royal Academy of Sciences, Spain
david.rios@urjc.es

Abstract

We introduce a model that describes the decision making process of an autonomous synthetic agent which interacts with another agent and is influenced by affective mechanisms. This model would reproduce patterns similar to humans and regulate the behavior of agents providing them with some kind of emotional intelligence and improving interaction experience. We sketch the implementation of our model with an edutainment robot.

1 Introduction

We have recently introduced in [19] the framework of Adversarial Risk Analysis (ARA) to cope with risk analysis of situations in which risks stem from deliberate actions of intelligent adversaries. ARA has a Bayesian game theoretic flavor, as in [11] and [17]. In supporting one of the participants, the problem is viewed as a decision analytic one, but principled procedures which employ the adversarial structure, and other information available, are used to assess the probabilities on the opponents' actions. There is a potentially infinite analysis of nested decision models arrived at when using ARA. This is in the realm of incomplete Bayesian games, see [8], which avoids the infinite regress by using the common (prior) knowledge assumption. We feel that this is a very strong hypothesis which is not tenable in our application domain. We prefer to be realistic and accommodate as much information as we can from intelligence into our analysis, through a structure of nested decision models. Depending on the level we climb up in such hierarchy of nested models, we talk about 0-level analysis, 1-level analysis and so on, see [1] and the discussion [10]. [1], [18] and [19] have introduced different principles to end up the hierarchy. In this paper, we shall explore how the ARA framework may support the decision making of an autonomous emotional agent in its interaction with a user.

Over the last several years, researchers in the field of cognitive processes have shown that emotions have a direct impact on judgment and decision-making tasks. This has vertebrated fields such as affective computing [16], affective decision making ([13] and [4]) and neuroeconomics [6]. Based on some of their concepts, we develop a model that allows an agent to make decisions influenced by emotional factors within the ARA framework.

The model is essentially multi-attribute decision analytic, see [3], but our agent entertains also models forecasting the evolution of its adversary and the environment surrounding them. We also include models simulating emotions, which have an impact over the agent's utility function. In such a way, we aim at better simulating human decision making and, specially, improving interfacing and interaction with users.

2 Basic Elements

We start by introducing the basic elements of our model. We aim at designing an agent A whose activities we want to regulate and plan. There is another participant B, the user, which interacts with A. The activities of both A and B take place within an environment E. As a motivating example, suppose that we aim at designing a bot A which will interact with a kid B within a given room E.

A makes decisions within a finite set $\mathcal{A} = \{a_1, \ldots, a_m\}$, which possibly includes a *do nothing* action. B makes decisions within a set $\mathcal{B} = \{b_1, \ldots, b_n\}$, which also includes a *do nothing* action. $\mathcal{B}$ will be as complete as possible, while simplifying all feasible results down to a finite number. It may be the case that not all user actions will be known a priori. This set could grow as the agent learns, adding new user actions, as we discuss in our conclusions. The environment E changes with the user actions. The agent faces this changing environment, which affects its own behavior. We assume that the environment adopts a state within a finite set $\mathcal{E} = \{e_1, \ldots, e_r\}$.

A has q sensors which provide readings and are the window through which it perceives the world. Sensory information originating in the external environment plays an important role in the intensity of the agent's emotions, which, in turn, affect its decision-making process. Each sensor reading is attached to a time t, so that the sensor reading vector will be $s_t = (s_t^1, \ldots, s_t^q)$. The agent infers the external environmental state e, based on a transformation a function f, so that

$$\hat{e}_t = f(s_t).$$

A also uses the sensor readings to infer what the user has done, based on a (possibly probabilistic) function g

$$\hat{b}_t = g(s_t).$$

We design our agent with an embedded *management by exception* principle, see [23]. Under normal circumstances, its activities will be planned according to the basic loop shown in Figure 1. This is open to interventions if an exception occurs.

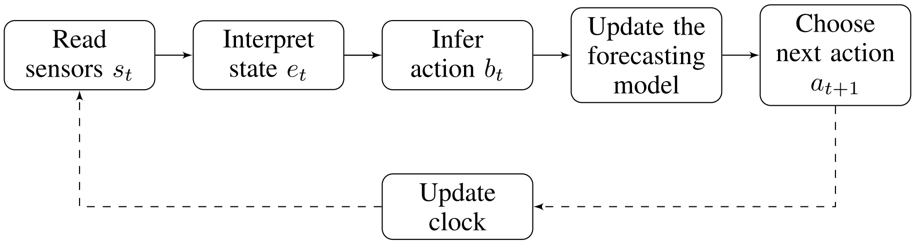


Figure 1: Basic Agent Loop

3 ARA Affective Decision Model

Essentially, we shall plan our agent's activities over time within the decision analytic framework, see [3]. We describe, in turn, the forecasting model (which incorporates the ARA elements), the preference model (which incorporates the affective elements) and the optimization part.

3.1 Forecasting Models

The agent maintains forecasting models which suggest him with which probabilities will the user act and the environment react, given the past history of its actions, the user's actions and the evolution of the environment. We describe the general structure of such models.

Assume that, for computational reasons, we limit the agent's memory to two instant times, so that we are interested in computing

$$p(e_t, b_t \mid a_t, (e_{t-1}, a_{t-1}, b_{t-1}), (e_{t-2}, a_{t-2}, b_{t-2})).$$

Extensions to k instants of memory or forecasting m steps ahead follow a similar path. The above may be decomposed through

$$p(e_t|b_t, a_t, (e_{t-1}, a_{t-1}, b_{t-1}), (e_{t-2}, a_{t-2}, b_{t-2})) \times p(b_t|a_t, (e_{t-1}, a_{t-1}, b_{t-1}), (e_{t-2}, a_{t-2}, b_{t-2})).$$

We assume that the environment is fully under the control of the user. As an example, the user controls the light, the temperature and other features of the room. Moreover, he may plug in the bot to charge its battery, and so on. Only the latest action of the user will affect the evolution of the environment. Thus, we shall assume that

$$p(e_t \mid b_t, a_t, (e_{t-1}, a_{t-1}, b_{t-1}), (e_{t-2}, a_{t-2}, b_{t-2})) = p(e_t \mid b_t, e_{t-1}, e_{t-2}).$$

We term this the environment model.

Similarly, we shall assume that the user has its own behavior evolution, that might be affected by how does he react to the agent actions, thus incorporating the ARA principle, so that

$$p(b_t \mid a_t, (e_{t-1}, a_{t-1}, b_{t-1}), (e_{t-2}, a_{t-2}, b_{t-2})) = p(b_t \mid a_t, b_{t-1}, b_{t-2}). \tag{1}$$

The agent will maintain two models for such purpose. The first one describes the evolution of the user by itself, assuming that he is in control of the whole environment, and he is not affected by the agent's actions. We call it the user model, and is described through

$$p(b_t \mid b_{t-1}, b_{t-2}).$$

The other one refers to the user's reactions to the agent's actions. Indeed, that the user is fully reactive to the agent, which we describe through

$$p(b_t \mid a_t).$$

We call it the classical conditioning model, with the agent conditioning the user.

We combine both models to recover (1). For example, under appropriate independence conditions we recover it through

$$p(b_t \mid a_t, b_{t-1}, b_{t-2}) = (p(b_t \mid b_{t-1}, b_{t-2})\ p(b_t \mid a_t))/p(b_t).$$

Other combinations are required if such conditions do not hold. In general, we could view the problem as one of model averaging, see [9]. In such case

$$p(b_t \mid a_t, b_{t-1}, b_{t-2}) = p(M_1)p(b_t \mid b_{t-1}, b_{t-2}) + p(M_2)p(b_t \mid a_t),$$

where $p(M_i)$ denotes the probability that the agent gives to model i, which, essentially, capture how much reactive to the agent's actions the user is. Learning about various models within our implementation is sketched in Section 4.

3.2 Affective Preference Model

We describe now the preference model, which incorporates the affective principles. We shall assume that the agent faces multiple consequences $c = (c_1, c_2, \ldots, c_l)$. At each instant t, such consequences depend on his action a_t, the user's action b_t and the future state e_t, realized after a_t and b_t. Therefore, the consequences will be of the form

$$c_i(a_t, b_t, e_t),\ i = 1, \ldots, l.$$

We assume that they are evaluated through a multi-attribute utility function, see [3]. Specifically, without much loss of generality, see [24], we shall adopt an additive form

$$u(c_1, c_2, \dots, c_l) = \sum_{i=1}^{l} w_i u_i(c_i),$$

with $w_i \geq 0,\ \sum_{i=1}^{l} w_i = 1$.

The consequences might be perceived differently depending on the current emotional state d_t of the agent. We shall define it in terms of the level of k basic emotions, through a mixing function

$$d_t = h(em_t^1, em_t^2, \dots, em_t^k).$$

[5] and [22] provide many pointers to the literature on mixing emotions. The intensity of these basic emotions, in turn, will be defined in terms of how desirable a situation is, i.e. how much utility $u(c_t)$ is gained, and how surprising the situation was, see [7] for an assessment of such models. The expectations, or surprise, will be defined by comparing the predicted and the actual (inferred) user action through some distance function

$$z_t = d(\bar{b}_t, \hat{b}_t),$$

where $\bar{b}^t$ is (the most likely) predicted user action. We assume some stability within emotions, in that current emotions influence future emotions. Thus, we assume a probabilistic evolution of emotions through

$$em_t^i = r_i(em_{t-1}^i, u(c_t), z_t).$$

Finally, following [13], we shall actually assume that the utility weights will depend on the emotional state, the stock of visceral factors in their notation, so that

$$u(c) = \sum_{i=1}^{l} w_i(d) u_i(c_i).$$

3.3 Expected Utility

The goal of our agent will be to maximize the predictive expected utility. Planning m instants ahead requires computing maximum expected utility plans defined through:

$$\max_{(a_t, \dots, a_{t+m})} \psi(a_t, \dots, a_{t+m}) = \sum_{(b_t, e_t) \dots, (b_{t+m}, e_{t+m})} \left[\sum_{i=1}^{m} (u(a_{t+i}, b_{t+i}, e_{t+i})) \right] \times$$

$$\times p((b_t, e_t), \dots, (b_{t+m}, e_{t+m}) \mid (a_t, a_{t+1}, \dots, a_{t+m}, (a_{t-1}, b_{t-1}, e_{t-1}), (a_{t-2}, b_{t-2}, e_{t-2}))).$$

assuming utilities to be additive over time. This could be done through dynamic programming. If planning m instants ahead turns out to be very expensive computationally, we could plan just one period ahead. In this case, we aim at solving

$$\max_{a_t \in \mathcal{A}} \psi(a_t) = \sum_{b_t, e_t} u(a_t, b_t, e_t) \times p(b_t, e_t \mid a_t, (a_{t-1}, b_{t-1}, e_{t-1}), (a_{t-2}, b_{t-2}, e_{t-2})).$$

We may mitigate the myopia of this approach by adding a term penalizing deviations from some ideal agent consequences, as in [21]. In this case, the utility would have the form $u(c) - \rho(c, c^*)$ where ρ is a distance and c^* is an ideal consequence value.

Agents operating in this way may end up being too predictable. We may mitigate such effect by choosing the next action in a randomized way, with probabilities proportional to the predictive expected utilities, that is

$$P(a_t) \propto \psi(a_t), \tag{2}$$

where $P(a_t)$ is the probability of choosing a_t.

4 Implementation

The above procedures have been implemented within the AISoy1 robot environment (http://www.aisoy.es). Some of the details of the model implemented are described next, with code developed in C++ over Linux.

The set $\mathcal{A}$ includes the robot's actions which include *cry, tell a joke, ask for being recharged, complain, talk, sing, argue, ask for playing, ask for help* and *do nothing*. On the user's side, set $\mathcal{B}$, the robot is able to detect several agent's actions, some of them in a probabilistic way. Among them, the robot detects *hug, hit, shout, speak, being recharged, play, being touched, stroke* or *no action*. Regarding the environment (set $\mathcal{E}$), the bot may recognize contextual issues concerning the presence of noise or music, the level of darkness, the temperature, or its inclination. To do so, the bot has several sensors including a camera to detect objects or persons within a scene, as well as the light intensity; a microphone used to recognize when the user talks and understand what he says, through a natural language processing component; some touch sensors, to interpret when it has been touched or hugged or hit; an inclination sensor so as to know when it is lying down or not; and a temperature sensor. The information provided by these sensors is used by the bot to infer the user's actions and environmental states. Some are based on simple deterministic rules; for example, the hit action is interpreted through a detection in a touch sensor and a variation in the inclination sensor. Others are based on probabilistic rules, like those involving voice recognition and processing.

The basic forecasting models (environment, user, classical conditioning) are Markov chains and we learn about their transition probabilities based on matrix beta models, see [20]. For expected utility computations and point forecasts we summarize the corresponding row-wise Dirichlet distributions through their means. Learning about probability models is done through Bayesian model averaging, as in [9].

The bot aims at satisfying four objectives, which, as in [14], are ordered in hierarchical order of importance. They go from a primary security objective, in which the bot cares mainly for its survival and security (in terms of not being hit, having a sufficient energy level, and being at the right temperature), to higher objective levels in relation with a social empathy layer, once basic objectives are sufficiently covered, in relation with social interests of the bot. Weights reflect the importance of the objectives and component utility functions reflect the aim of fulfilling as quickly as possible the primary objectives, up to a certain level. Emotion implementations are based on [7], who compare different appraisal models to obtain the intensity of an emotion. We use four basic emotions (joy, sadness, hope and fear), which are then combined to obtain more complex emotions. Emotions evolve as Dynamic Linear Models, as in [23].

The model is implemented in a synchronous mode. Sensors are read at fixed times (with different timings for various sensors). When relevant events are detected, the basic information processing and decision making loop is shot. However, as mentioned it is managed by exception in that if exceptions to standard behavior occur, the loop is open to interventions through various threads. We plan only one step ahead and choose the action with probabilities proportional to the computed expected utilities. Memory is limited to the two previous instants.

5 Discussion

We have described a model to control the behavior of an agent in front of an intelligent adversary. It is multi-attribute decision analytic at its core, but it incorporates forecasting models of the adversary (Adversarial Risk Analysis) and emotion-based behavior (Affective Decision Making). This was motivated by our interest in improving the user's experience interacting with a bot [2], [12] and [15]. We believe though that this model may find many other potential applications in fields like interface design, e-learning, entertainment or therapeutical devices.

The model should be extended to a case in which the agent interacts with several users, through a process of identification. It could also be extended to a case in which there are several agents, possibly cooperating or competing, depending on their emotional state. Dealing with the possibility of learning about new user actions, based on repeated readings, and, consequently, augmenting the set $\mathcal{B}$ is another challenging problem. Finally, we have shown what is termed a 0-level ARA analysis. We could try to undertake higher ARA levels in modeling the performance of adversaries.

Acknowledgments

Research supported by grants from the MICINN projects eCOLABORA and RIESGOS, the RIESGOS-CM project and the INNPACTO project HAUS. We are grateful to discussion with Diego Garcia, from AISoy Robotics, Jesus Rios and David Banks.

References

[1] Banks, D., Petralia, F., Wang, S. (2011), Adversarial risk analysis: Borel games. *Applied Stochastic Models in Business and Industry*, **27**: 72-86.

[2] Breazeal, C. (2002) *Designing Sociable Robots.* The MIT Press.

[3] Clemen, R.T., Reilly, T. (2004) *Making Hard Decisions with Decision Tools.* Duxbury: Pacific Grove, CA.

[4] Damasio, A. R. (1994) *Descartes' Error: Emotion, Reason, and the Human Brain.* New York: G.P. Putnam.

[5] El-Nasr, M.S., Yen, J., Ioerger, T.R. (2000) FLAME: Fuzzy Logic Adaptive Model of Emotions. *Autonomous Agents and Multi-Agent Systems* **3**(3):219-257.

[6] Glimcher, P. W., Camerer, C., Poldrack, R. A., Fehr, E. (2008). *Neuroeconomics: Decision Making and the Brain*, Academic Press.

[7] Gratch, J., Marsella, S., Wang, N., Stankovic, B. (2009) Assessing the validity of appraisal-based models of emotion. In Pantic, M., Nijholt, A., Cohn, J. (eds.), *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*. Amsterdam, Netherlands, ACII'09: IEEE Computer Society Press.

[8] Harsanyi, J. (1967) Games with incomplete information played by Bayesian players, I-III. Part I. The basic model. *Management Science*, 14, 3, 159-182.

[9] Hoeting, J., Madigan, D., Raftery, A., Volinsky, C. (1999) Bayesian model averaging: A tutorial, *Statistical Science*, **4**, 382-417.

[10] Kadane, J. B. (2011). Adversarial Risk Analysis: What's new, what isn't?: Discussion of Adversarial Risk Analysis: Borel Games. *Journal Applied Stochastic Models in Business and Industry*, **27**, 2 (March 2011), 87-88.

[11] Kadane, J. B., Larkey, P. D. (1982) Subjective probability and the theory of games. *Management Sci* **28**(2):113-120.

[12] Kirby, R., Forlizzi, J., Simmons, R. (2010) Affective social robots. *Robotics and Autonomous Systems* **58** 3:322-332.

[13] Loewenstein, G. (2000). Preferences, Behavior, and Welfare - Emotions in Economic Theory and Economic Behavior. *AEA Papers and Proceedings*, **90**(2), 426-32.

[14] Maslow, A. H. (1943) A theory of human motivation. *Psychological Review*, **50**, 4, 370-96.

[15] Parisi, D., Petrosino, G. (2010) Robots that have emotions. *Adaptive Behavior* **18**(6):453-469.

[16] Picard, R. W. (1997) *Affective Computing.* Cambridge, MA: MIT Press

[17] Raiffa, H. (2007) *Negotiation Analysis: The Science and Art of Collaborative Decision Making.* Cambridge, Massachusetts:Belknap Press of Harvard University Press

[18] Ríos, J., Rios Insua, D. (2012) Adversarial Risk Analysis: Applications to Counterterrorism Modeling. *Risk Analysis*, (to appear).

[19] Ríos Insua, D., Ríos, J., Banks, D. (2009) Adversarial risk analysis. *Journal of the American Statistical Association* **104**(486):841-854.

[20] Ríos Insua, D., Ruggeri, F., Wiper, M. (2012) *Bayesian Analysis of Stochastic Process Models*, Wiley.

[21] Ríos Insua, D., Salewicz, K. (1995) The operation of Kariba Lake: a multiobjective decision analysis. *Journal of Multicriteria Decision Analysis*, 1995, 4, 203-222.

[22] Velásquez, J. D. (1997) Modeling Emotion and Other Motivations in Synthetic Agents. *Proceedings, 14th National Conference on AI*, AAAI Press.

[23] West, M., Harrison, P. J. (1997) *Bayesian Forecasting and Dynamic Models.* New York: Springer.

[24] von Winterfeldt, D., Edwards, W. (1986). *Decision Analysis and Behavioral Research.* New York: Cambridge University Press.

Random belief learning

David Leslie
Department of Mathematics
University of Bristol
University Walk, Clifton, Bristol BS8 1TW
UK
david.leslie@bristol.ac.uk

Abstract

When individuals are learning about an environment and other decision-makers in that environment, a statistically sensible thing to do is form posterior distributions over unknown quantities of interest (such as features of the environment and 'opponent' strategy) then select an action by integrating with respect to these posterior distributions. However reasoning with such distributions is very troublesome, even in a machine learning context with extensive computational resources; Savage himself indicated that Bayesian decision theory is only sensibly used in reasonably "small" situations.

Random beliefs is a framework in which individuals instead respond to a single sample from a posterior distribution. There is evidence from the psychological and animal behaviour disciplines to suggest that both humans and animals may use such a strategy. In our work we demonstrate such behaviour 'solves' the exploration-exploitation dilemma 'better' than other provably convergent strategies. We can also show that such behaviour results in convergence to a Nash equilibrium of an unknown game.

Bayesian Combination of Multiple, Imperfect Classifiers

Edwin Simpson, Stephen Roberts, Ioannis Psorakis
Department of Engineering Science, University of Oxford, UK.
Arfon Smith , **Chris Lintott**
Department of Physics, University of Oxford, UK.

Abstract

Classifier combination methods need to make best use of the outputs of multiple, imperfect classifiers to enable higher accuracy classifications. In many situations, such as when human decisions need to be combined, the base decisions can vary enormously in reliability. A Bayesian approach to such uncertain combination allows us to infer the differences in performance between individuals and to incorporate any available prior knowledge about their abilities when training data is sparse. In this paper we explore Bayesian classifier combination, using the computationally efficient framework of variational Bayesian inference. We apply the approach to real data from a large citizen science project, Galaxy Zoo Supernovae, and show that our method far outperforms other established approaches to imperfect decision combination. We go on to analyse the putative community structure of the decision makers, based on their inferred decision making strategies, and show that natural groupings are formed.

1 Introduction

In many real-world scenarios we are faced with the need to aggregate information from cohorts of imperfect decision making agents (*base classifiers*), be they computational or human. Particularly in the case of human agents, we rarely have available to us an indication of how decisions were arrived at or a realistic measure of agent confidence in the various decisions. Fusing multiple sources of information in the presence of uncertainty is optimally achieved using Bayesian inference, which elegantly provides a principled mathematical framework for such knowledge aggregation. In this paper we provide a Bayesian framework for such imperfect decision combination, where the base classifications we receive are greedy preferences (i.e. labels with no indication of confidence or uncertainty). The classifier combination method we develop aggregates the decisions of multiple agents, improving overall performance. We present a principled framework in which the use of weak decision makers can be mitagated and in which multiple agents, with very different observations, knowledge or training sets, can be combined to provide complementary information. The preliminary application we focus on in this paper is a distributed *citizen science* project, in which human agents carry out classification tasks, in this case identifying transient objects from images as corresponding to potential supernovae or not. This application, *Galaxy Zoo Supernovae* [1], is part of the highly successful *Zooniverse* family of citizen science projects. In this application the ability of our base classifiers can be very varied and there is no guarantee over any individual's performance, as each user can have radically different levels of domain experience and have different background knowledge. As individual users are not overloaded with decision requests by the system, we often have little performance data for individual users (base classifiers). The methodology we advocate provides a scaleable, computationally efficient, Bayesian approach to learning base classifier performance thus enabling optimal decision combinations. The approach is robust in the presence of uncertainties at all levels and naturally handles missing observations, i.e. in cases where agents do not provide any base classifications.

1.1 Independent Bayesian Classifier Combination

Here we present a variant of Independent Bayesian Classifier Combination (IBCC), originally defined in [2]. The model assumes conditional independence between base classifiers, but performed as well as more computationally intense dependency modelling methods [2]. For the ith data point, we assume that the true label t_i is generated from a multinomial distribution with probability $\boldsymbol{\kappa}$: $p(t_i = j|\boldsymbol{\kappa}) = \kappa_j$. We assume that observed classifier outputs, $\boldsymbol{c}$, are discrete and are generated from a multinomial distribution dependent on the class of the true label, with parameters $\boldsymbol{\pi}$: $p(c_i^{(k)}|t_i = j, \boldsymbol{\pi}) = \pi^{(k)}_{jc_i^{(k)}}$. Thus there are minimal requirements on the type of base classifier output, which need not be probabilistic and could be selected from an arbitrary number of discrete values, indicating, for example, greedy preference over a set of class labels. The parameters $\boldsymbol{\pi}$ and $\boldsymbol{\kappa}$ have Dirichlet prior distributions with hyper-parameters $\boldsymbol{\alpha_j^{(k)}} = [\alpha_{0j,1}^{(k)}, ..., \alpha_{0j,L}^{(k)}]$ and $\boldsymbol{\nu} = [\nu_{01}, ...\nu_{0J}]$ respectively, where L is the number of possible outputs from classifier k and J is the number of classes. The joint distribution over all variables is

$$p(\boldsymbol{\kappa}, \boldsymbol{\pi}, \boldsymbol{t}, \boldsymbol{c}|\boldsymbol{\alpha}, \boldsymbol{\nu}) = \prod_{i=1}^{N}\{\kappa_{t_i}\prod_{k=1}^{K}\pi_{t_i,c_i^{(k)}}\}p(\boldsymbol{\kappa}|\boldsymbol{v})p(\boldsymbol{\pi}|\boldsymbol{\alpha}). \tag{1}$$

The graphical model for IBCC is shown in figure 1.

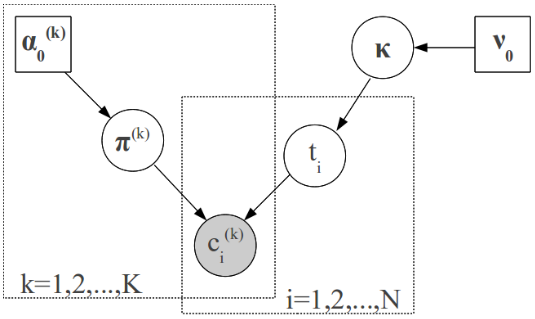


Figure 1: Graphical Model for IBCC. Shaded nodes are observed values, circular nodes are variables with a distribution and square nodes are variables instantiated with point values.

A key feature of IBCC is that $\boldsymbol{\pi}$ represents a *confusion matrix* that quantifies the decision-making abilities of *each* base classifier. This potentially allows us to ignore, or retrain, poorer classifiers and assign experts decision makers to data points that are highly uncertain. Such efficient selection of base classifiers is vitally important when obtaining a classification that has a cost related to the number of decision makers, for example. The IBCC model also allows us to infer values for missing observations of classifier outputs, $\boldsymbol{c}$, so that we can naturally handle cases in which only *partially observed* agents make decisions.

The IBCC model assumes independence between the rows in $\boldsymbol{\pi}$, i.e. the probability of each classifier's outputs is dependent on the *true label class*. In some cases it may be reasonable to assume that performance over one label class may be correlated with performance in another; indeed methods such as *weighted majority* [3] make this tacit assumption. However, we would argue that this is not universally the case, and IBCC makes no such strong assumptions.

The goal of the combination model is to perform inference for the unknown variables $\boldsymbol{t}$, $\boldsymbol{\pi}$, and $\boldsymbol{\kappa}$. The inference technique proposed in [2] was Gibbs Sampling. While this provides some theoretical guarantee of accuracy given the proposed model, it is often very slow to converge and convergence is difficult to ascertain. In this paper we consider the use of a principled approximate Bayesian methods, namely *variational Bayes* (VB) [4] as this allows us to replace non-analytic marginal integrals in the original model with analytic updates in the *sufficient statistics* of the variational approximation. This produces a model that iterates rapidly to a solution in a computational framework which can be seen as a Bayesian generalisation of the *Expectation-Maximization* EM algorithm.

In [2] an exponential prior distribution is placed over $\boldsymbol{\alpha_0}$. However, exponentials are not conjugate to the Dirichlet, and the conjugate prior to the Dirichlet is non-standard and its normalisation constant is not in closed form [5], requiring the use of an expensive adaptive rejection Gibbs sampling step for $\boldsymbol{\alpha}$ and making even the variational Bayesian solution intractable. We therefore alter the model, so as to using point values for $\boldsymbol{\alpha_0}$, as are used in other VB models [6, 7, 8]. The hyper-parameter values of $\boldsymbol{\alpha_0}$ can hence be chosen to represent any prior level of uncertainty in the values of the agent-by-agent confusion matrices, $\boldsymbol{\pi}$, and can be regarded as pseudo-counts of prior observations, offering a natural method to include any prior knowledge and a methodology to extend the method to sequential, on-line environments.

1.2 Variational Bayes

Given a set of observed data $\boldsymbol{X}$ and a set of latent variables and parameters $\boldsymbol{Z}$, the goal of variational Bayes (VB) is to find a tractable approximation $q(\boldsymbol{Z})$ to the posterior distribution $p(\boldsymbol{Z}|\boldsymbol{X})$ by minimising the KL-divergence between the approximate distribution and the true distribution. We can write the log of the model evidence $p(\boldsymbol{X})$ as

$$\begin{aligned} \ln p(\boldsymbol{X}) &= \int q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z})}{q(\boldsymbol{Z})} \mathrm{d}\boldsymbol{Z} - \int q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{Z}|\boldsymbol{X})}{q(\boldsymbol{Z})} \mathrm{d}\boldsymbol{Z} \qquad (2) \\ &= L(q) - \mathrm{KL}(q||p). \qquad (3) \end{aligned}$$

As $q(\boldsymbol{Z})$ approaches $p(\boldsymbol{Z}|\boldsymbol{X})$, the KL-divergence disappears and the lower bound $L(q)$ is maximised. Variational Bayes selects a restricted form of $q(\boldsymbol{Z})$ that is tractable to work with, then seeks the distribution within this restricted form that minimises the KL-divergence. A common restriction is to assume $q(\boldsymbol{Z})$ factorises into single variable factors $q(\boldsymbol{Z}) = \prod_{i=1}^{M} q_i(\boldsymbol{Z}_i)$. For each factor $q_i(\boldsymbol{Z}_i)$ we then seek the optimal solution $q_i^*(\boldsymbol{Z}_i)$ that minimises the KL-divergence. Mean field theory [9] then shows that the log of each optimal factor $\ln q_i^*(\boldsymbol{Z}_i)$ is the expectation with respect to all other factors of the log of the joint distribution over all hidden and known variables:

$$\ln q_i^*(\boldsymbol{Z}_i) = \mathbb{E}_{i \neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})] + \text{const}. \qquad (4)$$

We can evaluate these optimal factors iteratively by first initialising all factors, then updating each in turn using the expectations with respect to the current values of the other factors. Unlike Gibbs sampling, the each iteration is guaranteed to increase the lower bound on the log-likelihood, $L(q)$, converging to a (local) maximum in a similar fashion to standard EM algorithms. If the factors $q_i^*(\boldsymbol{Z}_i)$ are exponential family distributions, as is the case for the IBCC method we present in the next section, the lower bound is convex with respect to each factor $q_i^*(\boldsymbol{Z}_i)$ and $L(q)$ will converge to a *global* maximum of our approximate, factorised distribution. In practice, once the optimal factors $q_i^*(\boldsymbol{Z}_i)$ have converged to within a given tolerance, we can approximate the distribution of the unknown variables and calculate their expected values.

2 Variational Bayesian IBCC

To provide a variational Bayesian treatment of IBCC, VB-IBCC, we first propose the form for our variational distribution ($q(\boldsymbol{Z})$ in the previous section) that factorises between the parameters and latent variables.

$$q(\boldsymbol{\kappa}, \boldsymbol{t}, \boldsymbol{\pi}) \quad = \quad q(\boldsymbol{t}) q(\boldsymbol{\kappa}, \boldsymbol{\pi}) \qquad (5)$$

This is the only assumption we must make to perform VB on this model; the forms of the factors arise from our model of IBCC. We can use the joint distribution in equation 1 to find the optimal factors $q^*(\boldsymbol{t})$ and $q^*(\boldsymbol{\kappa}, \boldsymbol{\pi})$ it in the form given by equation 4. For the target labels we have

$$\ln q^*(\boldsymbol{t}) \quad = \quad \mathbb{E}_{\boldsymbol{\kappa}, \boldsymbol{\pi}}[\ln p(\boldsymbol{\kappa}, \boldsymbol{t}, \boldsymbol{\pi}, \boldsymbol{c})] + \text{const}. \qquad (6)$$

We rewrite this into factors corresponding to independent data points, with any terms not involving t_i being absorbed into the normalisation constant.

$$\ln q^*(t_i) = \mathbb{E}_{\boldsymbol{\kappa}}[\ln \kappa_{t_i}] + \sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{\pi}}[\ln \pi^{(k)}_{t_i, c_i^{(k)}}] + \text{const} \qquad (7)$$

To simplify the optimal factors in subsequent equations, we define expectations with respect to $\boldsymbol{t}$ of two statistics: the number of occurrences of each target class is given by

$$N_j = \sum_{i=1}^{N} \mathbb{E}_{\boldsymbol{t}}[t_i = j] = \sum_{i=1}^{N} q^*(t_i = j) \tag{8}$$

and the counts of each classifier decision, $c_i^{(k)} = l$, given the target label, $t_i = j$, given by

$$N_{jl}^{(k)} = \sum_{i=1}^{N} [c_i^{(k)} = l] \mathbb{E}_{\boldsymbol{t}}[t_i = j] = \sum_{i=1}^{N} [c_i^{(k)} = l] q^*(t_i = j). \tag{9}$$

where $[c_i^{(k)} = l]$ is unity if $c_i^{(k)} = l$ and zero otherwise.

For the parameters of the model we have the optimal factors given by:

$$\begin{align} \ln q^*(\boldsymbol{\kappa}, \boldsymbol{\pi}) &= \mathbb{E}_{\boldsymbol{t}}[\ln p(\boldsymbol{\kappa}, \boldsymbol{t}, \boldsymbol{\pi}, \boldsymbol{c})] + \text{const} \tag{10}\\ &= \mathbb{E}_{\boldsymbol{t}}[\sum_{i=1}^{N} \{\ln p_{t_i} + \sum_{k=1}^{K} \ln \pi_{t_i, c_i^{(k)}}^{(k)}\}] + \ln p(\boldsymbol{\kappa} | \boldsymbol{v_0}) \tag{11}\\ &\quad + \ln p(\boldsymbol{\pi} | \boldsymbol{\alpha}) + \text{const.} \tag{12} \end{align}$$

In equation 10 terms involving $\boldsymbol{\kappa}$ and terms involving each confusion matrix in $\boldsymbol{\pi}$ are separate, so we can factorise $q^*(\boldsymbol{\kappa}, \boldsymbol{\pi})$ further into

$$q^*(\boldsymbol{\kappa}, \boldsymbol{\pi}) = q^*(\boldsymbol{\kappa}) \prod_{k=1}^{K} \prod_{j=1}^{J} q^*(\boldsymbol{\pi}_j^{(k)}). \tag{13}$$

Considering the prior for $\boldsymbol{\kappa}$ is a Dirichlet distribution, we obtain the optimal factor

$$\begin{align} \ln q^*(\boldsymbol{\kappa}) &= \mathbb{E}_{\boldsymbol{t}}[\sum_{i=1}^{N} \ln \kappa_{t_i}] + \ln p(\boldsymbol{\kappa} | \boldsymbol{v}) + \text{const} \tag{14}\\ &= \sum_{j=1}^{J} N_j \ln \kappa_j + \sum_{j=1}^{J} (\nu_{0,j} - 1) \ln \kappa_j + \text{const.} \tag{15} \end{align}$$

Taking the exponential of both sides, we obtain a posterior Dirichlet distribution of the form

$$q^*(\boldsymbol{\kappa}) \propto \text{Dir}(\boldsymbol{\kappa} | \boldsymbol{\nu}_1, ..., \boldsymbol{\nu}_J) \tag{16}$$

where $\boldsymbol{\nu}$ is updated in the standard manner by adding the data counts to the prior counts ν_0:

$$\nu_j = \nu_{0,j} + N_j. \tag{17}$$

The expectation of $\ln \boldsymbol{\kappa}$ required to update equation 7 is therefore:

$$\mathbb{E}[\ln \kappa_j] = \Psi(\nu_j) - \Psi(\sum_{j'=1}^{J} \nu_{j'}) \tag{18}$$

where $\Psi(z)$ is the standard digamma function.

For the confusion matrices $\boldsymbol{\pi}_j^{(k)}$ the priors are also Dirichlet distributions giving us the factor

$$\begin{align} \ln q^*(\boldsymbol{\pi}_j^{(k)}, \boldsymbol{\alpha}_j^{(k)}) &= \sum_{i=1}^{N} \mathbb{E}_{t_i}[t_i = j] \ln \pi_{j, c_i^{(k)}}^{(k)} + \ln p(\boldsymbol{\pi} | \boldsymbol{\alpha}) + \text{const} \tag{19}\\ &= \sum_{l=1}^{L} N_{jl}^{(k)} \ln \pi_{jl}^{(k)} + \sum_{l=1}^{L} (\alpha_{jl}^{(k)} - 1) \ln \pi_{jl}^{(k)} + \text{const.} \tag{20} \end{align}$$

Again, taking the exponential gives a posterior Dirichlet distribution of the form

$$q^*(\boldsymbol{\pi}_j^{(k)}) \quad \propto \quad \text{Dir}(\boldsymbol{\pi}_j^{(k)} | \alpha_{j1}^{(k)}, ..., \alpha_{jL}^{(k)}) \tag{21}$$

where $\boldsymbol{\alpha}_j^{(k)}$ is updated by adding data counts to prior counts $\alpha_{0,j}^{(k)}$:

$$\alpha_{jl}^{(k)} = \alpha_{0,jl}^{(k)} + N_{jl}^{(k)}. \tag{22}$$

The expectation required for equation 7 is given by

$$\mathbb{E}[\ln \pi_{jl}^{(k)}] = \Psi(\alpha_{jl}^{(k)}) - \Psi(\sum_{l'=1}^{L} \alpha_{jl'}^{(k)}). \tag{23}$$

To apply the VB algorithm to IBCC, we initialise all the expectations over $\mathbb{E}[\ln \pi_{jl}^{(k)}]$ and $\mathbb{E}[\ln \kappa_j]$, either randomly or by choosing their prior expectations (if we have domain knowledge to inform this). We then iterate over a two-stage procedure similar to the *Expectation-Maximization* (EM) algorithm. In the variational equivalent of the *E-step* we use the current expected parameters, $\mathbb{E}[\ln \pi_{jl}^{(k)}]$ and $\mathbb{E}[\ln \kappa_j]$, to update the variational distribution in equation 5. First we evaluate equation 7, then use the result to update the counts N_j and $N_{jl}^{(k)}$ according to equations 8 and 9. In the variational *M-step*, we update $\mathbb{E}[\ln \pi_{jl}^{(k)}]$ and $\mathbb{E}[\ln \kappa_j]$ using equations 18 and 23.

3 Galaxy Zoo Supernovae

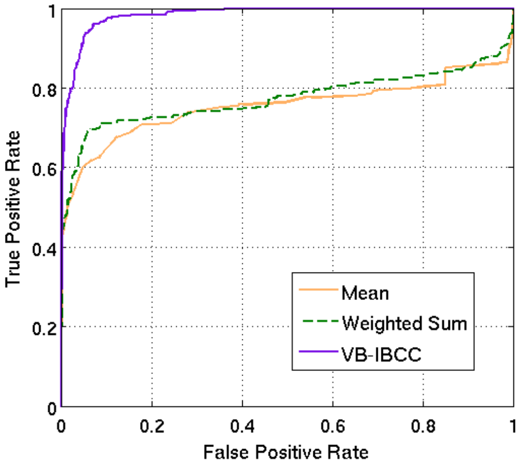
We tested the model using a dataset obtained from the Galaxy Zoo Supernovae citizen science project [1]. The dataset contains scores given by individual volunteer citizen scientists (base classifiers) to candidate supernova images after answering a series of questions, the aim being to classify each data sample (images) as either “supernova” or “not supernova”. A set of three linked questions are answered by the users, which are hard-coded in the project repository to scores of -1, 1 or 3, corresponding respectively to decisions that the data point is very unlikely to be a supernova, possibly a supernova and very likely a supernova.

In order to verfiy the efficacy of our approach and competing methods, we use “true” target classifications obtained from full spectroscopic analysis, undertaken as part of the Palomar Transient Factory collaboration [10]. We note that this information, is not available to the base classifiers (the users), being obtained retrospectively. This labelling is not made use of in our algorithms, save for the purpose of measuring performance. We compare IBCC using both variational Bayes (VB-IBCC) and Gibbs sampling (Gibbs-IBCC), using as output the expected values of t_i. We also tested simple majority voting, weighted majority voting & weighted sum [3] and mean user scores, which the Galaxy Zoo Supernovae currently uses to filter results. For majority voting methods we treat both 1 and 3 as a vote for the supernova class.

The complete dataset contains many volunteers that have provided very few classifications, particularly for positive examples, as there are 322 classifications of positive data points compared to 43941 “not supernova” examples. We therefore subsampled the dataset, selecting all positive data points, then selecting only negative data points that have at least 10 classifications from volunteers who have classified at least 50 examples, which produced a data set of some 1000 examples with decisions produced from around 1700 users. We tested all imperfect decision combination methods using five-fold cross validation.

Figure 2a shows the average *Receiver-Operating Characteristic* (ROC) curves taken across all cross-validation datasets for the mean score, weighted sum and VB-IBCC. The ROC curve for VB-IBCC clearly outperforms the mean of scores by a large margin. Weighted sum achieves a slight improvement on the mean by learning to discount base classifiers each time they make a mistake. The performance of the majority voting methods and IBCC using Gibbs sampling is summarised by the area under the ROC curve (AUC) in table 2b. Majority voting methods only produce one point on the ROC curve between 0 and 1 as they convert the scores to votes (-1 becomes a negative vote, 1 and 3 become positive) and produce binary outputs. These methods have similar results to the mean score approach, with the weighted version performing slightly worse, perhaps because too much information is lost when converting scores to votes to be able to learn base classifier weights correctly.

With Gibbs-sampling IBCC we collected samples until the mean of the sample label values converged. Convergence was assumed when the total absolute difference between mean sample labels



(a) Average Reciever operating characteristic (ROC) curves.

Method	AUC
Mean of Scores	0.7543
Weighted Sum	0.7722
Simple Majority Voting	0.7809
Weighted Majority Voting	0.7378
Gibbs-IBCC	0.9127
VB-IBCC	0.9840

(b) Area under the ROC curves (AUCs).

Figure 2: Galaxy Zoo Supernovae: ROC curves and AUCs with 5-fold cross validation.

of successive iterations did not exceed 0.01 for 20 iterations. The mean time taken to run VB-IBCC to convergence was 13 seconds, while for Gibbs sampling IBCC it was 349 seconds. As well as executing significantly faster, VB produces a better AUC than Gibbs sampling with this dataset.

4 Communities of decision makers

In this section we apply a recent community detection methodology to the problem of determining most likely groupings of base classifiers, the imperfect decision makers. Identifying overlapping communities in networks is a challenging task. In recent work [11] we have presented a novel approach to community detection that utilises a Bayesian factorization model to extract *overlapping* communities from a "similarity" or "interaction" network. The scheme has the advantage of soft-partitioning solutions, assignment of node participation scores to communities, an intuitive foundation and computational efficiency. We apply this approach to a similarity matrix calculated over all the citizen scientists in our study, based upon each users' confusion matrix. Denoting π_i as the (3×2) confusion matrix inferred for user i we may define a simple similarity measure between agents i and j as

$$V_{i,j} = \exp\left(-\mathcal{HD}(\pi_i, \pi_j)\right), \tag{24}$$

where $\mathcal{HD}()$ is the *Hellinger distance* between two distributions, meaning that two agents who have very similar confusion matrices will have high similarity.

Application of Bayesian community detection to the matrix $\mathbf{V}$ robustly gave rise to *five* distinct groupings of users. In figure 3 we show the centroid confusion matrices associated with each of these groups of citizen scientists. The labels indicate the "true" class (0 or 1) and the preference for the three scores offered to each user by the Zooniverse questions (-1, 1 & 3). Group 1, for example, indicates users who are clear in their categorisation of "not supernova" (a score of -1) but who are less certain regarding the "possible supernova" and "likely supernova" categories (scores 1 & 3). Group 2 are "extremists" who use little of the middle score, but who confidently (and correctly) use scores of -1 and 3. By contrast group 3 are users who almost always use score -1 ("not supernova") whatever objects they are presented with. Group 4 almost never declare an object as "not supernova" (incorrectly) and, finally, group 5 consists of "non-commital" users who rarely assign a score of 3 to supernova objects, preferring to stick with the middle score ("possible supernova"). It is interesting to note that all five groups have similar numbers of members (several hundred) but clearly each group indicates a very different approach to decision making.

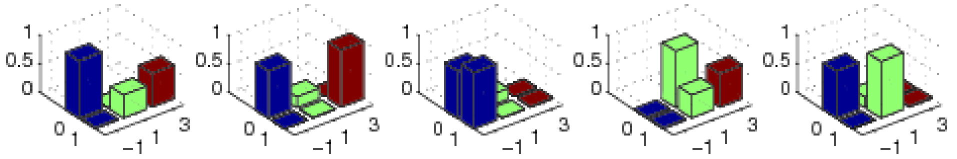


Figure 3: Prototypical confusion matrices for each of the five communities inferred using Bayesian social network analysis (see text for details).

5 Discussion

We present in this paper a very computationally efficient, variational Bayesian, approach to imperpect multiple classifier combination. We evaluated the method using real data from the Galaxy Zoo Supernovae citizen science project, with 963 data points and 1705 base classifiers. In our experiments, our method outperformed all other methods, including weighted sum and weighted majority, both of which are often advocated as they also learn weightings for the base classifiers. For our variational Bayes method the required computational overheads were far lower than those of Gibbs sampling approaches, thus giving much shorter compute time, which is particularly important for applications that need to make regular updates as new data is observed, such as our application here. Furthermore, on this data set at least, the performance was also better than the slower sample based method. We have shown that a sensible structure emerges from the cohort of decision makers via social network analysis and this provides valuable information regarding the decision-making of the groups' members.

Our current work considers how the rich information learned using this method can be exploited to improve the base classifiers, namely the human volunteer users. For example, we can use the confusion matrices, $\boldsymbol{\pi}$, to identify users groups who would benefit from more training, potentially from interaction with user groups who perform more accurate decision making (via extensions of *apprenticeship learning*, for example). We also consider, via selective object presentation, ways of producing user specialisation such that the overall performance of the human-agent collective is maximised. We note that this latter concept bears the hallmark traces of *computational mechanism design* and the incorporation of incentives engineering and coordination mechanisms into the model is one of our present challenges.

References

[1] A. M. Smith, S. Lynn, M. Sullivan, C. J. Lintott, P. E. Nugent, J. Botyanszki, M. Kasliwal, R. Quimby, S. P. Bamford, L. F. Fortson15, et al. Galaxy Zoo Supernovae. 2010.

[2] Z. Ghahramani and H. C. Kim. Bayesian classifier combination. *Gatsby Computational Neuroscience Unit Technical Report No. GCNU-T., London, UK:*, 2003.

[3] N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and computation*, 108(2):212–261, 1994.

[4] Hagai Attias. *Advances in Neural Information ... - Google Books*, chapter A Variational Bayesian Framework for Graphical Models, pages 209–215. 2000.

[5] S. Lefkimmiatis, P. Maragos, and G. Papandreou. Bayesian inference on multiscale models for poisson intensity estimation: Applications to Photon-Limited image denoising. *Image Processing, IEEE Transactions on*, 18(8):1724–1741, 2009.

[6] R. Choudrey and S. Roberts. Variational Mixture of Bayesian Independent Component Analysers. *Neural Computation*, 15(1), 2003.

[7] C. M. Bishop. *Pattern recognition and machine learning*. Information Science and Statistics. Springer Science+Business Media, LLC, 4 edition, 2006.

[8] W. D Penny and S. J Roberts. Dynamic logistic regression. In *Neural Networks, 1999. IJCNN'99. International Joint Conference on*, volume 3, pages 1562–1567, 1999.

[9] G. Parisi and R. Shankar. Statistical field theory. *Physics Today*, 41:110, 1988.

[10] N. M. Law, S. R. Kulkarni, R. G. Dekany, E. O. Ofek, R. M. Quimby, P. E. Nugent, J. Surace, C. C. Grillmair, J. S. Bloom, M. M. Kasliwal, et al. The palomar transient factory: System overview, performance, and first results. *Publications of the Astronomical Society of the Pacific*, 121(886):1395–1408, 2009.

[11] I Psorakis, S. Roberts, M. Ebden, and B. Shelden. Overlapping Community Detection using Bayesian Nonnegative Matrix Factorization. *Physical Review E*, 83, 2011.

Artificial Intelligence Design
for Real-time Strategy Games

Firas Safadi
University of Liège
fsafadi@ulg.ac.be

Raphael Fonteneau
University of Liège
raphael.fonteneau@ulg.ac.be

Damien Ernst
University of Liège
dernst@ulg.ac.be

Abstract

For now over a decade, real-time strategy (RTS) games have been challenging intelligence, human and artificial (AI) alike, as one of the top genre in terms of overall complexity. RTS is a prime example problem featuring multiple interacting imperfect decision makers. Elaborate dynamics, partial observability, as well as a rapidly diverging action space render rational decision making somehow elusive. Humans deal with the complexity using several abstraction layers, taking decisions on different abstract levels. Current agents, on the other hand, remain largely scripted and exhibit static behavior, leaving them extremely vulnerable to flaw abuse and no match against human players. In this paper, we propose to mimic the abstraction mechanisms used by human players for designing AI for RTS games. A non-learning agent for StarCraft showing promising performance is proposed, and several research directions towards the integration of learning mechanisms are discussed at the end of the paper.

1 Introduction

The real-time strategy (RTS) video game genre began to appear roughly two decades ago. Published in 1992, Dune II (Westwood Studios)[1] already featured the core concepts of RTS. However, the genre only started getting popular a few years later with the release of titles such as Warcraft (Blizzard Entertainment)[1] in 1994 or Command & Conquer (Westwood Studios)[1] in 1995. Other titles followed, each bringing new additions to the variety and, in 1998, StarCraft (Blizzard Entertainment)[1], one of the best-selling video games and now an acknowledged reference, cemented RTS in the industry. Part of the appeal of RTS games comes from their complexity and the control difficulty resulting from the intensive multitasking they require. In StarCraft tournaments for example, professional players routinely exceed 200 actions per minute. While multitasking poses no issues to computers, they are confronted with the intractable problem of learning in such convoluted environments. The large number of units to control as well as the diverse tasks to complete, coupled with partial observability, result in complex game dynamics.

Besides partial observability, the main cause for imperfect decision making in RTS games has to do with the way the complexity is managed, which is abstraction. Abstraction is an example mechanism used to summarize large quantities of information into a more compact and manageable, albeit abstract form. While it is an amazingly useful and powerful tool, there is at least one significant downside to using it: loss of information. By successively synthesizing numerous basic elements into abstract form, knotty details are eventually overlooked. Thus, abstraction could be seen as an ultimately lossy compression process. Because the quantity of information available at any moment and throughout the entire game is too large, part of it is inevitably discarded in the process of analyzing and rationalizing the game state.

[1] Westwood Studios and Blizzard Entertainment as well as Dune II: The Building of a Dynasty, Command & Conquer, Warcraft and StarCraft are registered trademarks.

Within an ideal framework, agents should not lose any information when analyzing the game state as this loss is only associated to the abstraction mechanisms used by human beings. In practice, this is not the case: agents also discard information because we are not capable of programming them to consider every information. Although this may seem counter-intuitive, the same flaw we suffer from is reintroduced in agents for the simple fact that we do not possess the technology to handle the game in raw form. Part of the objective of this work is thus to propose to the RTS community a simple and generic agent model based on abstract concepts to help efficiently build agents for different RTS games. In this document, we explore the RTS context and propose a modular and hierarchical agent design inspired by abstraction mechanisms used by human players with the aim of subsequently adding learning features. We then provide some promising experimental results in the particular case of StarCraft, and discuss some research directions opened by this work.

The following of the paper is structured as follows. Section 2 quickly covers some related work. Section 3 presents RTS games in detail. In Section 4, we provide an efficient modular and hierarchical agent design. Finally, we conclude and briefly discuss some future works in section 5.

2 Related Work

During the last decade, the scientific community has acknowledged that RTS games constitute rich environments for AI researchers to evaluate different AI techniques. Development frameworks for RTS agents such as the ORTS (Open RTS) project (Buro and Furtak, 2004) appeared and research work started to tackle some of the challenges offered by the RTS genre. Due to the inherent difficulty of designing good RTS agents able to address the multitude of problems they are confronted to, most work has been concerned with specific aspects of the game.

The strategy planning problem is probably the one that has received the most attention with the success of case-based planning methods to identify strategic situations and manage build orders. An approach based on case generation using behavioral knowledge extracted from existing game traces was tested on Wargus[2] (Ontañón et al., 2007). Other approaches for case retrieval and build order selection were tested on the same game (Aha et al., 2005; Weber and Mateas, 2009a). A case retrieval method based on conceptual neighborhoods was also proposed (Weber and Mateas, 2009b). The strategy planning problem has also been addressed with other techniques such as data mining (Weber and Mateas, 2009c). By analyzing a large collection of game logs, it is possible to extract building trends and timings which are then used with matching algorithms to identify a strategy or even predict strategic decisions. Evolutionary methods have been employed as well, mostly in strategy generation (Ponsen et al., 2006). Meanwhile, some work has focused on lower-level problems like micro-management. Monte Carlo planning was among other things applied to simple CTF (“Capture The Flag”) scenarios on the ORTS platform (Chung et al., 2005).

Although the above-mentioned works all showed interesting results, none actually takes on all the aspects of the RTS genre. Other works have instead considered the entire problem and present complete agents. A cognitive approach was tested using the ORTS infrastructure, though it suffered from artificial human multitasking limitations (Wintermute et al., 2007). Another approach was based on an integrated agent composed of distinct managers each responsible for a domain of competence (McCoy and Mateas, 2008). Although the design was clear, it lacked hierarchical structure and unit management was largely simplified.

While interesting, these do not offer a clear and simple design to develop and improve agents for RTS games. The model we suggest in this document is simple, efficient and generic and can potentially be used to add learning capabilities in new agents.

3 Real-time Strategy

In RTS games, players typically confront each other in a map with a unique terrain configuration. They start with a set number of units and must build a force to destroy all opponents. Players do not have access to the entire game state. Only areas where they have deployed units are visible. This is commonly referred to as the fog of war. Different types of units can be built throughout

[2] Wargus is a clone of Warcraft II, a RTS title published by Blizzard Entertainment in 1995.

the game. Each has its own attributes such as hit points, speed, range, whether it can fly, whether it is biological, or any other attribute part of the game mechanics. Each unit type also costs a certain amount of resources and requires some technology. Resources can be gathered from specific locations on the battlefield while technologies can be researched by players to unlock the desired unit types. Besides attributes, units also have special abilities (i.e., activating a shield). These abilities are either innate or need be unlocked. Depending on the game, some units may evolve during their lifespan, either acquiring new abilities or improving their base attributes. Some games also feature an upgrade system, which allows players to increase the performance of certain units (i.e., increase all infantry weapon damage).

Another characteristic of RTS games adding to their diversity and complexity is the concept of race. Players do not necessarily have access to the same units and technologies. Before the game starts, each player may choose a race. Depending on the choice, entirely different, yet balanced, sets of units and technologies can be available to different players. In StarCraft, players can choose among three races: the Terrans who excel at defense and adaptation, the Zerg with their overwhelming swarms and the Protoss, a humanoid species with unmatched individual fighting prowess.

One last, and important, aspect in RTS is diplomacy. Players can decide to form alliances or break them during a game to further their own interests. Extreme complexity may arise from such contexts and, as far as it pertains to this document, we instead focus on the simpler free-for-all setting when discussing agents.

Obviously, players must constantly take a multitude of decisions on different time scales. They must decide what and when units should be built, whether the current income is sufficient or new resource sites should be controlled, when to attack or to defend, when to retreat during an attack or whether a diversion is required, whether some unit should be on the front line, whether a special ability should be used, etc. It is also clear that players need to know what their opponents are planning in order to make good decisions and therefore have to constantly go on reconnaissance. By noting that an order can be sent to each unit at any time in the game and that the number of units a player owns is often larger than one hundred, it comes with no surprise that these games are challenging.

4 Agent Design

We first describe in Subsection 4.1 the concept of abstraction upon which the modular and hierarchical design detailed in Subsection 4.2 is based. We provide experimental results showing the performance of our agent playing StarCraft and discuss some limitations in Subsection 4.3.

4.1 Overcoming Complexity

Abstraction can be seen as the ability to reduce information into sets and structures for addressing a particular purpose. In the context of RTS games, the complexity and the large number of units to control cause human players to instinctively resort to abstraction. Instead of thinking about each unit and its abilities individually, they think about groups of units and more abstract abilities such as defense or harassment. Players thus use abstraction to efficiently control the environment. Also simplified by abstraction are objectives. Typically, the objective in an RTS game is to destroy all enemy units. Since there are many units, humans do not think about defeating each enemy unit. Rather, they move along abstract objectives like defeating an enemy outpost or primary base. Using abstract elements of this kind, we can then understand how RTS can be structured into a more manageable problem.

The modular and hierarchical agent model we next present is structured around the primary tasks we were able to identify. It is composed of different abstract managers generic enough to be used in a wide array of RTS games.

4.2 A Modular and Hierarchical Design

When looking at the tasks players must solve, we were inclined to identify to identify two categories based on task decomposition: production-related tasks and combat-related ones. Production tasks include everything from economy management and expansion to build order management and technology appraisal. On the other hand, combat tasks regroup all combat management elements such as attacking a base or defending an outpost. Furthermore, we divide tasks according to temporal factors. The first group deals with long-term objectives and is referred to as strategy. Decisions at this high level consist of abstract orders like performing a rush[3]. The second one, called tactics, consists of mid-term tasks such as winning a battle. Examples of decisions would be engaging on two fronts or retreating from combat to force the enemy to move to a designated location. The third and last group gathers the remaining short-term objectives. These involve more concrete orders like destroying a specific unit or using a special ability.

As a result, a hierarchical and modular model is proposed in Figure 1. It features a strategy manager at the top level taking all strategic decisions. Directly below come two other managers. A production manager handles construction as well as build orders and controls in turn a number of work squad managers. On the other side, a combat manager takes care of tactical decisions and controls several military squad managers. Information thus travels from the strategy manager and is successively translated into lesser abstract decisions until it reaches the managers at the lowest level which relay direct orders to units on the battlefield. This process is illustrated in Figure 2.

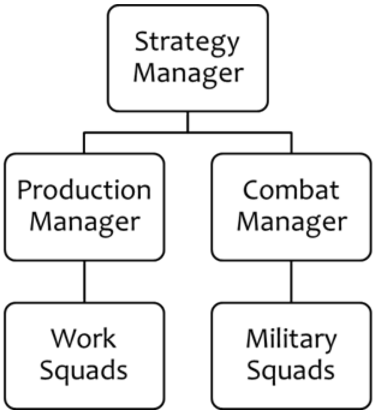


Figure 1: A hierarchical and modular design.

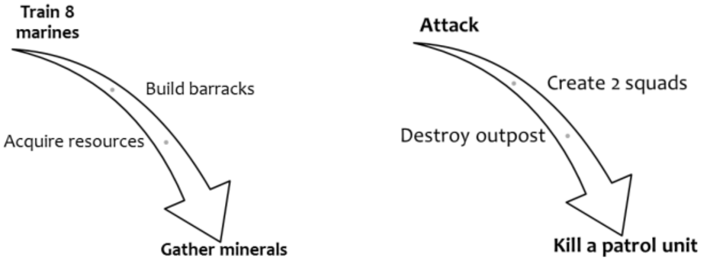


Figure 2: Order processing. In the production order example (left), the strategy manager decides to train 8 marines. This order is relayed to the production manager where it is processed and translated into a series of orders suchs as building a barracks and acquiring the necessary resources. The latter is in turn passed to a mining squad manager which sends workers out to a mineral field. The combat order example (right) can be analyzed is a similar fashion. When the strategy manager decides to attack, it raises an attack flag in the combat manager which responds by preparing some squads using the available units and locating a nearby target like an outpost. It then sends the squads there which run into a patrol unit on the way and receive the order to destroy it from their managers. We can thus see how orders spawn in the strategy manager as abstract decisions and are processed into more and more concrete decisions until they are eventually translated into specific unit commands.

[3] Quick attack against a supposedly unprepared opponent

4.3 Experimental Results and Limitations

In this section, we illustrate the performances of our agent in the particular case of StarCraft. The resulting agent is capable of playing full games as a Terran player facing a Zerg opponent. As can be seen in Tables 1 and 2, the agent controls units very efficiently compared to the default Zerg AI. It managed to score an average of 3 units lost in 10 games versus the opponent's 84.5 average. It even succeeded in winning without losing a single unit in a couple of games.

While this design allows for efficient agent implementations, it does not address the primary issue agents face in RTS games. Indeed, without learning, even an efficient agent will eventually lose against an adaptive opponent as the latter detects its inevitable flaws and starts exploiting them. Flaws can originate from multiple sources. First, it may be that our understanding of the way humans process game information is erroneous and omits important functions, resulting in an incomplete design. Furthermore, even a perfectly similar processing design would still fail to account for all possible game scenarios, leaving some situations uncovered and potentially exploitable.

Game	A	B	C	D	E
Units produced	73	62	72	69	68
Units killed	81	79	78	75	72
Units lost	10	2	0	0	1

Game	F	G	H	I	J
Units produced	77	76	63	85	78
Units killed	100	101	83	74	102
Units lost	2	1	0	13	1

Table 1: Unit statistics.

AUP	AUK	AUL	Games won	Total games
72.3	84.5	3	10	10

Table 2: Average units produced (AUP), killed (AUK) and lost (AUL).

5 Conclusions and Future Works

In this document, we have discussed some thoughts about RTS and the difficulties around it. We proposed a modular and hierarchical agent design inspired by human abstraction mechanisms for which promising experimental results were reported.

While the model described above can be used to efficiently implement effective agents, we still need to embed learning capabilities in order to address the main weakness of current agents, that is the lack of adaptation. Hence, with this model, we can imagine adding learning for carefully selected tasks in the different managers. For example, the strategy manager could learn new strategies by examining the opponents build orders and mimicking them. New tactics, such as squad formations, could also be learned from the opponents own unit grouping. At yet lower levels, military squad managers could learn to prioritize select targets based on the units the opponent takes out first. Although this has not been tested yet, the possibility of adding such features opens avenues to an exciting future for AI in RTS games. Yet further, another interesting step would be the automatic learning of control structures such as the one we proposed.

Acknowledgements

Raphael Fonteneau is a postdoctoral researcher of the FRS-FNRS from which he acknowledges the financial support. This paper presents research results of the Belgian Network DYSCO funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office.

References

Michael Buro and Timothy M. Furtak. RTS games and real-time AI research. In *Proceedings of the 13th Behavior Representation in Modeling and Simulation Conference (BRIMS-04)*, pages 51–58, 2004.

Santiago Ontañón, Kinshuk Mishra, Neha Sugandh, and Ashwin Ram. Case-based planning and execution for real-time strategy games. In *Proceedings of the 7th International Conference on Case-Based Reasoning (ICCBR-07)*, pages 164–178, 2007.

David W. Aha, Matthew Molineaux, and Marc J. V. Ponsen. Learning to win: case-based plan selection in a real-time strategy game. In *Proceedings of the 6th International Conference on Case-Based Reasoning (ICCBR-05)*, pages 5–20, 2005.

Ben Weber and Michael Mateas. Case-based reasoning for build order in real-time strategy games. In *Proceedings of the 5th Artificial Intelligence for Interactive Digital Entertainment Conference (AIIDE-09)*, 2009a.

Ben G. Weber and Michael Mateas. Conceptual neighborhoods for retrieval in case-based reasoning. In *Proceedings of the 8th International Conference on Case-Based Reasoning (ICCBR-09)*, pages 343–357, 2009b.

Ben G. Weber and Michael Mateas. A data mining approach to strategy prediction. In *Proceedings of the 5th IEEE Symposium on Computational Intelligence and Games (CIG-09)*, pages 140–147. IEEE Press, 2009c.

Marc Ponsen, Héctor Muñoz-Avila, Pieter Spronck, and David Aha. Automatically generating game tactics via evolutionary learning. *AI Magazine*, 27(3):75–84, 2006.

Michael Chung, Michael Buro, and Jonathan Schaeffer. Monte-Carlo planning in RTS games. In *Proceedings of the 1st IEEE Symposium on Computational Intelligence and Games (CIG-05)*, 2005.

Samuel Wintermute, Joseph Xu, and John E. Laird. SORTS: A human-level approach to real-time strategy AI. In *Proceedings of the 3rd Artificial Intelligence for Interactive Digital Entertainment Conference (AIIDE-07)*, pages 55–60, 2007.

Josh McCoy and Michael Mateas. An integrated agent for playing real-time Strategy games. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence (AAAI-08)*, pages 1313–1318, 2008.

Joseph Bates, A. Bryan Loyall, and W. Scott Reilly. Integrating reactivity, goals, and emotion in a broad agent. In *Proceedings of the 14th Annual Conference of the Cognitive Science Society*, 1992.

Michael Buro. Real-time strategy games: a new AI research challenge. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI-03)*, pages 1534–1535, 2003.

Danny C. Cheng and Ruck Thawonmas. Case-based plan recognition for real-time strategy games. In *Proceedings of the 5th International Conference on Computer Games: Artificial Intelligence, Design and Education (CGAIDE-04)*, pages 36–40, 2004.

Paul R. Cohen. Empirical methods for artificial intelligence. *IEEE Expert*, 11(6):88, 1996.

Alex Kovarsky and Michael Buro. A first look at build-order optimization in real-time strategy games. In *Proceedings of the 2006 GameOn Conference*, pages 18–22, 2006.

Distributed Decision Making by Categorically-Thinking Agents

Joong Bum Rhim
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139
jbrhim@mit.edu

Lav R. Varshney
IBM Thomas J. Watson Research Center
Hawthorne, NY 10532
lrvarshn@us.ibm.com

Vivek K Goyal
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139
vgoyal@mit.edu

Abstract

This paper considers group decision making by imperfect agents that only know quantized prior probabilities for use in Bayesian likelihood ratio tests. Global decisions are made by information fusion of local decisions, but information sharing among agents before local decision making is forbidden. The quantization scheme of the agents is investigated so as to achieve the minimum mean Bayes risk; optimal quantizers are designed by a novel extension to the Lloyd-Max algorithm. Diversity in the individual agents' quantizers leads to optimal performance.

1 Introduction

Consider a binary decision problem. The Bayes rational decision making strategy is to perform the likelihood ratio test (LRT). Decision makers first compute the likelihood ratio of states of an object based on an observation. Then they make a decision by comparing the ratio to a decision threshold determined by the prior probability of the state and their costs. Not only do LRTs minimize Bayes risk, but also psychology experiments suggest that human decision makers employ them [1, 2].

Optimal LRTs require precise knowledge of the prior probabilities of object states. Much previous research considers the prior probability to be a constant known to decision makers. However, decision makers may face a great variety of objects. For example, soccer referees handle more than twenty-two players in one game and salespeople at stores observe hundreds of customers in one day. This is problematic because players have different prior probabilities of committing fouls and customers have different prior probabilities of making purchases.

Decision makers should use different thresholds uniquely optimized to different objects of decision making, such as players or customers. Computing thresholds and then remembering them is an information processing burden, especially for human decision makers that often resort to categorical and coarse thinking [3, 4]. Human decision makers can afford around seven categories without getting confused [5]. Thus, we model decision makers as grouping similar objects together and treating them identically by applying a single decision threshold. By classifying all objects into a small number of categories, decision makers can handle infinitely many objects; however decision makers consequently have limited threshold precision, a type of bounded rationality.

In the context of LRTs, categorization of objects is equivalent to quantization of their prior probabilities. With this idea, we move from considering a single object with a constant prior probability to considering an ensemble of objects with performance averaged over the distribution of prior probabilities.

Consider a decision-making group of N agents that chooses between two hypotheses h_0 and h_1. Agents make local hard decisions without knowing other agents' decisions. Local decisions are combined by a fusion center to produce a global decision. The fusion center has a fixed symmetric fusion rule of the L-out-of-N form whereby the global decision is h_1 when L or more agents choose h_1. The symmetric fusion rule implies that all agents have an equal voice. Due to information-processing constraints, agents must quantize prior probabilities to one of K values. Our interest here is to design quantizers that lead to the smallest Bayes risk on average.

The study of quantization of prior probabilities originates from [6], which focuses on the minimum mean Bayes risk error (MBRE) quantizer of a single agent. Maximum Bayes risk error is considered in [7, 8]. Recent results and economic implications are reviewed in [9].

We have previously considered a distributed hypothesis testing problem with similar imperfect agents, but where each agent is assumed to know other agents' quantized prior probabilities, whether they have a common interest [8, 10] or whether they have conflicts of interest [11]. The assumption in these prior papers enables agents to optimize decision rules so as to minimize Bayes risk within either the collaboration or the conflict system.

Information about other agents' quantizers should not be taken for granted; it requires a coordination mechanism built on communication channels. Such communication may not be possible in human group decision-making scenarios due to geographic separation, desire to remain clandestine, or if N is too large. In engineering applications, memory or power constraints may prevent detectors from exchanging any information with neighboring detectors. In these scenarios, each agent has to make decisions based on its information—its quantized prior probability and observed signal—only. In this paper, agents do not know how other agents quantize prior probabilities.

Lack of knowledge about others makes it impossible for agents to collaborate by sharing a common goal. Hence, their quantizers need to be cleverly designed so that local decision making becomes harmonious with respect to the global mean Bayes risk (MBR), the distortion measure for quantization. A modified Lloyd-Max algorithm can design MBR-optimal quantizers. It is demonstrated that diversity among agents in quantization of prior probabilities can be helpful to improve the quality of group decision making.

The group decision-making model we consider is described in Section 2. In Section 3, we analyze the mean Bayes risk in terms of endpoints and representation points of quantizers. Then we propose an algorithm to design optimal quantizers. An example of optimal quantizers obtained from our algorithm is presented in Section 4. Section 5 concludes the paper.

2 Distributed Decision-Making Model with Imperfect Agents

We consider a team of N agents and an object in one of two binary states $H \in \{h_0, h_1\}$. The prior probability of the object being in state h_0, $p_0 = \mathbb{P}\{H = h_0\}$, is a realization of a random variable P_0 drawn from its distribution f_{P_0}. Since the prior probability of being in state h_1 is determined by p_0 through $p_1 = 1 - p_0$, by the term *prior probability* we simply mean p_0. The prior probability is important for good decision making but Agent i only knows its quantized value of the prior probability, $q_i(p_0)$.

Agent i makes a noisy state measurement Y_i with likelihood functions $f_{Y_i \mid H}(y_i \mid h_0)$ and $f_{Y_i \mid H}(y_i \mid h_1)$. Agent i then makes a hard decision $\widehat{H}_i$ whether the object is in state h_0 or in h_1 based on the quantized prior probability $q_i(p_0)$ and the observation Y_i. Its decision is transferred to a fusion center, which makes a global decision $\widehat{H}$ as h_1 if it receives h_1 from L or more agents

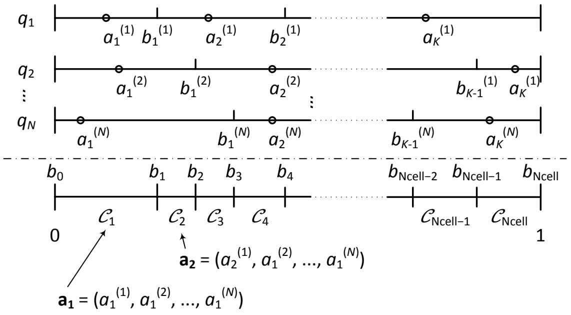


Figure 1: A set of quantizers $q_1, \dots, q_N$ categorizes cells $\mathcal{C}_1, \dots, \mathcal{C}_{N_{\text{Cell}}}$.

and as h_0 otherwise. The agents incur cost c_{10} for a false alarm ($\widehat{H} = h_1$ when $H = h_0$) and c_{01} for a missed detection ($\widehat{H} = h_0$ when $H = h_1$); costs c_{10} and c_{01} are common for all agents. For simplicity, correct global decisions incur zero cost.

Agent i optimizes its decision rule as if it is the only decision maker because it does not have any information about other agents:

$$\frac{f_{Y_i \mid H}(y_i \mid h_1)}{f_{Y_i \mid H}(y_i \mid h_0)} \underset{\widehat{H}_i(y)=h_0}{\overset{\widehat{H}_i(y)=h_1}{\gtrless}} \frac{c_{10} q_i(p_0)}{c_{01}(1 - q_i(p_0))}. \tag{1}$$

This decision rule yields an error with probability $P_{e,i}^{\mathrm{I}} = \mathbb{P}\{\widehat{H}_i = h_1 \mid H = h_0\}$ when $H = h_0$ and with probability $P_{e,i}^{\mathrm{II}} = \mathbb{P}\{\widehat{H}_i = h_0 \mid H = h_1\}$ when $H = h_1$.

The agents cannot collaborate to design a decision rule, but they still fuse their decisions to make a global decision. By using L-out-of-N fusion rules, the global decision is wrong if L or more agents send h_1 when $H = h_0$ or if $N - L + 1$ or more agents send h_0 when $H = h_1$. These error probabilities, P_E^{I} and P_E^{II}, are used in computing the Bayes risk

$$R = c_{10} p_0 P_E^{\mathrm{I}} + c_{01}(1 - p_0) P_E^{\mathrm{II}}.$$

3 Optimal Quantization of Prior Probabilities

Agent i has quantizer q_i for prior probability p_0, which has K cells $[0, b_1^{(i)}), [b_1^{(i)}, b_2^{(i)}), \dots, [b_{K-1}^{(i)}, 1]$ with corresponding representation points $a_1^{(i)}, a_2^{(i)}, \dots, a_K^{(i)}$, where $a_k^{(i)} = q_i(p_0)$ for all $p_0 \in [b_{k-1}^{(i)}, b_k^{(i)})$. We define a set of endpoints $\{0, b_1, b_2, \dots, b_{N_{\text{Cell}}-1}, 1\}$, $0 < b_1 < b_2 < \dots < b_{N_{\text{Cell}}-1} < 1$, as the union of endpoints of all quantizers $q_1, \dots, q_N$ and define cells $\mathcal{C}_k$ as the intervals $[b_{k-1}, b_k)$, where N_{Cell} is the number of the cells $\mathcal{C}_k$. The maximum number of N_{Cell} is $N(K - 1) + 1$. For cell $\mathcal{C}_k$, we define a vector of representation points $\mathbf{a}_k = (q_1(p_0), q_2(p_0), \dots, q_N(p_0))$, where $p_0 \in \mathcal{C}_k$, see Fig. 1. The necessary conditions of representation points and endpoints for local optimality of the quantizers are now derived.

3.1 Representation points

Quantization performance is measured by Bayes risk averaged over P_0:

$$\mathbb{E}[R] = \int_0^1 \left(c_{10} p_0 P_E^{\mathrm{I}}(q_1(p_0), \dots, q_N(p_0)) + c_{01}(1 - p_0) P_E^{\mathrm{II}}(q_1(p_0), \dots, q_N(p_0)) \right) f_{P_0}(p_0)\, dp_0.$$

Within cell $\mathcal{C}_k$, since $(q_1(p_0), \dots, q_N(p_0)) = \mathbf{a}_k$ is constant, the mean Bayes risk (MBR) is

$$\mathbb{E}[R]_k = \int_{\mathcal{C}_k} \left(c_{10} p_0 P_E^{\mathrm{I}}(\mathbf{a}_k) + c_{01}(1 - p_0) P_E^{\mathrm{II}}(\mathbf{a}_k) \right) f_{P_0}(p_0)\, dp_0 = c_{10} \pi_k^{\mathrm{I}} P_E^{\mathrm{I}}(\mathbf{a}_k) + c_{01} \pi_k^{\mathrm{II}} P_E^{\mathrm{II}}(\mathbf{a}_k),$$

where $\pi_k^{\mathrm{I}} = \int_{\mathcal{C}_k} p_0 f_{P_0}(p_0)\, dp_0$ and $\pi_k^{\mathrm{II}} = \int_{\mathcal{C}_k} (1 - p_0) f_{P_0}(p_0)\, dp_0$ are constants with respect to $\mathbf{a}_k$.

Let us fix all representation points except that of q_z, a_{kz}, in $\mathcal{C}_k$. The mean Bayes risk in $\mathcal{C}_k$ can be written as $\mathbb{E}[R]_k = \alpha_1 P_{e,z}^{\mathrm{I}} + \alpha_2 P_{e,z}^{\mathrm{II}} + \alpha_3$, where α_1, α_2, and α_3 are positive constants. Since $P_{e,z}^{\mathrm{II}}$ is strictly convex in the $P_{e,z}^{\mathrm{I}}$ and vice versa [12], $\mathbb{E}[R]_k$ is strictly convex in $P_{e,z}^{\mathrm{I}}(a_{kz})$ and $P_{e,z}^{\mathrm{II}}(a_{kz})$.

The convexity is preserved in the entire MBR $\mathbb{E}[R] = \mathbb{E}[R]_1 + \ldots + \mathbb{E}[R]_{N_{\mathrm{Cell}}}$ because the MBR in each cell is strictly convex in $P_{e,z}^{\mathrm{I}}(a_{kz})$ and $P_{e,z}^{\mathrm{II}}(a_{kz})$ or constant. Therefore, the value of a_{kz} that minimizes the MBR exists uniquely for any $1 \leq k \leq N_{\mathrm{Cell}}$ and $1 \leq z \leq N$. The value of a_{kz} should be the minimum point.

3.2 Endpoints

Let us fix all representation points and endpoints except an endpoint b_j. The endpoint b_j only affects two adjacent cells $\mathcal{C}_j$ and $\mathcal{C}_{j+1}$, whose boundary is b_j.

$$\begin{aligned}\mathbb{E}[R]_j + \mathbb{E}[R]_{j+1} = &\int_{b_{j-1}}^{b_j} \left(c_{10} p_0 P_E^{\mathrm{I}}(\mathbf{a}_j) + c_{01}(1-p_0) P_E^{\mathrm{II}}(\mathbf{a}_j)\right) f_{P_0}(p_0)\, dp_0 \\ &+ \int_{b_j}^{b_{j+1}} \left(c_{10} p_0 P_E^{\mathrm{I}}(\mathbf{a}_{j+1}) + c_{01}(1-p_0) P_E^{\mathrm{II}}(\mathbf{a}_{j+1})\right) f_{P_0}(p_0)\, dp_0\end{aligned}$$

Taking the derivative of the MBR, we have

$$\begin{aligned}\frac{d}{db_j}(\mathbb{E}[R]) &= \frac{d}{db_j}(\mathbb{E}[R]_j + \mathbb{E}[R]_{j+1}) \\ &= \left(c_{10} b_j (P_E^{\mathrm{I}}(\mathbf{a}_j) - P_E^{\mathrm{I}}(\mathbf{a}_{j+1})) - c_{01}(1-b_j)(P_E^{\mathrm{II}}(\mathbf{a}_{j+1}) - P_E^{\mathrm{II}}(\mathbf{a}_j))\right) f_{P_0}(b_j) \quad (2)\end{aligned}$$

If we compare each entry of two vectors $\mathbf{a}_j$ and $\mathbf{a}_{j+1}$, at least one entry has different values. For any entry that has a different value, $\mathbf{a}_{j+1}$ has a greater value than $\mathbf{a}_j$ does because the former represents larger P_0. A bigger representation point leads to a smaller local false alarm probability. Thus, $P_E^{\mathrm{I}}(\mathbf{a}_{j+1}) < P_E^{\mathrm{I}}(\mathbf{a}_j)$. On the contrary, $P_E^{\mathrm{II}}(\mathbf{a}_{j+1}) > P_E^{\mathrm{II}}(\mathbf{a}_j)$. Let $\beta_1 = P_E^{\mathrm{I}}(\mathbf{a}_j) - P_E^{\mathrm{I}}(\mathbf{a}_{j+1}) > 0$ and $\beta_2 = P_E^{\mathrm{II}}(\mathbf{a}_{j+1}) - P_E^{\mathrm{II}}(\mathbf{a}_j) > 0$.

$$\frac{d}{db_j}(\mathbb{E}[R]) = ((c_{10}\beta_1 + c_{01}\beta_2) b_j - c_{01}\beta_2) f_{P_0}(b_j). \quad (3)$$

This first derivative is zero at only one or no point if $f_{P_0}(p_0) > 0, \forall p_0 \in [0,1]$. This means that $\mathbb{E}[R]$ has only one or zero extreme point for $b_j \in (b_{j-1}, b_{j+1})$: if it has one extreme point, then it is the minimum point. Otherwise, either b_{j-1} or b_{j+1} is the minimum point. The value of b_j should be the minimum point.

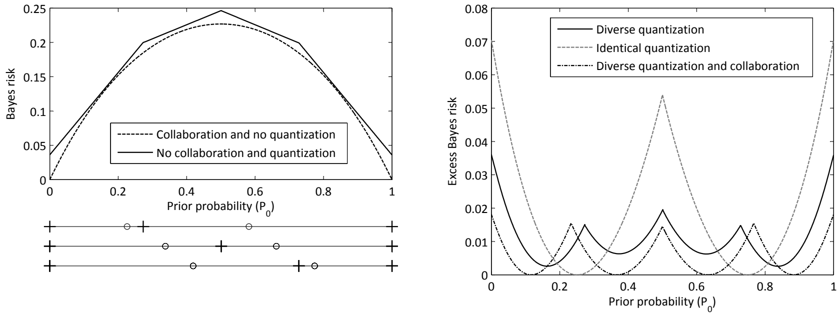
3.3 Algorithm

The iterative Lloyd-Max algorithm is applied to find an optimal quantizer in a single-agent decision-making model [6]. In this problem, however, the algorithm needs to be modified so as to optimize N different quantizers together. The key to the Lloyd-Max algorithm is alternating iterations of finding optimal endpoints while fixing representation points and finding optimal representation points while fixing endpoints.

In our group decision-making model, optimization steps are complicated because of dependency among variables; a change of one representation point also changes optimal values of other representation points. Hence, representation points are repeatedly adjusted until every representation point is optimal for the other representation points and given endpoints. Likewise for optimization of endpoints.

We use the following alternating nested-iteration optimization algorithm:

1. Assign initial values to endpoints and representation points.
2. (E-Step) Optimize endpoints with representation points fixed.
   (a) From the first endpoint variable $b_1^{(1)}$ to the last one $b_{K-1}^{(N)}$, successively optimize each variable.



(a) (b)

Figure 2: Bayes risk for uniformly distributed P_0 and $K = 2$. (a) An optimal set of quantizers (cell endpoints as +'s and representation points as ∘'s) and the resulting Bayes risk. (b) The performance loss in terms of Bayes risk due to the quantization of prior probabilities.

   (b) Repeat step (a) until all endpoints become stable, i.e., a new iteration does not change any endpoints.

3. (R-Step) Optimize representation points with endpoints fixed.
   (a) From the first representation point variable $a_1^{(1)}$ to the last one $a_{K-1}^{(N)}$, successively optimize each variable.
   (b) Repeat step (a) until all representation points become stable.
4. Iterate E-Step and R-Step until all endpoints and representation points become stable.

4 Example

As an example, let us consider the following measurement model for $N = 3$ agents:

$$Y_i = s_m + W_i, \quad i = 1, \dots, N, \quad m \in \{0, 1\}, \tag{4}$$

where $s_0 = 0$, $s_1 = 1$, and W_i is a zero-mean Gaussian random variable with variance $\sigma^2 = 1$. The Bayes costs are $c_{10} = c_{01} = 1$. The local decisions are fused by MAJORITY rule (2-out-of-3 rule).

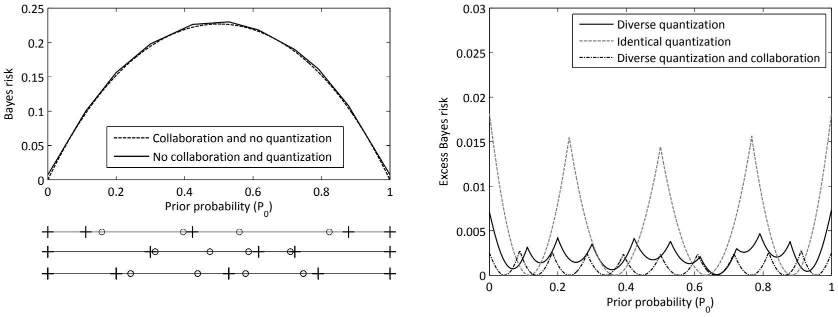
Fig. 2a shows Bayes risk when the agents can distinguish 2 categories, i.e., they use 2-level quantizers. The Bayes risk (solid piecewise line) is compared to the Bayes risk when the agents can distinguish any prior probability exactly and collaborate with others (dashed curve) like in [10], which is the best performance that the agents can achieve. The excess Bayes risk, the difference between the Bayes risks with and without quantization, is depicted in Fig. 2b. It shows the performance loss due to quantized prior probabilities compared to the best performance. For comparison, Fig. 2b also shows the performance loss when the agents are forced to use identical quantizers (gray dashed curve) and the performance loss when the agents use diverse quantizers and can collaborate by sharing their quantized values (dash-dot curve). The latter is the best performance that the agents can achieve with quantized prior probabilities [8, 10].

The Bayes risks when the agents use 4-level quantizers are depicted in Fig. 3.

5 Conclusion

We have discussed decision making by multiple agents that have imperfect perception ability. There are two factors that worsen the quality of global decisions. First, they perform local testing based on quantized prior probabilities. Second, they do not know how other agents quantize prior probabilities. We have determined the effect of these factors on Bayes risk in decision making.

To minimize the negative influence from these factors, we have defined mean Bayes risk as the optimization criterion for prior-probability quantizers. The Lloyd-Max algorithm is modified to an algorithm with double-iteration structure to design optimal quantizers. Using the algorithm, we



(a) (b)

Figure 3: Bayes risk for uniformly distributed P_0 and $K = 4$. (a) An optimal set of quantizers (cell endpoints as +'s and representation points as ∘'s) and the resulting Bayes risk. (b) The performance loss in terms of Bayes risk due to the quantization of prior probabilities.

have provided an example of additive white Gaussian noise model. The result shows that the MBR when the agents use diverse quantizers is lower than the MBR when they use identical quantizers. It is reasonable because N_{Cell} $(= N(K-1)+1)$ when they use diverse quantizers is greater than N_{Cell} $(= K)$ when they use identical quantizers. Therefore, we can conclude that the diversity among agents is still helpful even though they cannot fully utilize the diversity because of lack of knowledge about other agents.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grants 0729069 and 1101147 and the Korea Foundation for Advanced Studies (KFAS) Fellowship.

References

[1] J. A. Swets, T. Wilson P, Jr., and T. G. Birdsall, "Decision process in perception," *Psychological Review*, vol. 68, no. 5, pp. 301–340, Sep. 1961.

[2] M. Glanzer, A. Hilford, and L. T. Maloney, "Likelihood ratio decisions in memory: Three implied regularities," *Psychonomic Bulletin & Review*, vol. 16, no. 3, pp. 431–455, Jun. 2009.

[3] G. A. Miller, "Human memory and the storage of information," *IRE Trans. Inf. Theory*, vol. 2, no. 3, pp. 129–137, Sep. 1956.

[4] R. Fryer and M. O. Jackson, "A categorical model of cognition and biased decision making," *The B.E. Journal of Theoretical Economics*, vol. 8, no. 1, Feb. 2008.

[5] G. A. Miller, "The magical number seven, plus or minus two: Some limits on our capacity for processing information," *Psychological Review*, vol. 63, no. 2, pp. 81–97, Mar. 1956.

[6] K. R. Varshney and L. R. Varshney, "Quantization of prior probabilities for hypothesis testing," *IEEE Trans. Signal Process.*, vol. 56, no. 10, pp. 4553–4562, Oct. 2008.

[7] ——, "Multilevel minimax hypothesis testing," in *Proc. IEEE/SP Workshop Stat. Signal Process.*, Nice, France, Jun. 2011, pp. 109–112.

[8] J. B. Rhim, L. R. Varshney, and V. K. Goyal, "Quantization of prior probabilities for collaborative distributed hypothesis testing," *arXiv:1109.2567*, 2011.

[9] L. R. Varshney, J. B. Rhim, K. R. Varshney, and V. K. Goyal, "Categorical decision making by people, committees, and crowds," in *Proc. Information Theory and Applications Workshop*, La Jolla, CA, Feb. 2011.

[10] J. B. Rhim, L. R. Varshney, and V. K. Goyal, "Collaboration in distributed hypothesis testing with quantized prior probabilities," in *Proc. IEEE Data Compression Conf.*, Snowbird, UT, Mar. 2011, pp. 303–312.

[11] ——, "Conflict in distributed hypothesis testing with quantized prior probabilities," in *Proc. IEEE Data Compression Conf.*, Snowbird, UT, Mar. 2011, pp. 313–322.

[12] H. L. van Trees, *Detection, Estimation, and Modulation Theory, Part I.* New York: John Wiley & Sons, 1968.

Decision making and working memory in adolescents with ADHD after cognitive remediation

Michel Bader
Service Universitaire de Psychiatrie
de l'Enfant et de l'Adolescent (SUPEA)
CH-1011 Lausanne, Switzerland
Michel.Bader@chuv.ch

Sarah Leopizzi
Service Universitaire de Psychiatrie
de l'Enfant et de l'Adolescent (SUPEA)
CH-1011 Lausanne, Switzerland

Eleonora Fornari
Department of Radiology
Centre Hospitalier Universitaire Vaudois (CHUV) CH-1011 Lausanne, Switzerland
and
CIBM (Centre d'Imagrie Biomédicale), CHUV unit
Lausanne, Switzerland

Olivier Halfon
Service Universitaire de Psychiatrie
de l'Enfant et de l'Adolescent (SUPEA)
Lausanne, Switzerland

Nouchine Hadjikhani
Martinos Center for Biomedical Imaging
Harvard Medical School
Boston, USA
and
Brain and Mind Institute
EPFL-SV-BMI
CH-1015 Lausanne, Switzerland

Abstract

An increasing number of theoretical frameworks have incorporated an abnormal sensitivity response inhibition as to decision-making and working memory (WM) impairment as key issues in Attention deficit hyperactivity disorder (ADHD). This study reports the effects of 5 weeks cognitive training (RoboMemo®, Cogmed) with fMRI paradigm by young adolescents with ADHD at the level of behavioral, neuropsychological and brain activations. After the cognitive remediation, at the level of WM we observed an increase of digit span without significant higher risky choices reflecting decision-making processes. These preliminary results are promising and could provide benefits to the clinical practice. However, models are needed to investigate how executive functions and cognitive training shape high-level cognitive processes as decision-making and WM, contributing to understand the association, or the separability, between distinct cognitive abilities.

1 Introduction

Attention deficit hyperactivity disorder (ADHD) is one of the most common neurobehavioral disorder of childhood and impacts many aspects of development at home and at school, including social, emotional and cognitive functioning. ADHD is a highly prevalent disorder worldwide, thought to affect 5%-9% of children [1] and 3-4% of adults [2, 3]. Longitudinal follow-up studies show that the majority of ADHD children and teenagers experience persistent symptoms and functional impairments into early adultdhood [4, 5, 6]. Throughout the life cycle, patients with ADHD have high rates of comorbidity with oppositional defiant disorder, conduct disorder, mood disorder (both unipolar and bipolar), anxiety disorders, and learning disorders [7]. Several studies have consistently documented that ADHD is associated with high levels of grade retention, need for tutoring, lower levels of overall achievement when compared with control subjects [8]. The annual costs of ADHD in the US is substantial, amounting to a total excess cost of USD 31.6 billion in 2000 [9].

The most effective and widely used treatments for ADHD are medication and behavior modification. These empirically-supported interventions are generally successful in reducing ADHD symptoms, but treatment effects are rarely maintained beyond the active intervention. Because ADHD is now generally thought of as a chronic disorder that is often present well into adolescence and early adulthood, the need for continued treatment throughout the lifetime is both costly and problematic for a number of logistical reasons.

The evolving field of research on ADHD has now moved beyond the search of a common core dysfunction towards a recognition of ADHD as a heterogeneous disorder of multiple neuropsychological deficits and hypothesized causal substrates [10, 11, 12, 13, 14, 15]. For many years the focus of cognitive research has been on deficits in executive function [16, 17], especially inhibition [18]. Recently, however, an increasing number of theoretical frameworks have incorporated an abnormal sensitivity response inhibition as to reinforcement and WM impairment as three of the key issues in ADHD [19, 20, 13, 11]. An abnormal sensitivity to reinforcement may influence cognitive processes such as decision making through unconscious "somatic marker signals" that arise from bioregulatory processes [21]. WM capacity is an important factor to determining problem solving and reasoning ability. WM impairment is of central importance in ADHD, probably caused by impaired function of the prefrontal and parietal cortex [19, 22, 13, 23].

Recent studies observe the impact of decision-making and reinforcement contingencies on ADHD subjects on performance and on levels of motivation [20], i.e. strategies of reward and response cost [10], or of WM training [24, 25]. Children with ADHD have been found to show an increased sensitivity to instances of (immediate) gratification (see [20] for review). Otherwise, children with ADHD have been found to require more response cost than controls in order to perform accurately [26], suggesting that children with ADHD suffer from a diminished sensitivity to negative outcomes.

At the behavioral level a comparison between ADHD children and normal controls aged 7 to 10 years performing a simple probabilistic discounting task has observed children with ADHD opted more frequently for less likely but larger rewards than normal controls [27]. Shifts of the response category after positive or negative feedback, however, occurred as often in children with ADHD as in control children. In children with ADHD, the frequency of risky choices was correlated with neuropsychological measures of response time variability but unrelated to measures of inhibitory control. The authors have suggested that the tendency to select less likely but larger rewards possibly represents a separate facet of dysfunctional reward processing, independent of delay aversion or altered responsiveness to feedback.

WM is the ability to retain different pieces of information and to then access them to make a decision. Both, WM and inhibition are correlated with IQ. In turn, IQ is correlated with impatience and present-biased preferences [28, 29, 30, 31]. Risk preferences also correlate with IQ, with smarter individuals tending to make more risk-neutral choices and being less sensitive to losses [29]. When both measures are available, WM is often more strongly correlated with these measures than IQ, indicating that this aspect of executive function is particularly important [31]. Thus, these correlational studies suggest that executive function affects choices in the direction of making them more compatible with the economic model.

Recent studies have shown that WM can be substantially improved with an intensive training regimen. Such training has been shown to have an impact on non-trained tasks directly related to WM [25, 32] and IQ tests [32, 31]. More intriguingly, WM training also appears to have transfer effects less directly related to measures that correlate with executive function. WM and inhibition task activate common brain regions [33], including right inferior frontal gyrus and right middle frontal gyrus, as well as right parietal regions. Increased brain activity is found in several of these regions after intensive WM training [24, 34], thus leading to the hypothesis that WM training may also have spillover effects onto inhibition.

Consistent with this hypothesis, children with ADHD, after a sequence of intensive WM training, are reported to be less inattentive and less impulsive by their parents, and have improved performance in WM as well as inhibitory tasks [25]. Similarly, no transfer effects of more generalized cognitive training in an elderly subject population were found [35]. Thus, evidence on transfer effects of cognitive trainings is mixed on the score. Results seem to indicate that cognitive training has stronger transfer effects in ADHD populations.

2 Methods

2.1 Participants

ADHD participants were 11-15 years old adolescents with ADHD (DSM-IV criteria) and with methylphenidate treatment. They were subdivided in two groups.

- Adolescents ADHD with methylphenidate and with cognitive training (N=18 : 4 girls and 14 boys ; 13±0.9 years old)
- Adolescents ADHD with methylphenidate without cognitive training (N=10 : 1 girl and 9 boys ; 12±1.1 years old)

The controls were healthy subjects (N=21 : 5 girls and 14 boys ; 12±1.8 years old) who did the cognitive training. The following preliminary results investigate only the adolescents ADHD with methylphenidate and with cognitive training, and the control subjects with the cognitive training. Regarding previous studies [25, 34] and research fundings no control group without training was included at this study.

Mini International Neuropsychiatric Interview (M.I.N.I.) [36] is a structured instrument completed by the parents (or primary caregiver) probing the most common child psychiatric diagnoses. This instrument allows the diagnosis of ADHD and co-morbid symptoms, including oppositional defiant disorder, conduct disorders, internalising disorders. The MINI is relatively easy to use for researchers and does not require comprehensive clinical training. This instrument allows the diagnosis of ADHD with DSM-IV criteria (DSM-IV, 1994) [37] and co-morbid symptoms, including oppositional defiant disorder, conduct disorders, internalizing disorders. The MINI is relatively easy to use for researchers and does not require comprehensive clinical training. ADHD characteristic were also assessed using the Revised Conners' Parent Rating Scale (CPRS-R; [38]) and the Conners-Wells' Adolescent Self-Report Scale CASS; [39],and the BRIEF [40, 41] specifically designed to assess child and adolescent everyday executive skills in natural, everyday environments, including home and school.

The Conners' Parents Rating Scale (CPRS-R) [38] is a validated 82-items parental assessment containing the following 9 subscales: family problems, emotional problems, conduct problems, anger control problems, hyperactivity, ADHD index. DSM-IV total score, DSM-IV ADHD inattention, and DSM-IV ADHD hyperactivity-impulsivity. Items will be scored on a 4-point scale (from 0, not true, to 3, very much true). The psychometric properties of this revised scale appear adequate as demonstrated by good internal reliability coefficient, high test-retest reliability, and effective discriminatory power.

The Conners-Wells Adolescent Self Report of Symptom Scale (CAARS-L) [39] is a validated 87-items self report assessment containing the following 9 subscales: family problems, emotional problems, conduct problems, anger control problems, hyperactivity, ADHD index, DSM-IV ADHD total score, DSM-IV ADHD inattention, and DSM-IV ADHD hyperactivity-impulsivity. Items will be scored on a 4-point scale (from 0, not true, to 3, very much true).

The neuropsychological measures were investigated by following tasks. The span-board task from the WAIS-RNI testing battery [42] will be used to measure visuospatial WM. The mean performance from trials with forward and backward repeating of the memoranda will be used in the analysis to provide a more reliable measure. The WISC-III digit span subtest [43] is derived from scores of the two subtests: digits forwards and digit backward. The first subtest clearly indexes verbal storage processes, whereas the reverse version requires the subject to recall the series of digits in reverse order, wich entails manipulation of information [22].

The Stroop Color and Word Test [44, 45] measures a participant's ability to respond selectively to one dimension of a multidimentional stimulus. Each of the three trials of the Stroop Test uses a card containing five columns of 20 stimuli. The participant will be asked to complete as many of the stimuli as possible within 45 s. The Word trial requires the participant to read as rapidly as possible the names of colours (red, blue, and green) printed on the card in black ink. The subsequent Color trial requires the participant to name consecutively the color of each stimuli on a card containing colored patches of red, blue, or green ink. Finally, the card for the Interference trial contains the words red, blue, and green printed in a different colour, and the participant is asked to name the colour of the ink of each stimulus. The three trials of the Stroop are hypothesized to assess reading speed, naming speed, and interference control [46]

The visuo-spatial Corsi [47] is used to determine the visual-spatial memory span and the implicit visual-spatial learning abilities. Participants sit with nine wooden 3x3 cm blocks fastened before them on a 25 x 30 cm baseboard in a standard random order. The subject taps a sequence pattern onto the blocks with participants must the replicate. The sequence length increases each trial until the participant is no longer able to correctly replicate the pattern.

The Test Battery of Attentional Performrmance (TAP 2.2) [48] is used to investigate the divided attention which is the ability to successfully execute more than one action at a time, while paying attention to two or more channels of information. The TAP 2.2 in a ”dual-task” paradigm, in which two stimuli have to be attenuated simultaneously.

The Go/Nogo paradigm (Go/Nogo - TAP 2.2) [48] is used to test this form of behavioral control, in which it is important to suppress a reaction triggered by an external stimulus to the benefit of an internally controlled behavioral response. Go/Nogo test is used to measure a subject capacity for sustained attention and response control. The test requires a participant to perform an action given certain stimuli (press a button - Go) and inhibit that action under a different set of stimuli (not press the same button - NoGo).

The Trailmaking Test (TMT B) from the Halstead-Reitan Neuropsychological Battery [49]. The Trailmaking Test requires subjects to trace a path between consecutive letters scattered randomly on the page (Form A) and alternative letters and numbers (Form B) as rapidly as they can without making errors. The difference between B and A time is viewed as an index of set shifting ability [13].

The Raven's Colored Progressive Matrices (RCPM) [50] measures clear-thinking ability and consists of 36 items in 3 sets, with 12 items per set (A and B from the standard matrices, with a further set of 12 items between the two, as set Ab). The RCPM items are arranged to assess cognitive development up to the stage when a person is sufficiently able to reason by analogy and adopt this way of thinking as a consistent method of inference. Most items are presented on a colored background to make the test visually stimulating for participants. However, the very last items in set B are presented as black-on-white; in this way, if a subject exceeds the tester's expectations, transition to sets C, D, and E of the standard matrices is eased. The stability of the measure over time and across cultures has been demonstrated [51].

After the first visit (time=T1), a brain magnetic resonance imaging data collection including WM stimuli was conducted for all subjects at the University Radiology Department of the CHUV in Lausanne (Professor Reto Meuli) within the Unit of the Centre d'imagerie biomédicale (CIBM) in Lausanne (Professor Matthias Stuber). After the 5 weeks cognitive training, the second visit (time=T2) was performed including post-test assessments (ADHD scales, neuropsychological post-test and post-test fMRI). After 2 months, the parents have filled an assement evaluation executive functioning at home.

2.2 Cognitive Training

Participants do a systematic training of performing executive functions tasks during a 5-week period, implemented in a computer program (RoboMemo®, Cogmed). The cognitive training includes visuospatial WM tasks as well as verbal tasks (remembering phonemes, letters, or digits). Each of the games is designed to have an element of pressure and excitement, to maximize the subject's motivation to participate and the “rush” that has been shown to assist performance and decision making processes in subjects with ADHD. The games are designed so that each participant c an do his/her remediation each day at home, thereby accruing including time and date stamps.

The subjects perform 115 WM trials during each session. Total time will depend on the level and the time taken between trials. Each session will last approximately 30-45 minutes (excluding breaks).

Difficulty level will be automatically adjusted, on a trail-by-trial basis, to match the WM span of the subject on each task. Response to each trial is logged to a file computer.

Once every week during the training period, a research psychologist calls participating the subjects, and the teenager's parents, to ask about technical difficulties and to check the number of sessions to be uploaded. This procedure helps target compliance, serves to prevent lagtime due to technical difficulties, and reinforces progress linked to working memory and decision making during the training period.

3 Results

The following preliminary results investigate only the adolescents ADHD with methylphenidate and with cognitive training vs. the control subjects with the cognitive training.

The results shown in Figure 1 indicate that the adolescents with ADHD have a lower digit span compared to control subjects. The digit span uses numbers. The subjects has to repeat back in correct order immediately after presentation. Backward digit span is a more challenging variation which involves recalling items in reverse order. In this test the examinee required to repeat 3 - 9 digits forward and 2 - 9 digits backwards. This difference is particularly evident in the backward form, suggesting difficuties in verbal WM in accordance with Klingberg studies. These observations are in concordance with observations of the literature on ADHD [19, 22, 14, 23] and seems to reflect impaired function of the prefrontal and parietal cortex.

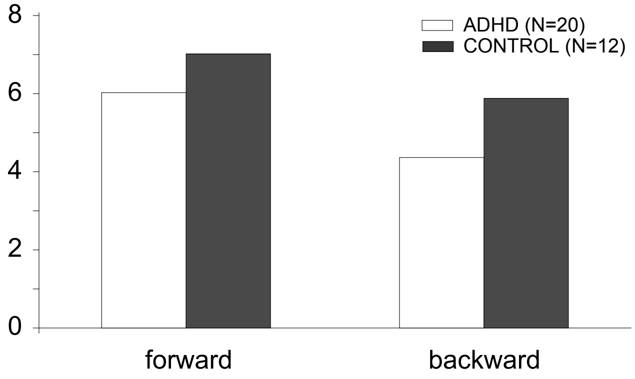


Figure 1: Digit Span of ADHD (N=20) vs. CONTROL (N=12) participants before cognitive remediation.

After the cognitive remediation, we observed an increase of digit span (Fig. 2). Both groups made equivalent progress and the index improvement of Cogmed training did not show significant difference between groups of ADHD participants (0.08, Mann-Whitney test). The effects of the training were more effective for the adolescents with ADHD than in the control participants. In particular, the clinical subjects describe a subjective positive decision making improvement at the end of the cognitive training. The current data analysis do not show higher significant risky choices during the 2'800 WM trials of the cognitive training and the high flood of working memory. The cognitive training seems to improve the ability to decision-making through the frequency of stimuli without a stress of magnitude of penalty.

The follow up inquiry showed that most of the parents of the ADHD subjects perceived a positive effect in the daily life (Fig. 3). The improvements are observed at school and at home relating school performances, improvements in attention and concentration. The training's effects are described partially as effective, but the evaluation is globally positive. The parents describe a tendency to be more engaged in intellectual work, to have more endurance and to evaluate better the ability to engage into cognitive processes and to take breaks.

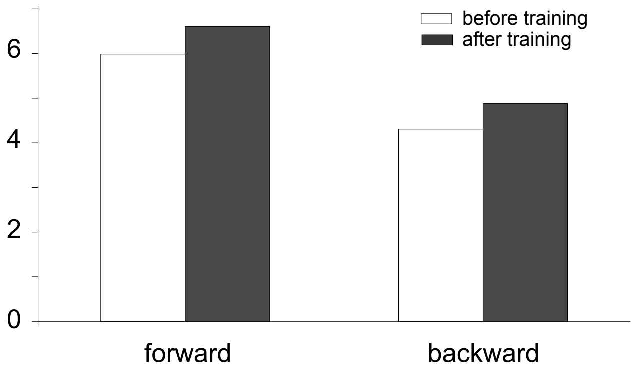


Figure 2: Digit Span of ADHD vs. CONTROL before and after cognitive remediation.

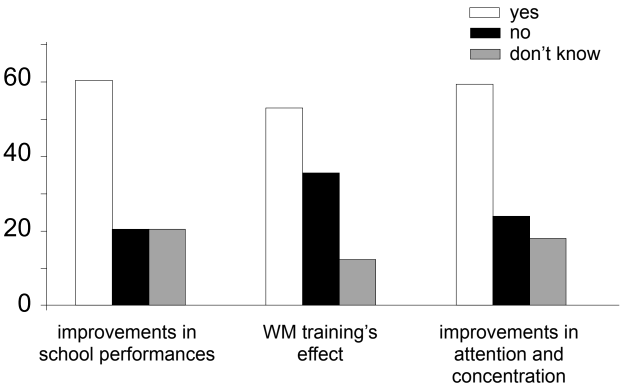


Figure 3: ADHD WM training perception: follow-up (2 months).

4 Discussion

The evolving field of research on ADHD has now moved beyond the search of a common core dysfunction towards a recognition of ADHD as a heterogeneous disorder of multiple neuropsychological deficits and hypothesized causal substrates. An increasing number of theoretical frameworks have incorporated an abnormal sensitivity response inhibition as to decision-making and working memory (WM) impairment as key issues in Attention deficit hyperactivity disorder (ADHD). The most effective and widely used treatments for ADHD are medication and behavior modification. New interventions without drug therapy, as cognitive remediation and neurofeedback, are now necessary to be explored as suggested by recents studies on WM in children and adolescents with ADHD [24, 25].

This study reports the effects of 5 weeks cognitive training (RoboMemo®, Cogmed) with fMRI paradigm by young adolescents with ADHD at the level of behavioral, neuropsychological and brain activations. These preliminary results are limited on neuropsychological data and on feedbacks from parents and adolescents subjects. These preliminary results tend to confirm a lower digit span in the ADHD group. This is particularly evident in the backward form, suggesting difficuties in verbal WM in accordance with Klingberg studies. The current data analysis do not show significant higher risky choices during the cognitive training and the high flood of WM. The cognitive training seems to improve the ability to decision-making through the frequency of stimuli without a stress of magnitude of penalty. The follow up inquiry showed that most of the parents of the ADHD subjects perceived a positive effect in the daily life. These preliminary results are promising and could provide benefits to the clinical practice, using for instance cognitive remediation training or groups focuses on executive functions and decision-making with ADHD children and adolescents.

Overall, the current literature leaves several questions unanswered: it is not clear to what extent WM training can affect behavior in ADHD patients or in normal subjects. While the evidence suggests that time preferences and risk preferences cannot be affected by WM training in normal subjects the current literature has not tested whether WM training in normal subjects affects attention and response inhibition in a more basic task such as the ANT or the go/no go task. Furthermore, while some evidence exists on how WM training changes brain activity in a WM task, nothing is known how it affects brain activity in other tasks. The previous paradigms don't allow a clear assessment of how WM training affects decision making in other areas such as risky choices or choices involving delayed rewards–all areas in which ADHD patients are known to differ from the rest of the population. Our future work will explore fMRI data which could provide an objective measure of the impact of WM and decision-making to others tasks by ADHD patients and controls. New researches are needed to investigate in novel ways how executive functions and cognitive training shape high-level cognitive processes as decision-making and WM, contributing to understand the association, or the separability, between distinct cognitive abilities. Cognitive training could be of direct relevance to improve abilities in cognitive process and choice evaluations in economic decision. Cognitive training inducing brain plasticity that could be a reciprocal interplay between behavior, cognition and brain biochemistry and should be relevant for psychiatry disorders as well applications in psychology and others domains, as for instance computer sciences.

Acknowledgments

The Authors acknowledge the support of the Swiss National Science Foundation grants 3200BO-118373, CR13I1-138032, and the Novartis Foundation of the scientific research. They wish to thank Professor Reto Meuli of the University Radiology Department of the CHUV and Professor Matthias Stuber of the Unit of the Centre d' imagerie biomédicale (CIBM) in Lausanne. The Authors thank also Dr. Alessandra Lintas of the Neuroheuristic Research Group at the Dept. of Information Systems of UNIL for helping with editing and layout of the manuscript.

References

[1] S V Faraone, J Sergeant, C Gillberg, and J Biederman. The worldwide prevalence of ADHD: is it an american condition? *World Psychiatry*, 2(2):104–113, Jun 2003.

[2] R C Kessler, L Adler, R Barkley, J Biederman, C K Conners, O Demler, S V Faraone, L L Greenhill, M J Howes, K Secnik, T Spencer, T B Ustun, E E Walters, and A M Zaslavsky. The prevalence and correlates of adult ADHD in the United States: results from the National Comorbidity Survey Replication. *Am J Psychiatry*, 163(4):716–723, Apr 2006.

[3] J Fayad and R de Graaf. Cross-national prevalence and correlates of adult attention-deficit hyperactivity disorder. *Brit J of Psychiatry*, 19:402–409, 2007.

[4] R A Barkley, M Fischer, L Smallish, and K Fletcher. Young adult outcome of hyperactive children: adaptive functioning in major life activities. *J Am Acad Child Adolesc Psychiatry*, 45(2):192–202, Feb 2006.

[5] R A Barkley. Global issues related to the impact of untreated attention-deficit/hyperactivity disorder from childhood to young adulthood. *Postgrad Med*, 120(3):48–59, Sep 2008.

[6] J Biederman, C R Petty, M Evans, J Small, and S V Faraone. How persistent is ADHD? A controlled 10-year follow-up study of boys with ADHD. *Psychiatry Res*, 177(3):299–304, May 2010.

[7] J Biederman and S V Faraone. Attention-deficit hyperactivity disorder. *Lancet*, 366(9481):237–248, Jul 2005.

[8] T W Frazier, E A Youngstrom, J J Glutting, and M W Watkins. ADHD and achievement: meta-analysis of the child, adolescent, and adult literatures and a concomitant study with college students. *J Learn Disabil*, 40(1):49–65, Jan-Feb 2007.

[9] H G Birnbaum, R C Kessler, S W Lowe, K Secnik, P E Greenberg, S A Leong, and A R Swensen. Costs of attention deficit-hyperactivity disorder (ADHD) in the US: excess costs of persons with ADHD and their family members in 2000. *Curr Med Res Opin*, 21(2):195–206, Feb 2005.

[10] F X Castellanos and R Tannock. Neuroscience of attention-deficit/hyperactivity disorder: the search for endophenotypes. *Nat Rev Neurosci*, 3(8):617–628, Aug 2002.

[11] E J Sonuga-Barke. Psychological heterogeneity in AD/HD–a dual pathway model of behaviour and cognition. *Behav Brain Res*, 130(1-2):29–36, Mar 2002.

[12] E J Sonuga-Barke. The dual pathway model of ad/hd: an elaboration of neuro-developmental characteristics. *Neurosci Biobehav Rev*, 27(7):593–604, Nov 2003.

[13] J T Nigg. Neuropsychologic theory and findings in attention-deficit/hyperactivity disorder: the state of the field and salient challenges for the coming decade. *Biol Psychiatry*, 57(11):1424–1435, Jun 2005.

[14] J T Nigg, E G Willcutt, A E Doyle, and E J Sonuga-Barke. Causal heterogeneity in attention-deficit/hyperactivity disorder: do we need neuropsychologically impaired subtypes? *Biol Psychiatry*, 57(11):1224–1230, Jun 2005.

[15] E J Sonuga-Barke and J Sergeant. The neuroscience of adhd: multidisciplinary perspectives on a complex developmental disorder. *Dev Sci*, 8(2):103–104, Mar 2005.

[16] R A Barkley. Behavioral inhibition, sustained attention, and executive functions: constructing a unifying theory of ADHD. *Psychol Bull*, 121(1):65–94, Jan 1997.

[17] B F Pennington and S Ozonoff. Executive functions and developmental psychopathology. *J Child Psychol Psychiatry*, 37(1):51–87, Jan 1996.

[18] M Sakagami, X Pan, and B Uttl. Behavioral inhibition and prefrontal cortex in decision-making. *Neural Netw*, 19(8):1255–1265, Oct 2006.

[19] A Diamond. Attention-deficit disorder (attention-deficit/ hyperactivity disorder without hyperactivity): a neurobiologically and behaviorally distinct disorder from attention-deficit/hyperactivity disorder (with hyperactivity). *Dev Psychopathol*, 17(3):807–825, 2005.

[20] M Luman, J Oosterlaan, and J A Sergeant. The impact of reinforcement contingencies on AD/HD: a review and theoretical appraisal. *Clin Psychol Rev*, 25(2):183–213, Feb 2005.

[21] A R Damasio. The somatic marker hypothesis and the possible functions of the prefrontal cortex. *Philos Trans R Soc Lond B Biol Sci*, 351(1346):1413–1420, Oct 1996.

[22] R Martinussen, J Hayden, S Hogg-Johnson, and R Tannock. A meta-analysis of working memory impairments in children with attention-deficit/hyperactivity disorder. *J Am Acad Child Adolesc Psychiatry*, 44(4):377–384, Apr 2005.

[23] E G Willcutt, A E Doyle, J T Nigg, S V Faraone, and B F Pennington. Validity of the executive function theory of attention-deficit/hyperactivity disorder: a meta-analytic review. *Biol Psychiatry*, 57(11):1336–1346, Jun 2005.

[24] P J Olesen, H Westerberg, and T Klingberg. Increased prefrontal and parietal activity after training of working memory. *Nat Neurosci*, 7(1):75–79, Jan 2004.

[25] T Klingberg, E Fernell, P J Olesen, M Johnson, P Gustafsson, K Dahlström, C G Gillberg, H Forssberg, and H Westerberg. Computerized training of working memory in children with ADHD–a randomized, controlled trial. *J Am Acad Child Adolesc Psychiatry*, 44(2):177–186, Feb 2005.

[26] M Slusarek, S Velling, D Bunk, and C Eggers. Motivational effects on inhibitory control in children with ADHD. *J Am Acad Child Adolesc Psychiatry*, 40(3):355–363, Mar 2001.

[27] R Drechsler, P Rizzo, and H C Steinhausen. Decision making with uncertain reinforcement in children with attention deficit/hyperactivity disorder (ADHD). *Child Neuropsychol*, 16(2):145–161, Mar 2010.

[28] Shane Frederick. Cognitive reflection and decision making. *J. Econ. Perspect.*, 19(4):25–42, 2005.

[29] S V Burks, J P Carpenter, L Goette, and A Rustichini. Cognitive skills affect economic preferences, strategic behavior, and job attachment. *Proc Natl Acad Sci U S A*, 106(19):7745–7750, May 2009.

[30] T Dohmen, A Falk, D Huffman, and U Sunde. Are risk aversion and impatience related to cognitive ability? *The American Economic Review*, 120(54):256–271, 2010.

[31] W K Bickel, R Yi, R D Landes, P F Hill, and C Baxter. Remember the future: working memory training decreases delay discounting among stimulant addicts. *Biol Psychiatry*, 69(3):260–265, 2011.

[32] S M Jaeggi, M Buschkuehl, J Jonides, and W J Perrig. Improving fluid intelligence with training on working memory. *Proc Natl Acad Sci U S A*, 105(19):6829–6833, May 2008.

[33] F McNab, A Varrone, L Farde, A Jucaite, P Bystritsky, H Forssberg, and T Klingberg. Changes in cortical dopamine D1 receptor binding associated with cognitive training. *Science*, 323(5915):800–802, Feb 2009.

[34] P J Olesen, J Macoveanu, J Tegnér, and T Klingberg. Brain activity related to working memory and distraction in children and adults. *Cereb Cortex*, 17(5):1047–1054, May 2007.

[35] A M Owen, A Hampshire, J A Grahn, R Stenton, S Dajani, A S Burns, R J Howard, and C G Ballard. Putting brain training to the test. *Nature*, 465(7299):775–778, Jun 2010.

[36] Y. Lecrubier, E. Weiler, T. Hergueta, P. Amorin, and J. P. Lépine. *Mini International Neuropsychiatric Interview (M.I.N.I.). French Version 5.0.0.* Hôpital de la Salpétrière, Paris, 1998.

[37] American Psychiatric Association, Washington, DC, USA. *American Psychiatric Association. Diagnostic and Statstical Manual of Mental Disorders*, fourth edition edition, 1994.

[38] C K Conners, G Sitarenios, J D Parker, and J N Epstein. The revised conners' parent rating scale (cprs-r): factor structure, reliability, and criterion validity. *J Abnorm Child Psychol*, 26(4):257–268, Aug 1998.

[39] C K Conners, K C Wells, J D Parker, G Sitarenios, J M Diamond, and J W Powell. A new self-report scale for assessment of adolescent psychopathology: factor structure, reliability, validity, and diagnostic sensitivity. *J Abnorm Child Psychol*, 25(6):487–497, Dec 1997.

[40] G A Gioia, P K Isquith, P D Retzlaff, and K A Espy. Confirmatory factor analysis of the behavior rating inventory of executive function (brief) in a clinical sample. *Child Neuropsychol*, 8(4):249–257, Dec 2002.

[41] P K Isquith, G A Gioia, and K A Espy. Executive function in preschool children: examination through everyday behavior. *Dev Neuropsychol*, 26(1):403–422, 2004.

[42] D. Wechsler. *WAI-R Manual*. The Psychological Corporation, New York, NY, USA, 1981.

[43] D. Wechsler. *Wechsler Intelligence Scale for children UK*. The Psychological Corporation Ltd Foots Cray, Sidcup, Kent, England, 1992.

[44] J. R. Stroop. Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18(6):643–662, 1935.

[45] C. J. Golden. *Stroop Color and Word Test: A Manual for Clinical and Experimental Uses*. Skoelting, Chicago, Illinois, USA, 1978.

[46] C M MacLeod. Half a century of research on the stroop effect: an integrative review. *Psychol Bull*, 109(2):163–203, Mar 1991.

[47] R P Kessels, E van den Berg, C Ruis, and A M Brands. The backward span of the Corsi Block-Tapping Task and its association with the WAIS-III Digit Span. *Assessment*, 15(4):426–434, Dec 2008.

[48] P. Zimmermann and B. Fimm. *The Test of Attentional Performance (TAP 2.2.)*. Psychologische Testsysteme, Herzogenrath, Germany, 2009.

[49] R. M. Reitan and D. Wolfson. *The Halstead–Reitan Neuropsycholgical Test Battery: Therapy and clinical interpretation*. Neuropsychological Press, Tucson, AZ, USA, 1985.

[50] J. Raven. *Coloured Progressive Matrices*. Oxford Psychological Press, Oxford, 1995.

[51] J Raven. The raven's progressive matrices: change and stability over culture and time. *Cogn Psychol*, 41(1):1–48, Aug 2000.

Towards Distributed Bayesian Estimation
A Short Note on Selected Aspects

K. Dedecius
Department of Adaptive Systems
Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
Pod Vodárenskou věží 4, Prague, 182 08, Czech Republic
dedecius@utia.cas.cz

V. Sečkárová
Department of Adaptive Systems
Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
Pod Vodárenskou věží 4, Prague, 182 08, Czech Republic
seckarov@utia.cas.cz

Abstract

The rapid development of ad-hoc wireless networks, sensor networks and similar calls for efficient estimation of common parameters of a linear or nonlinear model used to describe the operating environment. Therefore, the theory of collaborative distributed estimation has attained a very considerable focus in the past decade, however, mostly in the classical deterministic realm. We conjecture, that the consistent and versatile Bayesian decision making framework, whose applications range from the basic probability counting up to the nonlinear estimation theory, can significantly contribute to the distributed estimation theory.

The limited extent of the paper allows to address the considered problem only very superficially and shortly. Therefore, we are forced to leave the rigorous approach in favor of a short survey indicating the arising possibilities appealing to the non-Bayesian literature. First, we introduce the problem in a general Bayesian decision making domain and then narrow the scope to the estimation problem. In the ensuing parts, two mainstream approaches to common-objective distributed estimation are presented and the constraints imposed by the environment are studied.

1 Introduction

From the immense distributed decision making framework, we consider only the fully distributed decision problem (specifically parameter estimation) in which different agents (network nodes) obtain rather slightly different measurements from the environment. These measurements ideally obey the same distribution and differ only with respect to a realization of a noise variable. In this regard, the prescriptive methodology for designing agents proof of systematic error or bias is necessary. Furthermore, we assume that the agents have the same objective function. Our (practically unreachable) goal is to achieve the *general consensus on the decision*. In other words, if the decision is the evaluation of the posterior probability of some event, the goal is to achieve the state when all agents agree on it [22]. The solution has been proposed in [3]. Its time-dependent reformulation follows [22]:

Each time instant, the agents first communicate their own distributions among themselves and then update own distribution by reflecting the obtained distributions from the others. Following this procedure, the consensus is achieved as its limit case.

Since we deal with distributed estimation problem, it should be emphasized that our distributed decision problem differs from the team decision making [18],[19] and others. In our case, the *common consensus* on the estimate is the primary goal.

The theory of distributed estimation of an unknown common variable of interest has attained the prevailing focus in the last decade. The main cause was the increasing spatial complexity of large-scale ad-hoc wireless and sensor networks consisting of heterogeneous devices. Such an environment, more or less limited with respect to the energy, communication and processing resources, calls for efficient computational paradigms. The main tasks of interest in these networks, closely related to estimation, comprise in-network routing, signal processing, management, load balancing, sensor management, change point estimation etc.

From the Bayesian viewpoint, the theory of parameter estimation belongs to the decision theory. Suppose, that the considered network consists of $n \in \mathbb{N}$ nodes whose respective scalar or multidimensional measurements $y_1, \ldots, y_n$ are related to some unobservable quantity Θ, called parameter, some scalar or multidimensional input variables $u_1, \ldots, u_n$ and the task consists in estimation of Θ. This basically means, that the nodes seek the probability distribution of Θ given measurements and inputs, mostly in the form of a probability density function

$$f_\Theta\left(\Theta | y_0, y_1, \ldots, y_n, u_0, u_1, \ldots, u_n\right), \quad n \in \mathbb{N},$$

where y_0 and u_0 form the prior information, e.g. obtained from an expert, from past measurements or a noninformative prior is used.

The distributed systems can actively benefit from the higher number of participating sensors in the network and, potentially, from their technical heterogeneity, allowing to measure and compute with different performance according to the actual state of the observed reality.[1]

Some examples of distributed estimation problems comprise:

- collective estimation of a physical variable. This case is very important in sensor networks. Furthermore, it becomes popular in large physical experiments as well.
- fault tolerant systems with the voting circuit, in which several units collectively decide about a failure of a redundant device. Currently, the fuzzy approach dominates these solutions;
- classification networks, in which several nodes estimate the parameters of classifiers (represented, e.g., by beta-binomial or Dirichlet-multinomial models). This case is very important in bandwidth-limited networks;
- and many others.

From the communication and evaluation strategy, the estimation task in the distributed systems can be run in two different basic concepts, obviously influencing the network topology:

- The centralized approach in which the network nodes send their data to a dedicated unit responsible for computations;
- The decentralized approach in which all the network nodes posses and actively exploit own computing ability.

We will further describe them below.

[1] A particularly interesting case is the existence of the need of very precise measurements under several different conditions, preventing the user from using single measuring device.

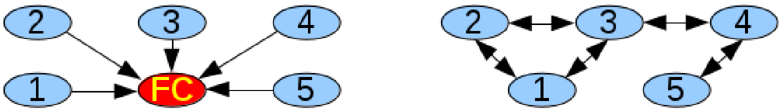


Figure 1: Principles of centralized and decentralized schemes. The centralized one (left) embodies a fusion center (FC) responsible for computation of estimates; in the decentralized concept (right) the computation task lies on the network nodes disseminating the available information, possibly partial. Remind, that the network topology of the decentralized scheme may differ from case to case.

2 Centralized and decentralized approaches

In this section, we describe the centralized and decentralized approaches to distributed parameter estimation, highlight the principal differences and mention several existing concepts. The purpose of this section is twofold: to induce contemplation on the pros and cons of the respective approaches and (maybe more importantly) to let the Bayesians take inspiration from the "disregarded deterministic world". Anyway, the good aspect is that in the Bayesian framework, the estimation mostly abstracts from the centralization or decentralization of information processing. In both cases, the basic methods can be the same. Some differences will arise with respect to the convergence properties of the estimators. Again, we stress that we focus only on the case of *distributed estimation of common parameter*.

3 Centralized approach

The centralized approach with a fusion center processing the measurements from the nodes in the network and potentially propagating the results back to them has appeared with the first occurrence of the distributed networks. This popular approach is widely used, e.g., in industrial applications, Internet services etc. However, it significantly suffers from high communication resources and high-availability (HA) demands to be able to transmit the data between the fusion center and the network nodes. The plethora of data is likely to saturate the (short term) memory and can lead to high system load. To some degree, these problems can be solved by data aggregation and quantization, e.g. [22] and references below.

On top of this, the fusion center represents a potential single point of failure (SPoF) requiring a special treatment, e.g., redundant hot and cold spare devices, data replications, graceful degradation systems etc., [28], which in turn increases the complexity of the whole system.

4 Decentralized approach

In the decentralized approach, the estimation is run directly in the nodes which share their information with other nodes. The decentralized approach can be further divided to incremental and diffusion approaches [4]. The former one is similar to the token-ring network topology, in which a closed cyclic path is performed, i.e., each node communicates with its neighbours within this path. In this setting, a failure of a link between two adjacent nodes can prevent the network from operation, hence the problem of SPoFs can be even worse. On the other hand, the diffusion approaches solve this issue by letting the nodes communicate within their closed neighbourhood, i.e., with their adjacent neighbours. In this case, the failure of a single link or node does not cause failure of the whole network, hence the estimation is preserved. It was shown, that the information propagating through the network leads in the limit case to the consensus [5], [4]. The diffusion approach possesses the adaptivity property that is very important for ad-hoc network, in which the topology of the network can dynamically change with time.

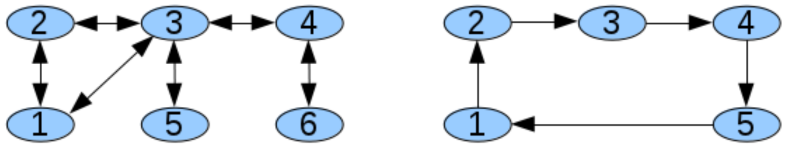


Figure 2: Two basic decentralized approaches. Left: diffusion approach, in which, e.g. node 3 communicates with nodes 1,2,4 and 5, while node 6 communicates only with node 4. Right: incremental approach (token-ring topology), the information sequentially circulates via all the network nodes which incorporate their contributions to it.

5 Communication constraints

The communication constraints still represent the most frequent restriction in distributed systems. In the centralized approach, the fusion center is often connected to other nodes via a backbone network, which enforces high bandwidth requirement. However, if the nodes are connected directly to the center via dedicated lines, the center must be able to effectively handle the incoming data traffic using a sort of an efficient switching protocol. This issue is far beyond the scope of the paper, we only reveal the need for as little traffic as possible.

In the decentralized estimation, the situation is much easier, since although the load of individual point-to-point communication lines remains the same, the high communication load typical for the fusion center is avoided. By using efficient network topology, the load can be decreased to a very low level, respecting the constraints of individual nodes.

Several possible communication strategies comprise:

- communication of all data, i.e., y_i, u_i and estimates $\hat{\Theta}$, possibly the parameters of distribution of the latter;;
- communication of sufficient statistics of respective distributions or, if applicable, non-sufficient statistics whose ancillary complements are naturally known to the network nodes;
- down to 1 bit communication strategy.

The communication of all data is a trivial case. The data quantization, i.e., the compression of data to be transmitted among the nodes in the network is an interesting option. Such a problem has been treated, for instance, in [8] and [1] for single node. For a decentralized distributed network, [13] restrict the local nodes to be data quantizers and develop the optimal design minimizing the estimation error. These results were further used by [2], [6], [9], [14], [16], [20], [27] and many others, dealing mostly with few or even one-bit messages. In this light, the assumption of possibility to communicate, for instance, sufficient statistics, so fundamental in the Bayesian framework, can simply fail. Therefore, it would be necessary to find out a way to fulfill the potential communication constraints.

6 Information fusion strategies

There exist several possible strategies for fusion of information obtained from nodes, irrespectively of the network scheme, topology or constraints. The Bayesian paradigm imposes constraints on entropy transform between the prior and posterior distribution, measured mostly by the information entropy (in terms of its maximization) or the Kullback-Leibler divergence and the cross-entropy (minimization). There are two main representation of the information from several information sources (i.e., network nodes), namely mixture, i.e. a convex combination of probability density functions [7], [10], [17], or a weighted likelihood [23], [24] and [25]. The latter has been proved suitable for distributed dynamic estimation in [21]. The mixture-based information treatment is a traditional and well-established way. The fusion of incompletely compatible probabilistic representations of information still represents a challenge. First steps towards this, exploiting the minimum cross-entropy principle, can be found, e.g. in [11].

7 Concluding remarks

We have outlined the possible future research trends towards the Bayesian distributed estimation of common parameter of interest using similar decision makers (estimators) with the same model. Unlike the traditional single problem oriented solutions employing mostly the non-stochastic methods, the Bayesian reasoning leads rather to a *methodology*, abstracting from a particular problem view. Being applicable to a large class of problems comprising, among others, dynamic estimation of least-squares problems, classification problems and many others, the distributed Bayesian framework only expects a suitable formulation of the problem. The application of the basic prescribed methods arising from the methodology can be done usually in a straightforward way.

The future research directions in the field of Bayesian distributed estimation will almost certainly deal with networks considerably constrained from specific aspects. This issue has been only recently dealt with from the non-Bayesian viewpoint, e.g., the energy-constrained networks in [12],[15],[26], the bandwidth constrained networks mentioned above etc.

Another very special issue is the design of intelligent nodes, able to agree on the form of information to be disseminated. However, this case is rather a part of the multi-agent systems (MAS) theory and definitely does not belong to the estimation theory.

Acknowledgement

This research has been supported by the Grant Agency of the Czech Republic, project number 102/08/0567.

References

[1] E. Ayanoglu. On optimal quantization of noisy sources. *Information Theory, IEEE Transactions on*, 36(6):1450–1452, November 1990.

[2] T. C. Aysal and K. E. Barner. Constrained decentralized estimation over noisy channels for sensor networks. *Signal Processing, IEEE Transactions on*, 56(4):1398–1410, April 2008.

[3] V. Borkar and P. Varaiya. Asymptotic agreement in distributed estimation. *Automatic Control, IEEE Transactions on*, 27(3):650–655, June 1982.

[4] F. S. Cattivelli, C. G. Lopes, and A. H. Sayed. Diffusion recursive Least-Squares for distributed estimation over adaptive networks. *IEEE Transactions on Signal Processing*, 56(5):1865–1877, May 2008.

[5] F. S. Cattivelli and A. H. Sayed. Diffusion LMS strategies for distributed estimation. *IEEE Transactions on Signal Processing*, 58(3):1035–1048, March 2010.

[6] Hao Chen and P. K. Varshney. Nonparametric One-Bit quantizers for distributed estimation. *Signal Processing, IEEE Transactions on*, 58(7):3777–3787, July 2010.

[7] Morris H. Degroot. *Optimal Statistical Decisions (Wiley Classics Library)*. Wiley-Interscience, April 2004.

[8] Y. Ephraim and R. M. Gray. A unified approach for encoding clean and noisy sources by means of waveform and autoregressive model vector quantization. *Information Theory, IEEE Transactions on*, 34(4):826–834, July 1988.

[9] Jun Fang and Hongbin Li. Distributed adaptive quantization for wireless sensor networks: From delta modulation to maximum likelihood. *Signal Processing, IEEE Transactions on*, 56(10):5246–5257, October 2008.

[10] M. Kárný, J. Böhm, T. V. Guy, L. Jirsa, I. Nagy, P. Nedoma, and L. Tesař. *Optimized Bayesian Dynamic Advising: Theory and Algorithms*. Springer, London, 2006.

[11] M. Kárný and T.V. Guy. Sharing of knowledge and preferences among imperfect Bayesian participants. In *Proc. of the NIPS Workshop 'Decision Making with Multiple Imperfect Decision Makers'*, 2010.

[12] Alexey Krasnopeev, Jin J. Xiao, and Zhi Q. Luo. Minimum energy decentralized estimation in a wireless sensor network with correlated sensor noises. *EURASIP J. Wirel. Commun. Netw.*, 2005:473–482, September 2005.

[13] Wai-Man Lam and A. R. Reibman. Design of quantizers for decentralized estimation systems. *Communications, IEEE Transactions on*, 41(11):1602–1605, November 1993.

[14] Junlin Li and G. AlRegib. Rate-Constrained distributed estimation in wireless sensor networks. *Signal Processing, IEEE Transactions on*, 55(5):1634–1643, May 2007.

[15] Junlin Li and G. AlRegib. Distributed estimation in Energy-Constrained wireless sensor networks. *Signal Processing, IEEE Transactions on*, 57(10):3746–3758, October 2009.

[16] Zhi-Quan Luo. Universal decentralized estimation in a bandwidth constrained sensor network. *Information Theory, IEEE Transactions on*, 51(6):2210–2219, June 2005.

[17] G. J. McLachlan and D. Peel. *Finite mixture models*. Wiley series in probability and statistics: Applied probability and statistics. Wiley, 2000.

[18] Michael and Bacharach. Interactive team reasoning: A contribution to the theory of cooperation. *Research in Economics*, 53(2):117 – 147, 1999.

[19] R. Radner. Costly and bounded rationality in individual and team decision-making. *Industrial and Corporate Change*, 9(4):623, 2000.

[20] A. Ribeiro and G. B. Giannakis. Bandwidth-constrained distributed estimation for wireless sensor networks-part II: unknown probability density function. *Signal Processing, IEEE Transactions on*, 54(7):2784–2796, July 2006.

[21] V. Sečkárová and K. Dedecius. Distributed Bayesian Diffused Estimation. In *IFAC SYSID 2012, submitted.* IFAC, 2012.

[22] J. N. Tsitsiklis and M. Athans. Convergence and asymptotic agreement in distributed decision problems. 21:692–701, December 1982.

[23] X. Wang. Asymptotic properties of maximum weighted likelihood estimators. *Journal of Statistical Planning and Inference*, 119(1):37–54, January 2004.

[24] X. Wang. Approximating Bayesian inference by weighted likelihood. *Can J Statistics*, 34(2):279–298, 2006.

[25] Xiaogang Wang and James Zidek. Derivation of mixture distributions and weighted likelihood function as minimizers of KL-divergence subject to constraints. *Annals of the Institute of Statistical Mathematics*, 57:687–701, 2005.

[26] Jwo-Yuh Wu, Qian-Zhi Huang, and Ta-Sung Lee. Minimal energy decentralized estimation via exploiting the statistical knowledge of sensor noise variance. *Signal Processing, IEEE Transactions on*, 56(5):2171–2176, May 2008.

[27] J. J. Xiao and Z. Q. Luo. Decentralized estimation in an inhomogeneous sensing environment. *Information Theory, IEEE Transactions on*, 51(10):3564–3575, October 2005.

[28] B.Y. Zhao, J. Kubiatowicz, and A.D. Joseph. Tapestry: An infrastructure for fault-tolerant wide-area location and routing. *Computer*, 74(11-20):46, 2001.

Variational Bayes in Distributed Fully Probabilistic Decision Making

Václav Šmídl and Ondřej Tichý,
Institute of Information Theory and Automation,
Pod vodárenskou věží 4, Prague 8, Czech republic,
{smidl,otichy}@utia.cas.cz

Abstract

We are concerned with design of decentralized control strategy for stochastic systems with global performance measure. It is possible to design optimal centralized control strategy, which often cannot be used in distributed way. The distributed strategy then has to be suboptimal (imperfect) in some sense. In this paper, we propose to optimize the centralized control strategy under the restriction of conditional independence of control inputs of distinct decision makers. Under this optimization, the main theorem for the Fully Probabilistic Design is closely related to that of the well known Variational Bayes estimation method. The resulting algorithm then requires communication between individual decision makers in the form of functions expressing moments of conditional probability densities. This contrasts to the classical Variational Bayes method where the moments are typically numerical. We apply the resulting methodology to distributed control of a linear Gaussian system with quadratic loss function. We show that performance of the proposed solution converges to that obtained using the centralized control.

1 Introduction

Complexity of large-scale uncertain systems, such as traffic light signalization in urban areas, prevents effective use of centralized design of control strategy. The technology of multi-agent systems [1] offers technical background how to build a distributed control system. The mainstream multi-agent theory is concerned with deterministic systems for which the majority of results on communication protocols and negotiation strategies are established. As a result, many stochastic problems are converted into deterministic formulation and solved as such. This is typical e.g. in design of distributed traffic light control, where the certainty equivalence assumption is used in all agents [2].

Design methodologies for optimal control strategies of large-scale decentralized stochastic systems are available, e.g. [3], however, the complexity of the decision maker is rather high. In this paper, we propose to design suboptimal (imperfect) decision makers by imposition of additional restrictions within an established centralized design methodology. Specifically, we focus on the theory of Fully Probabilistic Control Design (FPD) [4, 5] for centralized control strategies. This theory is based on minimization of Kullback-Leibler divergence (KLD) [6] and it has been extended to multiple participants using heuristic arguments [7, 8]. An independently developed variant of this approach was used in multi-agent setup in [9]. In this paper, we enforce distribution of control between decision makers via the constraint of conditional independence. Minimization of Kullback-Leibler divergence under this constraint is well known as the Variational Bayes approach [10, 11]. Generalization of these results yields a design methodology of approximate decision makers that are capable to design their own control strategy using probabilistic moments obtained from their neighbors.

We study two computation schemes in this contribution. The first scheme allows unlimited communication with small messages. The second scheme allows much lower number of messages, however, the messages contain much more information than in the first case. In both cases, the Variational Bayes approach is capable to compute approximate results in limited time depending on the number of iterations.

2 Review of Centralized Fully Probabilistic Design

Consider a probabilistic model of a stochastic system

$$y_t \backsim f(y_t|u_t, d_{t-m:t-1}), \tag{1}$$

where symbol $y \backsim f$ denotes that y is a realization from probability density f; vector y_t denotes system output at discrete time t; vector u_t is system input; $d_t = [y_t', u_t']'$ is an aggregation of output and input, where $(.)'$ denotes a transposition of vector or matrix; and $d_{t-m:t-1} = [d_{t-m}, \dots, d_{t-1}]$ is a matrix of the last m observation vectors. Our aim is to design a probabilistic control strategy $f\left(u_t|d_{1:t-1}\right)$ such that the closed loop behavior is as close to the desired behavior as possible.

The Fully Probabilistic Design is based on probabilistic description of the desired behavior represented by the target (ideal) probability density, ${}^I f(d_{1:t+h})$, which expresses its aim and constraints. Closeness of the real and the target behavior is measured by the Kullback-Leibler divergence. The optimal control strategy on a horizon of length h is then found recursively for $\tau = t+h, \dots, t+1$,

$$\begin{aligned} {}^o f(u_\tau|d_{1:\tau-1}) &= \arg\min_{f(u_\tau|d_{1:\tau-1})} KLD\left[f(d_{t+1:t+h})||{}^I f(d_{t+1:t+h})\right], &(2)\\ &= \arg\min_{f(u_\tau|d_{1:\tau-1})} E_{f(d_\tau|d_{1:\tau-1})}\left[\ln\frac{f(d_{t+1:t+h})}{{}^I f(d_{t+1:t+h})}\right], &(3)\end{aligned}$$

where $E_{f(x)}(.)$ is the expected value of the argument with respect to probability density $f(x)$; it is abbreviated as $E_{f(x)}(x) \equiv E(x)$ when no confusion can arise. $KLD(.||.)$ is the Kullback-Leibler divergence between the first and the second argument. The optimal solution can be found in the following form, [12]:

$${}^o f\left(u_\tau|d_{1:\tau-1}\right) = {}^I f(u_\tau|d_{1:\tau-1})\frac{\exp[-\omega(u_\tau, d_{1:\tau-1})]}{\gamma(d_{1:\tau-1})}. \tag{4}$$

Here, functions $\omega(.)$ and $\gamma(.)$ are recursively evaluated as

$$\begin{aligned} \omega(u_\tau, d_{1:\tau-1}) &= E_{f(y_\tau|u_\tau,d_{1:\tau-1})}\left(\ln\frac{f(y_\tau|u_\tau, d_{1:\tau-1})}{\gamma(d_{1:\tau}){}^I f(y_\tau|u_\tau, d_{1:\tau-1})}\right), &(5)\\ \gamma(d_{1:\tau-1}) &= \int {}^I f(u_\tau|d_{1:\tau-1})\exp[-\omega(u_\tau, d_{1:\tau-1})]\,du_\tau, &(6)\end{aligned}$$

initialized at time $\tau = t+h$ as $\gamma(d_{1:t+h}) = 1$.

2.1 Special case of Linear Quadratic design

Linear Quadratic Gaussian (LQG) control arise as a special case of FPD (4)–(6), when both the model and the target probability densities are Gaussian with linear function of their mean value:

$$\begin{aligned} f(y_t|u_t, d_{1:t-1}) &= \mathcal{N}(\Theta\psi_t, R), &(7)\\ {}^I f(y_t, u_t|d_{1:t-1}) &= \mathcal{N}\left(\left[\begin{array}{c}\overline{y}_t\\ \overline{u}_t\end{array}\right], \left[\begin{array}{cc}Q_y & 0\\ 0 & Q_u\end{array}\right]\right). &(8)\end{aligned}$$

Here, $\mathcal{N}(\mu, \Sigma)$ denotes Gaussian probability density with mean value μ and covariance Σ; Θ is a matrix of known parameters; ψ_t is a vector composed from an arbitrary combination of elements of $y_{t-m:t-1}$ and $u_{t-m:t}$, and any deterministic transformation of these elements.

Substitution of (8) into (5) at $\tau = t+h$, i.e. $\gamma(d_{1:t+h}) = 1$, yields:

$$\begin{aligned} \omega(u_\tau, d_{1:\tau-1}) &= \frac{1}{2}\left[\ln(Q_y R^{-1}) - n_y + tr(RQ_y^{-1}) + (\Theta\psi_\tau - \overline{y}_\tau)' Q_y^{-1}(\Theta\psi_\tau - \overline{y}_\tau)\right], \qquad (9) \\ &= [\psi'_\tau, 1]\Psi_\tau[\psi'_\tau, 1]' \qquad (10) \end{aligned}$$

where n_y denotes dimension of vector y_τ. Note that the first three terms in $\omega(.)$ are independent of u_τ and y_τ making them irrelevant to this time step. Evaluation of probability ${}^o f(u_\tau|\phi_\tau)$ from (4) is achieved by reordering the quadratic form in (10) into

$$[\psi'_\tau, 1]\Psi_\tau[\psi'_\tau, 1]' = [u_\tau, \phi'_\tau, 1]\Psi_{\omega,\tau}[u_\tau, \phi'_\tau, 1]', \qquad (11)$$

where u_τ was extracted from ψ_τ (the rest of the elements from ψ_τ are in vector with time-delayed values, ϕ_τ, related to the time τ), and $\Psi_{\omega,\tau}$ is composed of the same elements as Ψ_τ in adapted order with respect to vector $[u_\tau, \phi'_\tau, 1]$. Since (8) is independent in y_τ and u_τ, the marginal on u_τ can be written as

$$f(u_\tau|d_{1:\tau-1}) \propto \exp\left(-\frac{1}{2}[u_\tau, \phi'_\tau, 1]\Psi_{u,\tau}[u_\tau, \phi'_\tau, 1]'\right), \quad \Psi_{u,\tau} = \left[\begin{array}{ccc} Q_u^{-1} & 0 & Q_u\overline{u}_\tau \\ 0 & 0 & 0 \\ \overline{u}'_\tau Q_u^{-1} & 0 & \overline{u}_\tau Q_u^{-1}\overline{u}_\tau \end{array}\right].$$

The joint probability density (4) is then a quadratic form (11) with kernel $\Psi_{f,\tau} = \Psi_{\omega,\tau} + \Psi_{u,\tau}$. The kernel can be decomposed using Cholesky factorization into $\Psi_{f,\tau} = L_\tau L'_\tau$ where lower triangular matrix L_τ is decomposed into $L_\tau = \begin{pmatrix} \Upsilon_\tau & 0 \\ \Omega_\tau & \Lambda_\tau \end{pmatrix}$, with Υ_τ being triangular matrix of the same dimension as u_τ. Probability density (4) has form

$${}^o f(u_\tau|\phi_\tau) = \mathcal{N}(-(\Upsilon'_\tau)^{-1}\Omega_\tau[\phi'_\tau, 1]', (\Upsilon_\tau\Upsilon'_\tau)^{-1}). \qquad (12)$$

and the remainder

$$\gamma(d_{1:\tau-1}) = \exp\left(-\frac{1}{2}[\phi'_\tau, 1]\Lambda_\tau\Lambda'_\tau[\phi'_\tau, 1]'\right). \qquad (13)$$

The recursion from $\tau = t+h$ to t reveals the same quadratic forms with the exception that there are additional element in $\Psi_{f,\tau}$ from function $\gamma(d_{1:\tau-1})$.

The mean value of (12), i.e. $\hat{u}_\tau = -(\Upsilon'_\tau)^{-1}\Omega_\tau[\phi'_\tau, 1]'$, is equivalent to LQG designed strategy with loss function given by the quadratic form from (9) in exp(.) [4].

3 Distributed FPD via Variational Bayes

Consider a case where (1) describes a complex system, with vector inputs $u_t = [u_{1,t}, \ldots, u_{n,t}]$, where vectors $u_{i,t}$, $i = 1, \ldots, n$ are logically separated so that they represent independent decision makers. Without any additional assumptions on the model (1), solution (2) would be a complex probability density with no guide how to implement it in a distributed way.

As a first step to decentralization of the control strategy, we impose the restriction of conditional independence of control inputs

$$f(u_t|.) = \prod_{i=1}^{n} f(u_{i,t}|.), \forall t. \qquad (14)$$

If the solution is in this form, each decision maker can handle its own inputs via $f(u_{i,t}|\cdot)$. The task is to find a way how to design it.

We repeat minimization (3), under constraint (14)

$$\prod_{i=1}^{n} {}^o f(u_{i,\tau}|d_{1:\tau-1}) = \arg\min_{\prod_i f(u_{i,\tau}|\cdot)} E_{f(d_\tau|d_{1:\tau-1})}\left[\ln\frac{f(d_{t+1:t+h})}{{}^I f(d_{t+1:t+h})}\right]. \qquad (15)$$

Using the chain rule of probability calculus and definitions (5)–(6) we obtain

$$\prod_{i=1}^{n} {}^o f(u_{i,\tau}|d_{1:\tau-1}) = \arg\min_{\prod_i f(u_{i,\tau}|\cdot)} KLD\left[f(u_\tau|d_{1:\tau-1})||{}^o f(u_\tau|d_{1:\tau-1})\right]. \qquad (16)$$

Minimum of (16) is well known from the Variational Bayes method [11] to satisfy the following set of conditions:

$$ {}^{o}f(u_{i,\tau}|d_{1:\tau-1}) \propto \exp\left(E_{f(u_{/i,\tau}|d_{1:\tau-1})}\left[\ln {}^{o}f(u_{\tau}|d_{1:\tau-1})\right]\right), i = 1,\dots,n. \tag{17} $$

Here, $u_{/i,\tau}$ denotes a subset of elements of vector u_τ without the element $u_{i,\tau}$, i.e. $u_{/i,\tau} = [u_{1,\tau},\dots,u_{i-1,\tau},u_{i+1,\tau},\dots,u_{n,\tau}]$, and $\propto$ is equality up to normalizing constant.

Substitution of (4) into (17) at each step on the horizon, $\tau = t+h,\dots,t+1$, yields the following set of implicit equations for $i = 1,\dots,n$,

$$ {}^{o}f(u_{i,\tau}|d_{1:\tau-1}) \propto \exp\left(E_{f(u_{/i,\tau}|d_{1:\tau-1})}\left(\ln {}^{I}f(u_{\tau}|d_{1:\tau-1}) - \omega(u_\tau, d_{1:\tau-1})\right)\right), \tag{18} $$

The normalizing constant of (18) is

$$ \gamma_i(d_{1:\tau-1}) = \int \exp\Big(E_{f(u_{/i,\tau}|d_{1:\tau-1})}\big[\ln {}^{I}f(u_{\tau}|d_{1:\tau-1}) - \omega(u_\tau, d_{1:\tau-1})\big]\Big)\mathrm{d}u_{i,\tau}, \tag{19} $$

hence $\gamma(d_{1:\tau-1})$ required in (5) of the previous step factorizes into $\gamma(d_{1:\tau-1}) = \prod_{i=1}^{n}\gamma_i(d_{1:\tau-1})$.

Typically, set (18) does not have a closed form solution and must be solved iteratively using the iterative VB (IVB) algorithm. It has been shown that the IVB algorithm monotonically decrease the KLD in each iteration and thus converging to a local minimum [13].

Note that $d_{1:\tau-1}$ in $f(u_{i,\tau}|d_{1:\tau-1})$ are symbolic random variables. This contrasts to the typical application of the Variational Bayes where $d_{1:\tau-1}$ are measured data.

3.1 Special case of LQG

For the special case of linear Gaussian system discussed in Section 2.1, the Variational Bayes method [11] is to be applied to Gaussian probability density (12) with logarithm

$$ \ln f(u_\tau|d_{1:\tau-1}) = c - \frac{1}{2}(u_\tau - (\Upsilon_\tau')^{-1}\Omega_\tau[\phi_\tau',1]')'(\Upsilon_\tau\Upsilon_\tau')^{-1}(u_\tau - (\Upsilon_\tau')^{-1}\Omega_\tau[\phi_\tau',1]') \tag{20} $$

$$ = c - \frac{1}{2}[u_\tau',\phi_\tau',1]\Phi_\tau[u_\tau',\phi_\tau',1]', \tag{21} $$

Here, we use the same notation as in the previous section for ϕ_τ, Υ_τ, and Ω_τ, $c = \ln|\Upsilon_\tau|$ which is independent of control action u_τ, and Φ_τ is the kernel of quadratic form (21). For simplicity, we consider partitioning $u_\tau = [u_{1,\tau}, u_{2,\tau}]'$, generalization to n partitions is straightforward. The expected value of (20) with respect to $f(u_{2,\tau}|d_{1:\tau-1})$ is again a quadratic form

$$ E_{f(u_{2,\tau}|d_{1:\tau-1})}(\ln f(u_\tau|d_{1:\tau-1})) = E_{f(u_{2,\tau}|d_{1:\tau-1})}\left(c - \frac{1}{2}[u_{2,\tau},\zeta_\tau]\begin{bmatrix}\Phi_{uu,\tau} & \Phi_{u\zeta,\tau}\\ \Phi_{\zeta u,\tau} & \Phi_{\zeta\zeta,\tau}\end{bmatrix}[u_{2,\tau}',\zeta_\tau]'\right) \tag{22} $$

$$ = c - \frac{1}{2}[E_{f(u_{2,\tau}|d_{1:\tau-1})}(u_{2,\tau}\Phi_{uu,\tau}u_{2,\tau}) + \zeta_\tau\Phi_{\zeta u,\tau}E(u_{2,\tau}) + E(u_{2,\tau})\Phi_{u\zeta,\tau}\zeta_\tau + \zeta_\tau\Phi_{\zeta\zeta,\tau}\zeta_\tau] \tag{23} $$

$$ = c - \frac{1}{2}\zeta_\tau\overline{\Phi}_{u_1,\tau}\zeta_\tau, \tag{24} $$

where $\Phi_{uu,\tau}, \Phi_{u\zeta,\tau}, \Phi_{\zeta\zeta,\tau}$ are composed of elements of Φ_τ restructured to match the new decomposition of the $[u_\tau,\phi_\tau',1]$ to $u_{2,\tau}$, $\zeta_\tau = [u_{1,\tau},\phi_\tau',1]$ and $\overline{\Phi}_{u_1,\tau}$ is given by reordering to match the quadratic form in ζ_τ.

Note that (24) is equivalent to (10) and the control law can be obtained using the same derivation that lead to (12). In this case

$$ f(u_{1,\tau}|d_{1:\tau-1}) = \mathcal{N}(Q_{1,\tau}[\phi_\tau',1]',\sigma_{1,\tau}), \tag{25} $$

$$ f(u_{2,\tau}|d_{1:\tau-1}) = \mathcal{N}(Q_{2,\tau}[\phi_\tau',1]',\sigma_{2,\tau}), \tag{26} $$

Algorithm 1 DP-VB variant of the distributed control design.

```
Off-line:
Choose control horizon $h$, target probability density ${}^I f\left(d_{1:t+h}\right)$, and initial value of
$f^{(0)}(u_{i,\tau}|\cdot)$ for each decision maker $i = 1 \dots n$.
On-line:
At each time $t$, for each decision-maker $i$, do:
    1. For each $\tau = t+h, t+h-1, \dots, t$ do
        (a) Start negotiation with counter $j = 1$, and initial guess $f^{(0)}(u_{i,\tau}|\cdot)$,
        (b) Compute moments required by the neighbors and communicate them,
        (c) Compute $j$th value of control strategy $f^{(j)}(u_{i,\tau}|\cdot)$ using moments obtained from
            the neighbors,
        (d) If the strategy convergence is not reached and $j < j_{IVB}$, increase $j$ and goto
            (a), stop otherwise.
    2. Apply designed control action $u_{i,t}$ from the converged strategy.
```

where $\sigma_{i,\tau}$ is given using Cholesky decomposition of $\overline{\Phi}_{u_i,\tau}$ in the same form as in (12) and the second line follows from equivalent derivation for $u_{2,\tau}$. Now, we can formulate the necessary moments for substitution into (22):

$$E_{f(u_{i,\tau}|d_{1:\tau-1})}(u_{i,\tau}) = Q_{i,\tau}[\phi'_\tau, 1]', \tag{27}$$

$$E_{f(u_{i,\tau}|d_{1:\tau-1})}(u_{i,\tau}\Phi_{uu,\tau}u'_{i,\tau}) = Q_{i,\tau}[\phi'_\tau, 1]\Phi_{uu,\tau}[\phi'_\tau, 1]'Q'_{i,\tau} + \Phi_{uu,\tau}\sigma_{i,\tau}. \tag{28}$$

This finalizes the list of results that are necessary to run the IVB algorithm in each time step of the horizon $\tau = t+h, \dots, t+1$. This variant will be denoted as **DP-VB** algorithm, Algorithm 1.

3.2 Alternative evaluations

Note that the set of conditions (18) has to be met for each time of the horizon, τ. Put together, we may interpret it as a set of $n \times (h+1)$ conditions of optimality. If the control strategies $f(u_{i,\tau}|\cdot)$ were conditionally independent from $f(u_{i,\tau-1}|\cdot)$, then the iterations could be performed in any order and still guaranteed to converge to a local minimum. This would be a great property since it would allow asynchronous communication between the decision makers, and guarantee robustness against lost messages. However, this is not automatically guaranteed due to dependence $f(u_{i,\tau}|u_{i,\tau-1})$. Therefore, a change of order of the time index can lead to an increase of the KL divergence within one iteration due to inaccurate $\gamma(d_{1:\tau-1})$ from (19). However, similar difficulty arise in the case of on-line variational Bayes and the convergence is still guaranteed by means of stochastic approximations [14]. The only drawback is slower convergence in comparison to the standard IVB algorithm. We conjecture that it is also the case in our approach.

If our conjecture holds, then we may change the order of dynamic programming and IVB iterations. Specifically, each decision maker first exchange messages about expected values $[Q_{i,t}, \dots, Q_{i,t+h}, \sigma_{i,t}, \dots, \sigma_{i,t+h}]$ with its neighbors and then designs its strategy using backward evaluation (5), see Algorithm 2 for details. The new moments are send to the neighbors for the next iterations. This algorithm will be denoted as **VB-DP**.

4 Example

Consider the following 3-output 2-input system:

$$f(y_t|\psi_t, \Sigma) = \mathcal{N}(\Theta\psi_t, \Sigma_y), \tag{29}$$

where

$$y_t = [y_{1,t}, y_{2,t}, y_{3,t}]', \qquad \psi_t = [y_{1,t-1}, y_{2,t-1}, y_{3,t-1}, u_{1,t}, u_{2,t}, u_{1,t-1}, u_{2,t-1}]',$$

Algorithm 2 VB-DP variant of the distributed control design.

```
Off-line:
Choose control horizon $h$, target probability density ${}^{I}f(d_{1:t+h})$, and initial value of
$f^{(0)}(u_{i,\tau}|\cdot)$ for each decision maker $i = 1 \dots n$.
On-line:
At each time $t$, for each decision-maker $i$, do:
    1. Start negotiation with counter $j = 1$, and $f^{(0)}(u_{i,\tau}|\cdot)$.
    2. Compute $j$th value of control strategy $f^{(j)}(u_{i,\tau}|\cdot)$ on the whole horizon $\tau = t+h, \dots t$
       using moments obtained from the neighbors, evaluate moments required by the
       neighbors and communicate them,
    3. If the strategy convergence is not reached and $j < j_{IVB}$, increase $j$ and goto 2, stop
       otherwise.
    4. Apply designed control action $u_{i,t}$ from the converged strategy.
```

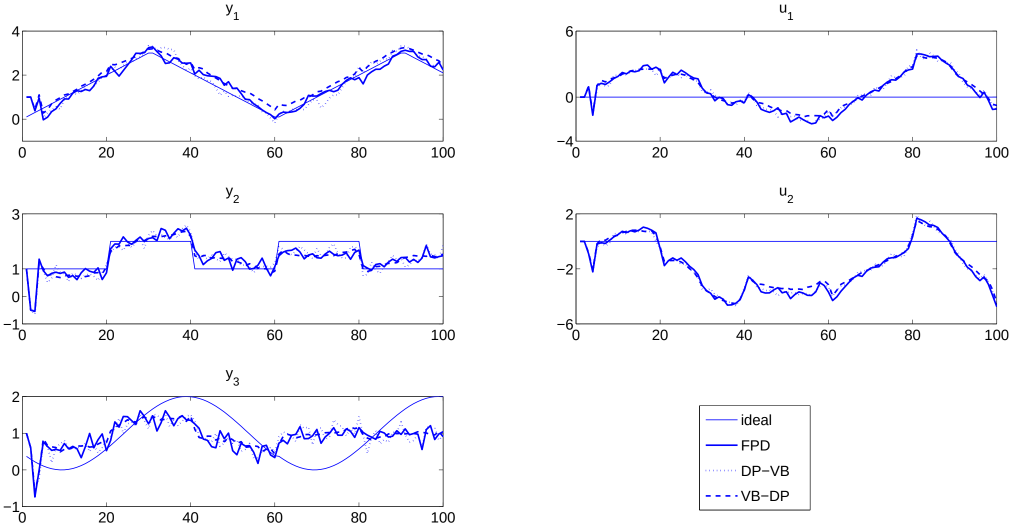


Figure 1: Example run of the controlled system. The target values for the inputs and the outputs are displayed in thin full line. The typical realization of outputs and inputs for all tested algorithms are also displayed for illustration.

$$\Theta = \begin{bmatrix} 0.8 & 0.2 & 0 & -0.3 & 0.4 & 0 & 0 & 0 \\ -0.2 & 0.5 & -0.8 & 0.2 & 0.5 & -0.2 & -0.5 & 0 \\ 0 & 1.1 & -0.5 & 0 & 0 & -0.2 & 0.3 & 0 \end{bmatrix}. \tag{30}$$

The target probability densities are

$$ {}^{I}f(y_t) = \mathcal{N}\left(\begin{bmatrix} \overline{y}_{1,t} \\ \overline{y}_{2,t} \\ \overline{y}_{3,t} \end{bmatrix}, \begin{bmatrix} 0.01 & & \\ & 0.01 & \\ & & 0.01 \end{bmatrix} \right), \quad {}^{I}f(u_t) = \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 100 & \\ & 100 \end{bmatrix} \right), \tag{31}$$

with values of $\overline{y}_{1,t}, \overline{y}_{2,t}, \overline{y}_{3,t}$ displayed in Fig 1 (solid lines). The choice of diagonal covariance matrices in (31) allows the convergence of algorithms from Section 3 to the centralized solution, Section 2.

Three control strategies were tested:

FPD: as the centralized strategy (Section 2.1),

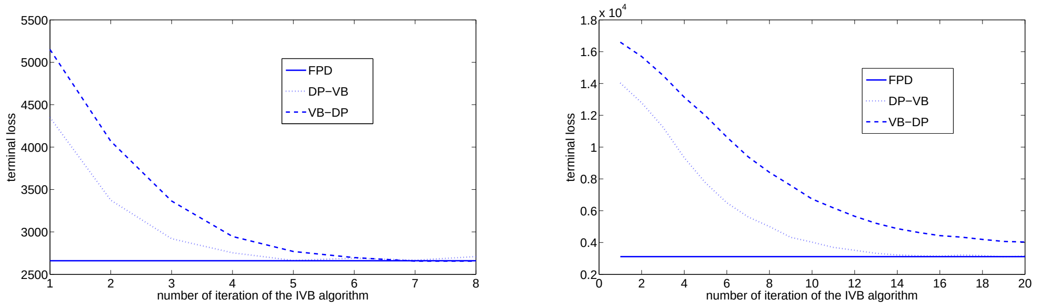


Figure 2: Convergence of the terminal loss (the sum of differences from target values) of the decentralized DP-VB and VB-DP algorithms to the terminal loss of the centralized FPD solution as a function of the number of IVB iterations for two variants of system parameters.

DP-VB: decentralized evaluation of the FPD control via multiple VB algorithms, one at each time τ on the horizon (Section 3.1),

VB-DP: decentralized evaluation of the FPD control via a single the VB algorithm on the whole horizon (Section 3.2).

A comparative Monte Carlo study with 15 runs of the system with parameter (30) was performed to establish convergence of the decentralized strategy design to the centralized one. An example run of the controlled system is shown in Fig. 1. Results of the study are displayed in Fig. 2 via dependence of the terminal loss on the number of iterations of the IVB algorithm, j_{IVB}. Note that the results converged to the centralized solution after a few iterations; the full convergence is allowed using diagonal covariance matrices in (31). As expected, the DP-VB variant converges faster than the VB-DP. Suitability of each strategy than depends on the quality of communication between agents. The VB-DP algorithm may be attractive especially for systems with higher latency in communication.

The difference is even more visible on a more demanding system with parameters

$$\Theta = \begin{bmatrix} 0.8 & 0.2 & 0.5 & -0.3 & 0 & 0.4 & 0 & 0 \\ -0.2 & 0.5 & -0.2 & 0.2 & -0.2 & 0.5 & -0.5 & 0 \\ 0.5 & 1.1 & -0.5 & 0 & -0.2 & 0 & 0.3 & 0 \end{bmatrix}. \tag{32}$$

The results of the same Monte Carlo experiment for the new value of parameter Θ are displayed in Fig. 2, right. While the DP-VB algorithm reaches performance of the centralized FPD after 14 iterations, the VB-DP algorithms requires more than 20 iterations to converge. The number of iterations required to reach the centralized solution is rather high, since the IVB algorithm was initialized with $f^{(0)}(u_{i,t}|\cdot) = {}^{I}f(u_{i,t})$ for both variants. The purpose of this choice was to verify if the algorithm converges to the correct solution even from poor initial conditions.

5 Conclusion

The presented methodology for design of approximate decision makers is based on fully probabilistic control and decentralization is achieved by imposing conditional independence between control inputs. The general method yields two principle outputs: (i) an iterative algorithm that is known to systematically decrease the loss function, and (ii) the moments that needs to be exchanged to achieve optimum performance. Under the condition of diagonal covariance matrices of target probability densities, the simulation results suggest that the decentralized control is able to reach the same performance as the centralized one. This was achieved at the price of all decision makers having full model of the system and intensive negotiation with high volume of communication. We have shown in simulation that the intensity of communication can be lowered by an alternative order of evaluation and communication. The Variational Bayes approach can cope with limited computational time, the quality of the solution depends on the number of iterations in the IVB algorithm.

Further simplifications can be achieved by imposing additional restrictions (e.g. in the form of conditional independence) on the solution.

Acknowledgment

Support of grants MŠMT 1M0572 and TAČR TA01030603 is gratefully acknowledged.

References

[1] G. Weiss, ed., *Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence*. The MIT Press, 2000.

[2] R. Negenborn, B. De Schutter, and J. Hellendoorn, "Multi-agent model predictive control for transportation networks: Serial versus parallel schemes," *Engineering Applications of Artificial Intelligence*, vol. 21, no. 3, pp. 353–366, 2008.

[3] M. Hutter, *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Berlin: Springer, 2005.

[4] M. Kárný, "Towards fully probabilistic control design," *Automatica*, vol. 32, no. 12, pp. 1719–1722, 1996.

[5] M. Kárný and T. Kroupa, "Axiomatisation of fully probabilistic design," *Information Sciences*, 2011. DOI=10.1016/j.ins.2011.09.018.

[6] S. Kullback and R. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, pp. 79–87, 1951.

[7] V. Šmídl and J. Andrýsek, "Merging of multistep predictors for decentralized adaptive control," in *Proc. of the American Control Conference*, (Seattle, USA), 2008.

[8] V. Šmídl, "On adaptation of loss functions in decentralized adaptive control," in *Proc. of the 12th IFAC symposium on Large Scale Systems*, (Villeneuve d'Ascq, France), 2010.

[9] B. van den Broek, W. Wiegerinck, and B. Kappen, "Graphical model inference in optimal control of stochastic multi-agent systems," *J. of Artificial Intelligence Research*, vol. 32, no. 1, pp. 95–122, 2008.

[10] Z. Ghahramani and M. Beal, "Propagation algorithms for variational Bayesian learning," *Advances in Neural Information Processing Systems*, vol. 13, pp. 507–513, 2001.

[11] V. Šmídl and A. Quinn, *The Variational Bayes Method in Signal Processing*. Springer, 2005.

[12] M. Kárný, J. Böhm, T. V. Guy, L. Jirsa, I. Nagy, P. Nedoma, and L. Tesař, *Optimized Bayesian Dynamic Advising: Theory and Algorithms*. London: Springer, 2006.

[13] S. Amari, S. Ikeda, and H. Shimokawa, "Information geometry of α-projection in mean field approximation," in *Advanced Mean Field Methods* (M. Opper and D. Saad, eds.), (Cambridge, Massachusetts), The MIT Press, 2001.

[14] M. Sato, "Online model selection based on the variational Bayes," *Neural Computation*, vol. 13, pp. 1649–1681, 2001.

Towards a Supra-Bayesian Approach to Merging of Information

Vladimíra Sečkárová[*]
Department of Adaptive Systems
Institute of Information Theory and Automation
Prague, CZ 182 08
seckarov@utia.cas.cz

Abstract

Merging of information given by different decision makers (DMs) has become a much discussed topic in recent years and many procedures were developed towards it. The main and the most discussed problem is the incompleteness of given information. Little attention is paid to the possible forms in which the DMs provide them; in most of cases arising procedures are working only for a particular type of information. Recently introduced Supra-Bayesian approach to merging of information brings a solution to two previously mentioned problems. All is based on a simple idea of unifying all given information into one form and treating the possible incompleteness. In this article, beside a brief repetition of the method, we show, that the constructed merger of information reduces to the Bayesian solution if information calls for this.

1 Introduction

In this article we bring the answer to a consistency question regarding the final result of information merging method based on Supra-Bayesian approach (introduced in [7]).

Method itself deals with problem of incomplete and incompatible (having different forms) data from sources – decision makers. People are trying to solve the incompleteness by developing various methods, bases of which are, e.g. semantics, entities and trust [1], reduction of the combination space by representing the notion of source redundancy or source complementarity [2] or Bayesian networks and factor graphs [3]. Altogether they often lack one thing – they are usable only if the information has unified form. The Supra-Bayesian merging solves previously mentioned problems in three steps. First, we focus on the incompatibility of forms of input data and transform them into a probabilistic form. Second, we fill in the missing information (in the paper it is called extension) to resolve the problem of incompleteness. After that we will construct the merger of already transformed and extended data. Articles related to the proposed topic can be found in [4] and [5].

Section 2 briefly describes the construction of the merger. Section 3 presents an important check of the solution's logical consistency: the final merger reduces to the standard Bayesian learning when the processed data meets standard conditions leading to it. Throughout the text a discrete case is considered.

2 Basic terms and notation

In the beginning of this section we introduce the basic terms and notation used through the text, then we give the main steps of the method.

[*] Department of Probability and Mathematical Statistics, Faculty of Mathematics and Physics, Charles University in Prague

The basic terms we use are as follows:

- a source – a decision maker (e.g. human being), which gave us the information,
  – we now pick one source, denoted by S; the explained setup can be, of course, applied to other sources as well,
- a domain (of the source) – a state space, about which the source provides the information,
  – since a domain can be difficult to describe, we use (discrete) random variable to map it onto preferable space; the range of this mapping consists of finite number of elements
  – every source can describe more than one domain; the relation between them and their ranges maps random vector
- a neighbour of the source S – another source, which has at least one domain (of its considered random variables) same as S
  (note that the range of these variables can differ, so the arising probability measure can be different for each source)
  – we assume that the number of neighbours is finite (for each considered source S); they are labeled by $j = 1, \ldots, s - 1 < \infty$,
  – altogether, we have a set consisting of s sources (source S and its $s - 1$ neighbours),
  – we denote random vector of j^{th} source by $\mathbf{Y}_j$, set of its realizations by $\{\mathbf{y}_j\}$.

2.1 Transformation of information into probabilistic type

I. Consider that the j^{th} source expressed the information about its domain as a realization of its random vector $\mathbf{Y}_j$ denoted by $\mathbf{x}_j$.

The transformation to probability mass function (pmf) $g_{\mathbf{Y}_j}$ will be done via Kronecker delta as follows:

$$g_{\mathbf{Y}_j}(\mathbf{y}_j) = \delta(\mathbf{x}_j - \mathbf{y}_j) = \begin{cases} 1 & \text{if } \mathbf{x}_j = \text{ a particular realization } \mathbf{y}_j \text{ of } \mathbf{Y}_j \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

II. Let j^{th} source give us the conditional expectation of the function of $\mathbf{Y}_j$. We would like to determine the pmf $g_{\mathbf{Y}_j}(\mathbf{y}_j) = g_{\mathbf{F}_j}(\mathbf{f}_j|\mathbf{p}_j)$, where $\mathbf{P}_j$ denotes a part of $\mathbf{Y}_j$, which is specified by source's past experience with realizations $\{\mathbf{p}_j\}$ and $\mathbf{F}_j$ denotes a part of $\mathbf{Y}_j$ expressing source's uncertainty (ignorance) with realizations $\{\mathbf{f}_j\}$. We will use the maximum entropy principle (see [8], [9]): we construct a set of all possible pmfs describing $\mathbf{Y}_j$ and satisfying the expectations, then we choose the pmf with the highest entropy.

III. Let j^{th} source give us the expectation of the function of $\mathbf{Y}_j$. Similarly as in the previous case we will use the maximum entropy principle to determine $g_{\mathbf{Y}_j}(\mathbf{y}_j) = g_{\mathbf{P}_j}(\mathbf{p}_j)$.

IV. and V. Let the source give a pmf of $\mathbf{Y}_j$ denoted by $g_{\mathbf{P}_j}(\mathbf{p}_j)$ or conditional pmf of a part of $\mathbf{Y}_j$ conditioned by the remaining part denoted by $g_{\mathbf{F}_j}(\mathbf{f}_j|\mathbf{p}_j)$. These types of information already are in the targeted probabilistic form.

2.2 Extension

Our first step in constructing the extensions is a unification of the ranges of considered sources $j = 1, \ldots, s < \infty$, which means construction of random vector $\mathbf{Y}$ involving all different random variables considered by sources. A set of realizations of $\mathbf{Y}$ will be denoted by $\mathcal{Y} = \{\mathbf{y}\}$ and their number will be finite (since we assumed range of each source has finite number of realizations).

The decomposition of $\mathbf{Y}$ according to j^{th} source then arises naturally:

- if j^{th} source has its random vector decomposed into two parts $\mathbf{Y}_j = (\mathbf{F}_j, \mathbf{P}_j)$ (as introduced in previous section), the decomposition of $\mathbf{Y}$ will be: $\mathbf{Y} = (\mathbf{U}_j, \mathbf{F}_j, \mathbf{P}_j)$, where $\mathbf{U}_j$ (with realizations $\{\mathbf{u}_j\}$) stands for the remaining realizations in $\mathbf{Y}$ unconsidered by j^{th} source;
- if for j^{th} source holds that $\mathbf{Y}_j = \mathbf{P}_j$, then the decomposition of $\mathbf{Y}$ will be: $\mathbf{Y} = (\mathbf{U}_j, \mathbf{P}_j)$, where again the part $\mathbf{U}_j$ denotes the remaining random variables in $\mathbf{Y}$ unconsidered by the source.

The yet unconstructed merger $\widetilde{h}$ serves us for the extension of pmfs $g_{\mathbf{P}_j}$ and $g_{\mathbf{F}_j|\mathbf{P}_j}$ to $g_{\mathbf{Y}}^{(j)}$ describing the union of neighbours' ranges. If the conditional pmf $g_{\mathbf{F}_j|\mathbf{P}_j}$ is available, then the extension is: $g_{\mathbf{Y}}^{(j)}(\mathbf{y}) = \widetilde{h}(\mathbf{u}_j|\mathbf{f}_j, \mathbf{p}_j) g_{\mathbf{F}_j}(\mathbf{f}_j|\mathbf{p}_j)\widetilde{h}(\mathbf{p}_j)$, where $\widetilde{h}(\mathbf{p}_j)$, $\widetilde{h}(\mathbf{u}_j|\mathbf{f}_j, \mathbf{p}_j)$ and $\widetilde{h}(\mathbf{u}_j|\mathbf{p}_j)$ are marginal and conditional versions of $\widetilde{h}$. We proceed similarly if the marginal pmf $g_{\mathbf{P}_j}$ is available.

2.3 Final merger

After successfully dealing with the transformation and extension of given information we can derive the merger. According to the Bayesian framework [10] our merger will be following pmf:

$$\widetilde{h} = \arg\min_{\hat{h}\in\widehat{H}} \mathrm{E}_{\pi(h|D)}[\mathrm{L}(h, \hat{h})|D],$$

where: $\widehat{H}$ denotes a set of all possible estimates $\hat{h}$ of h, D stands for a matrix consisting of extended probability vectors $g_{\mathbf{Y}}^{(j)}$, $\pi(h|D)$ is the posterior pdf of h based on D, $\mathrm{L}(.\,,.)$ is a loss function.

Since h and $\hat{h}$ are pmfs, the loss function should measure the distance between them. In particular, we choose the Kerridge inaccuracy $\mathrm{K}(.\,,.)$ (see [11]). We then get the following identity (after using Fubini's theorem and a little bit of computation; H is a probabilistic simplex containing h-values):

$$\arg\min_{\hat{h}\in\widehat{H}} \mathrm{E}_{\pi(h|D)}\left[\mathrm{K}(h, \hat{h})|D\right] = \ldots = \arg\min_{\hat{h}\in\widehat{H}} \mathrm{K}\left(\mathrm{E}_{\pi(h|D)}(h|D), \hat{h}\right).$$

Kerridge inaccuracy reaches the minimal value if its arguments are equal almost everywhere (a.e.) (see [11]). Then the following equation holds:

$$\widetilde{h} = \arg\min_{\hat{h}\in\widehat{H}} \mathrm{E}_{\pi(h|D)}\left[\mathrm{K}(h, \hat{h})|D\right] = \mathrm{E}_{\pi(h|D)}(h|D).$$

The only problem is we do not have the posterior pdf $\pi(h|D)$ of h, so before we actually get to the formula expressing the final merger $\widetilde{h}$ (final estimate of h) we have to choose $\pi(h|D)$. Again we will use maximum entropy principle. This time we are looking for the element with highest entropy subject to additional constraints. The constraints will be connected with the opinion of source S about the distance of j^{th} source from the unknown pmf h using Kerridge inaccuracy (for all $j = 1, \ldots, s$). They are expressed by

$$\mathrm{E}_{\pi(h|D)}\left(\mathrm{K}(g_{\mathbf{Y}}^{(j)}, h)|D\right) \leq \beta_j(D). \tag{2}$$

Thus, to obtain the optimal $\widetilde{\pi}(h|D)$ we have to solve following optimization task:

$$\widetilde{\pi}(h|D) = \arg\min_{\pi(h|D)\in\mathrm{M}} \left[\int_H \pi(h|D)\log\pi(h|D)\mathrm{d}h\right], \tag{3}$$

where $\mathrm{M} = \left\{\pi(h|D) :\ \mathrm{E}_{\pi(h|D)}(\mathrm{K}(g_{\mathbf{Y}}^{(j)}, h)|D) - \beta_j(D) \leq 0,\ j = 1, \ldots, s,\ \int_H \pi(h|D)\mathrm{d}h - 1 = 0\right\}$.

By constructing and rearranging the Lagrangian $\mathrm{L}(.\,,.)$ of the task (3) we get that its minimum is reached for pdf of Dirichlet distribution $Dir(\{\nu_{\mathbf{y}}\}_{\mathbf{y}\in\mathcal{Y}})$:

$$\widetilde{\pi}(h|D) = \frac{1}{Z(\boldsymbol{\lambda}(D))}\prod_{\mathbf{y}\in\mathcal{Y}} h(\mathbf{y})^{\nu_{\mathbf{y}}-1} \quad \text{with parameters } \nu_{\mathbf{y}} = 1 + \sum_{j=1}^{s}\lambda_j(D) g_{\mathbf{Y}}^{(j)}(\mathbf{y}), \quad \forall\, \mathbf{y}\in\mathcal{Y}.$$

Once we have computed the posterior pdf, we can go back to the expressing the final merger (the optimal estimate $\widetilde{h}$ of h). Denote the number of realizations of $\mathbf{Y}$ by n $(< \infty)$ and use the properties of Dirichlet distribution, particularly

$$\mathrm{E}_{\widetilde{\pi}(h|D)}[h(\mathbf{y})|D] = \frac{\nu_{\mathbf{y}}}{\nu_0}, \quad \text{where } \nu_0 = \sum_{\mathbf{y}\in\mathcal{Y}}\nu_{\mathbf{y}} = \sum_{\mathbf{y}\in\mathcal{Y}} 1 + \sum_{j=1}^{s}\lambda_j(D)\overbrace{\sum_{\mathbf{y}\in\mathcal{Y}} g_{\mathbf{Y}}^{(j)}(\mathbf{y})}^{=1}.$$

We get following result:

$$\widetilde{h}(\mathbf{y}) = \frac{1 + \sum_{j=1}^{s}\lambda_j(D) g_{\mathbf{Y}}^{(j)}(\mathbf{y})}{n + \sum_{j=1}^{s}\lambda_j(D)}. \tag{4}$$

3 Connection to the Bayesian solution

As promised earlier (see Section 1) we will now check if the final merger (4) reduces to a standard Bayesian learning if merging scenario meets conditions leading to it. First we will derive the empirical pmf via Bayesian approach, second we will reformulate the problem so that our merger can be applied, compute the empirical pmf and compare the results.

3.1 A Bayesian view

Let

- Y be a discrete random variable with finite number of realizations $\{y\} = \mathcal{Y}$,
- $\boldsymbol{\theta}$ be a following random vector: $\boldsymbol{\theta} = (\mathrm{P}(Y = y))_{y\in\mathcal{Y}} = (\theta_y)_{y\in\mathcal{Y}}$. Then let $X_1, \dots, X_s$, ($s < \infty$), denote the sequence of observations about Y, which will be considered as independent random variables with the same distribution as Y (depending on $\boldsymbol{\theta}$).

If we assume that

- the prior distribution of $\boldsymbol{\theta} = (\theta_y)_{y\in\mathcal{Y}}$ is Dirichlet distribution $Dir(\{\alpha_y\}_{y\in\mathcal{Y}})$, meaning $q(\boldsymbol{\theta}) \propto \prod_{y\in\mathcal{Y}} \theta_y^{\alpha_y - 1}$,
- the conditional probability of X_j, $j = 1, \dots, s$, conditioned by $\boldsymbol{\theta}$ is $f_{X_j}(x_j|\boldsymbol{\theta}) = \prod_{y\in\mathcal{Y}} \theta_y^{\delta(x_j - y)}$, where $\delta(.)$ stands for Kronecker delta (see (1)),

the posterior pmf of $\boldsymbol{\theta}$ based on $X_1, \dots, X_s$ is then

$$\pi(\boldsymbol{\theta}|X_1 = x_1, \dots, X_s = x_s) \propto q(\boldsymbol{\theta}) \prod_{j=1}^{s} f_{X_j}(x_j|\boldsymbol{\theta})$$

$$= \prod_{y\in\mathcal{Y}} \theta_y^{\alpha_y - 1} \prod_{j=1}^{s} \prod_{y\in\mathcal{Y}} \theta_y^{\delta(x_j - y)} = \prod_{y\in\mathcal{Y}} \theta_y^{\alpha_y + \sum_{j=1}^{s} \delta(x_j - y) - 1} \tag{5}$$

Since the formula (5) is the pdf of Dirichlet distribution $Dir\left(\left\{\alpha_y + \sum_{j=1}^{s} \delta(x_j - y)\right\}_{y\in\mathcal{Y}}\right)$, we can easily compute the conditional expectation of θ_y conditioned by $X_1, \dots, X_s$ as follows:

$$\mathrm{E}_{\pi(\boldsymbol{\theta}|X_1,\dots,X_s)}(\theta_y|X_1 = x_1, \dots, X_s = x_s) = \widetilde{P}(Y = y) = \frac{\alpha_y + \sum_{j=1}^{s} \delta(x_j - y)}{\sum_{y\in\mathcal{Y}} \left[\alpha_y + \sum_{j=1}^{s} \delta(x_j - y)\right]}$$

$$= \frac{\alpha_y + \sum_{j=1}^{s} \delta(x_j - y)}{\sum_{y\in\mathcal{Y}} \alpha_y + s} \tag{6}$$

Under the following choice:

$$\alpha_y = 1 \quad \forall\, y \in \mathcal{Y} \tag{7}$$

formula (6) will look as follows

$$\widetilde{P}(Y = y) = \frac{1 + \sum_{j=1}^{s} \delta(x_j - y)}{\sum_{y\in\mathcal{Y}} 1 + s}. \tag{8}$$

If n denotes the number of realizations of Y, then: $\quad \widetilde{P}(Y = y) = \frac{1 + \sum_{j=1}^{s} \delta(x_j - y)}{n + s}$.

Note: The first part of (8) – $\frac{1}{\sum_{y\in\mathcal{Y}} 1 + s}$ – can be considered as the prior pmf of Y, because if there is no available information, we will get: $\widetilde{P}_0(Y = y) = \frac{1}{\sum_{y\in\mathcal{Y}} 1 + s}$. Then, the choice (7) coincides with the statement, that the prior pmf for Y is a pmf of Uniform distribution.

Illustrative example:

Assume we are interested in the changes of the stock price. Y will now be an identity mapping from the set consisting of 3 elements: 1 (increase), 0 (stagnation), -1 (decrease). We now get a sequence of data - opinions from independent experts: $\{1, -1, 1, 1, 0, 1, 1, -1, 1, 1\}$. Then the estimate of the probabilities will be (regarding the (7)):

$\widehat{P}(Y = 1) = \frac{1+7}{3+10} = \frac{8}{13}, \quad \widehat{P}(Y = 0) = \frac{1+1}{3+10} = \frac{2}{12}, \quad \widehat{P}(Y = -1) = \frac{1+2}{3+10} = \frac{3}{13}.$

3.2 Merging approach

Now we reformulate and handle the same information scenario as in Subsection 3.1 by using the proposed information merging.

Let us have a group of s (independent) sources, all of them describing the same domain and range. Therefore sources are neighbours and so the merging can be applied on them. The relation between domain and range maps discrete random variable Y, realizations of which are denoted by $\{y\} = \mathcal{Y}$.

Assume also that the information they gave are the values of Y, denoted by $x_1, \ldots, x_s$. Now we can follow the steps introduced in the previous sections:

1. transformation: (non-probability form into probability form)

– x_j will be expressed (in the probability form) as follows: $g_{Y_j = Y}(y_j = y) = \delta(x_j - y)$,

2. extension: (from particular domains to the union of all considered domains)

– since the sources have the same domain, Y, the union is also Y,

– because of that, the extended version of probabilistic form of given information will be:

$g_Y^{(j)}(y) = \delta(x_j - y)$,

3. merging: now that we have probabilistic information extended on Y, we can use the merger (4):

$$\widetilde{h}(y) = \frac{1 + \sum_{j=1}^{s} \lambda_j(D) g_Y^{(j)}(y)}{\sum_{y \in \mathcal{Y}} 1 + \sum_{j=1}^{s} \lambda_j(D)} = \frac{1 + \sum_{j=1}^{s} \lambda_j(D) \delta(x_j - y)}{\sum_{y \in \mathcal{Y}} 1 + \sum_{j=1}^{s} \lambda_j(D)},$$

which for particular choice $\lambda_1(D) = \ldots = \lambda_j(D) = \ldots = \lambda_s(D) = 1$ has following form:

$$\widetilde{h}(y) = \frac{1 + \sum_{j=1}^{s} \delta(x_j - y)}{\sum_{y \in \mathcal{Y}} 1 + s}, \tag{9}$$

which coincides with (8). That is if we assume that the sources have the same reliability factor (see subsection 2.3) and it is equal to 1, the final merger (4) will reduce to the standard Bayesian learning considered in Subsection 3.1.

Illustrative example:

Our results can be easily applied on the example in the previous section: we have 9 independent sources, which have the same domain. Therefore they are neighbors. All given information are just values, we have to transform them into probabilities (see Section 2.1). Since they also have the same range no extension is needed, so we can directly proceed to the merging. According to (9) results are the same as in the example in Subsection 3.1.

Remark

In the note after the final merger (8) we brought the explanation of what should its first part represent – generally, it stands for the prior pmf for considered random vector $\mathbf{Y}$ (see (4)). In this paper, the prior pmf of $\mathbf{Y}$ is a pmf of uniform distribution. But we will be allowed to use another prior distribution if we choose constrained minimum cross entropy principle (see [9]) for determination of the posterior pdf (see subsection 2.3) instead of constrained maximum entropy principle. It is because the maximum entropy principle coincides with minimum cross entropy principle when prior distribution is uniform.

4 Conclusion

This paper brings an important conclusion regarding a new method for merging of information, which successfully deals with the different types of given partially overlapping information and also with problem of missing data. Since the method is based on Bayesian framework, we showed that it reduces to a standard Bayesian learning if independent identically distributed data are at disposal for parameter estimation. Still there are some open problems and topics of the future work, e.g. the choice of constraints $\beta_j(D)$ in (2), choice of prior distribution (see previous remark) and the extension to the continuous space.

Acknowledgement

This research has been partially supported by GAČR 102/08/0567 and DAR 1M0572.

References

[1] L. Šubelj, D. Jelenc, E. Zupančič, D. Lavbič, D. Trček, M. Krisper, and M. Bajec. Merging data sources based on semantics, contexts and trust. *The IPSI BgD Transactions on Internet Research*, 7(1):18–30, 2011.

[2] B. Fassinut-Mombot and J.B. Choquel. A new probabilistic and entropy fusion approach for management of information sources. *Information Fusion*, 5(1):35–47, 2004.

[3] G. Pavlin, P. de Oude, M. Maris, J. Nunnik, and T. Hood. A multi-agent systems approach to distributed bayesian information fusion. *Information Fusion*, 11(3):267–282, 2010.

[4] M. Kárný, T. Guy, A. Bodini, and F. Ruggeri. Cooperation via sharing of probabilistic information. *International Journal of Computational Intelligence Studies*, pages 139–162, 2009.

[5] M. Kárný and T.V. Guy. Sharing of Knowledge and Preferences among Imperfect Bayesian Participants. In *Proceedings of the NIPS Workshop 'Decision Making with Multiple Imperfect Decision Makers'*. UTIA, 2010.

[6] C. Genest and J. V. Zidek. Combining probability distributions: a critique and an annotated bibliography. With comments, and a rejoinder by the authors. *Stat. Sci.*, 1(1):114–148, 1986.

[7] V. Sečkárová. Supra-Bayesian Approach to Merging of Incomplete and Incompatible Data. In *Decision Making with Multiple Imperfect Decision Makers Workshop at 24th Annual Conference on Neural Information Processing Systems*, 2010.

[8] E.T. Jaynes. Information theory and statistical mechanics. I, II. 1957.

[9] J. E. Shore and R. W. Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Trans. Inf. Theory*, 26:26–37, 1980.

[10] M. H. DeGroot. *Optimal statistical decisions.* Wiley-Interscience; Wiley Classics Library. Hoboken, NJ: John Wiley and Sons. xx, 489 p., 1970.

[11] D.F. Kerridge. Inaccuracy and inference. *J. R. Stat. Soc., Ser. B*, 23:184–194, 1961.

Ideal and non-ideal predictors in estimation of Bellman function

Jan Zeman
Institute of Information Theory and Automation
Pod Vodarenskou vezi 4
CZ-182 08 Prague 8
Czech Republic
zeman@utia.cas.cz

Abstract

The paper considers estimation of Bellman function using revision of the past decisions. The original approach is further extended by employing predictions coming from an imperfect predictor. The resulting algorithm speeds up the convergence of Bellman function estimation and improves the results quality. The potential of the approach is demonstrated on a futures market data.

1 Introduction

The dynamic programming (DP) is clever and effective framework in many problems [1, 2]. Unfortunately, it also suffers by many issues such as curse of dimensionality [6]. Moreover, the incomplete knowledge and uncertainty makes the dynamic programming task hardly solvable, although the analytical solution of DP task is known [5]. The approximate dynamic programming (ADP) tries to solve the DP tasks and fight with all the technical issues. There are many ways to approximate the dynamic programming, but all of them assume that the approximation is precise enough to solve the problem. There is only a few approaches implicitly assuming the non-ideal or imperfect approximation. This paper presents such a approach related to solution of Bellman equation[1], i.e. estimation of Bellman function.[1]

The value iteration algorithm [1, 2, 4] makes the theoretical ground for estimation of Bellman function. It also suffers by dimensionality [4], which is often solved by approximate methods (see [3] for a review). There is an duality relation between optimal decision rule and Bellman function [1] and value iteration uses the idea of the convergent improving of Bellman function and decision rule - both together. Unfortunately, the value iteration is difficult to solve for tasks with continuous variables. We try to approximate Bellman function in value iteration and to speed up the convergence by searching the samples of the optimal decision rule [7]. The present approach can be extended using the prediction. This extension can bring only restricted impact, therefore must be considered a non-ideal predictor and its properties. Hence, the paper presents the imagination of the ideal and non-ideal predictor and their influence to estimation of Bellman function. These imaginations are compared and we point out the break, where the non-ideal one stops working. Respecting such a property, we can improve the estimation of Bellman function.

The paper briefly introduces the dynamic programming and estimation of Bellman function in Sec. 2. Then introduces the revisions and the related optimality criterion (Sec. 3) and considers the extension of the approach by additional usage of ideal and non-ideal prediction (Sec. 4). Finally, tests the presented idea on the task of trading commodity futures (Sec. 5).

[1] also called *value* function or *cost-to-go* function

2 Dynamic programming and revisions

We consider discrete time $t \in \{1, 2, \ldots, T\} = t^*$, where T is *horizon*. We consider *decision maker*, which is human or machine with particular aims to a part of world, so-called *system*. The decision maker observes the system and obtains *observation* y_t, then designs *decision* u_t and applies it to the system, this process is repeated at each time $t \in t^*$. The information available to design decision u_t is called knowledge P_t and consists of past observation and past decisions, $P_t = \{y_1, y_2, \ldots, y_t\} \cup \{u_1, u_2, \ldots, u_{t-1}\}$. The decision maker designs a decision rule, $u_t = \pi_t(P_t)$. The aim of the dynamic programming is to design the sequence of decision rules $\pi_1, \pi_2, \ldots, \pi_T$, so-called *strategy*, in order to maximize the sum of the gain functions $\sum_{t=1}^{T} g_t$ under the conditions above.

We consider the following task properties: (i) The decision maker and its environment work in open loop, i.e. decisions have influence neither on the environment behavior nor future observation; (ii) gain function has form $g_t(P_t, u_t)$ and it depends at last n observations and decisions, i.e. $g_t(y_{t-n}, \ldots, y_t, u_{t-n}, \ldots, u_t)$, where n is finite number.

2.1 Finite and infinite horizon

The optimal rule for finite horizon T can be constructed value-wise (see [5])

$$\pi_t^o(P_t) = \arg\max_{u_t} \mathbb{E}\left[g_t + \mathcal{V}_{t+1}(P_{t+1})|u_t, P_t\right], \tag{1}$$

where function $\mathcal{V}_{t+1}(.)$ is *Bellman function* and it is given by recurrence

$$\mathcal{V}_t(P_t) = \max_{u_t} \mathbb{E}\left[g_t + \mathcal{V}_{t+1}(P_{t+1})|u_t, P_t\right] \tag{2}$$

with the terminal condition

$$\mathcal{V}_{T+1}(P_{T+1}) = 0. \tag{3}$$

The equations (1, 2, 3) form the algorithm of dynamic programing with finite horizon T. This algorithm is important for revisions.

We consider task with infinite horizon $T = +\infty$. Consequently, the solution has stationary form:

$$\pi^o(P_t) = \arg\max_{u_t} \mathbb{E}\left[g_t + \mathcal{V}(P_{t+1})|u_t, P_t\right], \tag{4}$$

where function $\mathcal{V}$ is stationary Bellman function and it is given by recurrence

$$\mathcal{V}(P_t) = \max_{u_t} \mathbb{E}\left[g_t + \mathcal{V}(P_{t+1})|u_t, P_t\right]. \tag{5}$$

The equation (4) contains two terms at right-hand side. In general, both therms in (4) can be calculated difficultly due to uncertainty, incomplete knowledge, demanded prediction etc. Hence, ADP considers approaches how to calculate the terms, or to approximate them adequately. We focus on the approximation of Bellman function.

2.2 Estimation of Bellman function

Let us consider the infinite horizon task. The equation (4) contains two terms at right-hand side. First term is gain function g_t, which can be evaluated under knowledge P_t for the considered shape of g_t. Second term is Bellman function $\mathcal{V}(.)$ in stationary form applied on unavailable knowledge P_{t+1}. Under knowledge P_t the decision maker must predict further knowledge P_{t+1} and estimate Bellman function.

Let us assume that we have ideal predictor $\mathcal{M}^I(.)$ such as it can predict $P_{t+1} = \mathcal{M}^I(P_t)$. Equation (5) can be written for each time index $i \in \{1, \ldots, t\}$

$$\mathcal{V}(P_i) = \max_{u_i} \mathbb{E}\left[g_i + \mathcal{V}(\mathcal{M}^I(P_i))|u_i, P_i\right]. \tag{6}$$

The obtained t-equations system contains the main information about Bellman function $\mathcal{V}(.)$. Assuming the knowledge of the optimal decisions $u_1^o, \ldots, u_t^o$, the system is transformed to final form:

$$\mathcal{V}(P_i) = \mathbb{E}\left[g_i + \mathcal{V}(\mathcal{M}^I(P_i))|u_i^o, P_i\right], \quad \text{for } i \in \{1, \ldots, t\}. \tag{7}$$

This equations system contains only unknown function $\mathcal{V}(.)$ and can be used to estimation Bellman function [7]. The information contained in (7) is not full, therefore this system can bring only an approximate solution. This approximate solution can be found considering approximation of Bellman function in parametrized form $\mathcal{V}(.) \approx V(., \Theta)$, where Θ is finite dimensional unknown parameter. The system (7) characterizes points of Bellman function and inserting the approximation $V(., \Theta)$ the system is transformed to system for unknown variable Θ. Typically, the number of equations in (7) is bigger than dimension of parameter Θ. Hence, the best estimation of parameter $\hat{\Theta}$ is searched by regression methods.

This approach originates from value iteration [4] and the system (7) can be interpreted as subsystem of the full system:

$$\mathcal{V}(P) = \mathbb{E}\left[g_i + \mathcal{V}(\mathcal{M}^I(P))|\pi^o(P), P\right], \quad \text{for } P \in P^*. \tag{8}$$

Bellman function $\mathcal{V}(.)$ is a solution of system (8). The formal difference between system (7) and (8) is in the used knowledge. While system (8) contains all possible values of P, the system (7) contains only the realizations passed during the decision process $P_1, P_2, \ldots, P_t$. We have assumed the knowledge of the optimal decisions related to these realizations, therefore the term $\pi^o(P)$ is known only for these realizations $u_i^o = \pi^o(P_i)$ for $i \in \{1, 2, \ldots, t\}$. All in all, the system (8) contains a full information about Bellman function, whereas the system (7) contains only t points of Bellman function.

3 Revisions

The previous approach depends on the possibility to find the optimal decisions $u_1^o, \ldots, u_t^o$ for the given knowledge P_t. This is possible to obtain by the revisions.

The revision is the reconsideration of the decision under another knowledge than was used to design it. To design decision the maximal available knowledge is used, but we can redesign the decision under higher knowledge; let us denote the rules and the decisions by superscript, which characterizes the knowledge used to design the rule, e.g. $u_t^{t+i} = \pi^{t+i}(P_t)$ is redesign of the tth rule/decision under knowledge P_{t+i}. But we omit the superscript, when the rule/decision is designed under natural conditions $u_t = u_t^t = \pi_t^t(P_t) = \pi_t(P_t)$ is rule/decision designed under the knowledge available to design it. This differs the revision and the decisions. One clever way of this redesign is solving the same task, but with the finite horizon $T \equiv t$. We can reconsider all decision using the equations (1, 2, 3). And obtain the revision based on the knowledge P_t:

$$U_t^t = \{u_1^t, u_2^t, u_3^t, \ldots, u_t^t\}, \tag{9}$$

where $u_i^t = \pi_i^t(P_t)$ for $i \in \{1, 2, \ldots, t\}$. The sequence U_t^t is called t-revision. Due to asymptotic properties of DP [5], the revisions tends to optimal values, i.e. $u_i^t = \pi_i^t(P_i) \to \pi_i^o(P_i) = u_i^o$.

3.1 Optimality of revision

For our special shape of the gain function, the convergence can be interpreted as weighting of the influence of the terminal condition (3) and information contained in data, inserted into g_t. The algorithm of searching of t-revision goes backwards and from design of $\pi_t^t(.)$ to $\pi_1^t(.)$. The terminal condition (3) influences a decision rule $\pi_i^t(.)$ via Bellman equation (2). But the information-rich data can quickly decrease the influence, such as $\pi_t^t(.)$ is influenced, but further $\pi_{t-1}^t(.), \pi_{t-2}^t(.), \ldots$ are influenced less and less. When the influence of terminal condition is lost for $l \in \{1, 2, \ldots, t\}$ and $\pi_l^t(.)$ is independent on terminal condition (3), then the decision rule $\pi_i^t(.)$ maps the knowledge to optimal decision, where $i \in \{l, l-1, l-2, \ldots, 1\}$. The optimality is given by the independence on the terminal condition and the absolute dependence on the data.

The issue is to recognize, whether the optimality was reached. This factor can negative influence the potential of estimation of Bellman function. The bad recognized optimality can lead in: learning from non-optimal decisions, i.e. adding the non-valid equations to system (7); or redundant omission of some optimal decisions, i.e. omission available equations of system (7). Hence, the preciseness of optimality recognition is required.

The possible way how to recognize the optimal decision lies in independence on the terminal condition (3). Let us consider the terminal condition in form:

$$\mathcal{V}_{T+1}(P_{T+1}) = f(P_{T+1}), \tag{10}$$

where $f(.)$ is general function of P_{T+1}. Let us denote the class of all those functions $\mathfrak{F}$. The revision algorithm (1, 2, 3) can be generalized by usage the terminal condition (10) instead of (3). Then, the revision can be written as function of knowledge and terminal condition $u_i^t = \pi_i^t(P_i, f)$.

Finally, *the revision of decision u_i^t equals optimal decision u_i^o, if the revision does not depend on terminal condition (10), i.e.*

$$\exists \tilde{u}_i \in u^* \quad \forall f \in \mathfrak{F} \quad \pi_i^t(P_i, f) = \tilde{u}_i. \tag{11}$$

and the constant $\tilde{u}_i$ is the optimal decision, $u_i^o = \tilde{u}_i$.

The impact of this proposition is great, because it represents the inter-connection between the finite and infinite horizon task. The proposition gives algorithm how to use the generalized finite horizon task (1, 2, 10) to find some optimal decisions of infinite horizon task (4, 5). Unfortunately, the proposition does not guarantee that any optimal decisions will be found. Typically, there exists an index t^o such that revisions $u_1^t, u_2^t, \ldots, u_{t^o}^t$ are optimal and independent on f; and revisions $u_{t^o+1}^t, \ldots, u_t^t$ cannot be decided, whether are optimal because of the dependence on f.

The proposition uses simply idea that the interconnection between two consequent decisions is done via Bellman equation and the connection term is Bellman function. Using the right Bellman function $\mathcal{V}_{t+1}(.)$ we could connect the finite horizon task and infinite horizon task easily. Unfortunately, Bellman function is unknown therefore the proposition must go over all possible candidates $f \in \mathfrak{F}$, i.e. over all possible interconnections. Having a bit information about Bellman function, it is possible to exclude the impossible candidates and use the proposition over subset $\mathfrak{F}' \subset \mathfrak{F}$ containing only the possible ones. This idea can be reached by usage of predictions.

4 Revision and prediction

As was mentioned above, we can operate with ideal predictor $\mathcal{M}^I(.)$ such as $P_{t+1} = \mathcal{M}^I(P_t)$. Having the ideal predictor, we can use it recursively to predict P_{t+i} for any $i > 0$ and use the P_{t+i} as information for revision and searching its optimality. Such a approach can help us to increase the value t^o and use all available equations of system (7).

Let us consider the revision algorithm. For P_t, the algorithm starts with terminal condition (10). For one-step prediction P_{t+1}, the algorithm has one more step and due to back recursion in (2) obtain $\mathcal{V}_{t+1}$, i.e. the restricted analogy of condition (10), after one step:

$$\begin{aligned} \mathcal{V}_{t+2}(P_{t+2}) &= f(P_{t+2}), && (12) \\ \mathcal{V}_{t+1}(P_{t+1}) &= \max_{u_{t+1}} \mathbb{E}\left[g_{t+1} + \mathcal{V}_{t+2}(P_{t+2}) | u_{t+1}, P_{t+1}\right], && (13) \end{aligned}$$

where we expect that $f \in \mathfrak{F}$, and $\mathcal{V}_{t+1}(P_{t+1}) \in \mathfrak{F}_1 \subseteq \mathfrak{F}$.

This expectation originates from properties of Bellman equation, which can be viewed as operator on class $\mathfrak{F}$, i.e. $\mathcal{V}_i = \mathcal{T}(\mathcal{V}_{i+1})$, see [1, 2]. The recursion (2) converges for each terminal condition. Consequently, the operator has following property $\lim_{n \to +\infty} \mathcal{T}^n(f) = \mathcal{V}$, where $\mathcal{T}^n(.)$ is operator $\mathcal{T}(.)$ n-times recursively applied onto $f \in \mathfrak{F}$ and $\mathcal{V}$ is Bellman function. Hence, we can expect that the operator $\mathcal{T}(.)$ applied on all functions in $\mathfrak{F}$ produces the subset of $\mathfrak{F}$:

$$\mathfrak{F}_1 = \mathcal{T}(\mathfrak{F}) \quad \text{and} \quad \mathfrak{F}_1 \subseteq \mathfrak{F}. \tag{14}$$

Furthermore, each prediction step can be used as one more application of operator $\mathcal{T}(.)$, which reduces the set of possible candidates to terminal $\mathcal{V}_{t+1}$. The h-step prediction generates h subsets of $\mathfrak{F}$ as is depicted at Fig. 1. The usage of prediction can be interpreted as starting the optimality criterion from less set $\mathfrak{F}_i$ instead of $\mathfrak{F}$, which can result in earlier recognizing the optimal decisions, i.e. obtaining higher value t^o.

Of course, we do not have the ideal predictor $\mathcal{M}^I(.)$, but often we can use an predictor $\hat{P}_{t+1} = \mathcal{M}(P_t)$. We assume that the predictor $\mathcal{M}(.)$ has some restricted preciseness and degenerates the

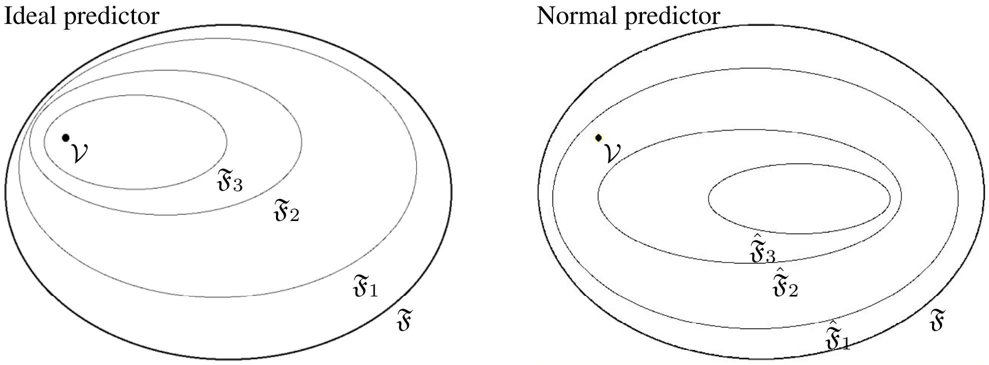


Figure 1: The convergence of operator J in space of possible Bellman functions: $\mathcal{V}$ is optimal Bellman function, $\mathfrak{F}$ is space of possible Bellman functions.

operator $\mathcal{T} \rightarrow \hat{\mathcal{T}}$. Let us denote: $\hat{\mathfrak{F}}_1 = \hat{\mathcal{T}}(\mathfrak{F})$ and $\hat{\mathfrak{F}}_{i+1} = \hat{\mathcal{T}}(\hat{\mathfrak{F}}_i)$. Typically, the non-ideal predictor predicts quite good first one or two steps and then the predictions go worse. The expected influence to Bellman function is depicted at Fig. 1, where the 'Normal predictor' gets lost in second step and the operator $\hat{\mathcal{T}}$ converges to the other function. Despite of this fact, the non-ideal predictor can be successfully used, when the number of prediction steps is restricted. The restriction equals to index of the last set including Bellman function; e.g. it can be used only as one-step predictor in case depicted at Fig. 1. We expect that this phenomenon would be observable as relation between length of used predictions and results quality. We expect slowly increase followed by rapid decrease of results quality according to the growing prediction length.

5 Experiment

The following experiment should demonstrate the expected properties. We compare: (i) the original method to estimation of Bellman function, where the optimal revisions were searched at available data, i.e. the terminal condition (10) was taken over whole set $\mathfrak{F}$; and (ii) the presented method, where the optimal revisions were searched at available data extended by prediction, i.e. the terminal condition (10) was taken over the subset $\hat{\mathfrak{F}}_i \subseteq \mathfrak{F}$. The set of experiments contains 11 experiments per data sequence. Each experiment is related with one prediction length 0-10, where zero length is the original task using $\mathfrak{F}$ and the other lengths $l \in \{1, 2, \ldots, 10\}$ corresponds with the systems of sets $\hat{\mathfrak{F}}_1, \ldots, \hat{\mathfrak{F}}_{10}$ (see previous section).

We expect that results quality will grow with length of prediction to some break value. Then, the result quality will decrease. This expectation is caused by the imagination of the non-ideal predictor analogical to Fig. 1, where Bellman function is in $\hat{\mathfrak{F}}_1$, but it is not in $\hat{\mathfrak{F}}_2$. Thus, we expect that each data set should have some length of prediction l, where Bellman function is in $\hat{\mathfrak{F}}_l$, but it is not in $\hat{\mathfrak{F}}_{l+1}$. The approach to estimate of Bellman function should work most effective, when starts with the smallest subset $\hat{\mathfrak{F}}_l$ containing the Bellman function $\mathcal{V}(.)$. Otherwise, when it starts with subset $\hat{\mathfrak{F}}_{l+1}$, it need more information to find Bellman function, because it got lost by irrelevant set $\hat{\mathfrak{F}}_{l+1}$ and the convergence is delayed. We expect that this phenomenon should be observable as results quality increase for prediction length $1, 2, \ldots, l$, followed by quality decrease for prediction length $l + 1, \ldots, 10$.

The experiment was done on trend prediction task based on the trading with commodities. The task is classical price speculation, where decision maker tries to predict future price trend and chose the decision to follow the trend. The gain function has shape:

$$g_t = (y_t - y_{t-1})u_{t-1} + C|u_{t-1} - u_t|, \tag{15}$$

where y_{t-1}, y_t are samples of price sequence, u_{t-1}, u_t are decisions and C is transaction cost. The decision can be chosen from two-values set $u_t \in \{-1, 1\}$, where $u_t = 1$ characterizes the future price increase and $u_t = -1$ characterizes decrease.

The data used for experiment are day samples of price, so-called close price. The used time series are related to following five commodities: Cocoa - CSCE (CC), Petroleum-Crude Oil Light - NMX (CL), 5-Year U.S. Treasury Note - CBT (FV2), Japanese Yen CME (JY), Wheat - CBT (W). The used data were collected between January 1990 and September 2005, which is about 4000 trading days.

The experiment designs the decisions via approximated (4). The predictor $\mathcal{M}(.)$ is based on the autoregressive model, $y_{t+1} = \alpha y_t + \beta y_{t-1} + e_t$, where α, β are model parameters and e_t is noise, $e_t \approx N(\mu, \sigma)$. The model parameters are estimated via Bayesian estimation [5]. The prediction is calculated recursively $\hat{P}_{t+1} = \mathcal{M}(P_t)$, and $\hat{P}_{t+i+1} = \mathcal{M}(\hat{P}_{t+i})$ for $i \in \{1, \ldots, 9\}$. And Bellman function is approximated in parametrized form:

$$\mathcal{V}(P_t) \approx V(P_t, \Theta) = \Theta' \Psi_t(P_t), \tag{16}$$

where Θ is vector of $n+1$ parameters and $\Psi_t(P_t) = (y_t, y_{t-1}, \ldots, y_{t-n+1}, 1)'$, which is 1st order Taylor expansion of Bellman function $\mathcal{V}(.)$. The parameters Θ are estimated via system (7) and the count of equation t^o in system (7) is estimated via revisions and the optimality proposition.

Table 1 contains experiment results. According to our expectation, the results written by bold font have the expected growing quality. As can be seen, the prediction improved the results quality in all datasets according to non-prediction experiment for $l = 0$. The length of growing trend is related with feasibility of the predictor to the dataset, and it was expected that the 1- or 2-step prediction can improve the results. Hence, the results of 3-step prediction at CC and JY can be viewed as unexpected success. A little surprising fact is the quality of results after the increase. We have expected the rapid decrease due to worse initial conditions, but a few experiments reached comparatively results or better results. The expected behavior can be demonstrated at CL dataset, where $l \in \{0, 1, 2\}$ the results quality grows and then fall down and stay under the value for $l = 0$. An representative of the surprising is JY dataset, which grows for $l \in \{0, 1, 2\}$, then it decrease, but then it increases and reaches better results than for $l \in \{0, 1, 2\}$. The mentioned facts lead to conclusion that the prediction improve the results for a few steps. But after these steps, there cannot be expected any property or trend related to the prediction length.

6 Conclusion

The paper presents the approach to estimation of Bellman function via revisions. The revisions are originally calculated from the knowledge available to design the decision. The paper considers extension of this approach by the usage of predictions. It is expected the better convergence to Bellman function. The idea is considered for ideal predictor and non-ideal predictor. The ideal predictor can simply improve the algorithm, but it is unavailable, whereas all available predictors can be classified as non-ideal. The imagination of the non-ideal predictor leads to expectation that the prediction can improve the approach, when is used a restricted number of prediction steps.

The idea is experimentally tested on trend prediction task, where works quite well. The results have verified the idea that the improvement is related to restricted number of prediction steps. But surprising was the fact that after these few steps, the improvement can be reached, but randomly. This opens the question of the better analysis of the problem: the paper describes only a raw imagination of the problem and the convergence in set $\mathfrak{F}$, and relations between sets $\hat{\mathfrak{F}}_i$ and $\hat{\mathfrak{F}}_{i+1}$ can be more complex than was presented. This fact is topic of the further consideration.

Moreover, the paper presents that the number of prediction steps should be restricted, but it does not give any guidelines how to estimate the right length of the prediction. The right guidelines can make the approach suitable for applications and should be also considered in future.

Ex.	0	1	2	3	4	5	6	7	8	9	10
CC	-13,0	**-13,0**	**-10,7**	**-3,5**	-10,9	-6,6	-12,8	-15,2	-15,2	-8,2	-6,3
CL	-14,2	**-9,6**	**-6,8**	-16,0	-23,8	-21,4	-14,9	-16,7	-21,7	-25,1	-24,8
FV2	2,6	**24,2**	23,1	19,2	22,1	24,8	24,7	27,1	27,4	23,7	16,5
JY	8,3	**20,4**	**22,5**	**40,6**	28,3	30,9	30,8	44,6	39,9	15,5	1,7
W	2,2	**16,0**	**17,5**	12,9	11,0	13,9	10,7	7,6	8,1	13,3	12,6

Table 1: Results of experiments Ap1-Ap10 in $1000 USD.

Acknowledgments

This work was supported by grants MŠMT 1M0572 and GAČR 102/08/0567.

References

[1] R. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, New Jersey, 1957.

[2] D.P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, Nashua, US, 2001. 2nd edition.

[3] M. Hauskrecht. Value-function approximations for partially observable markov decision processes. *Journal of Artificial Intelligence Research*, 13:33–94, 2000.

[4] Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1-2):99–134, 1998.

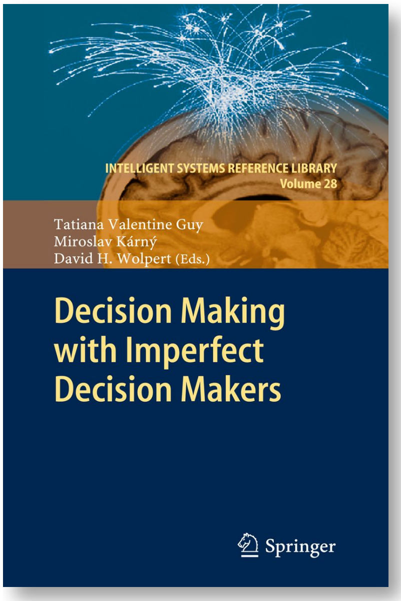
[5] M. Kárný, Böhm J., T. V. Guy, L. Jirsa, I. Nagy, P. Nedoma, and L. Tesař. *Optimized Bayesian Dynamic Advising: Theory and Algorithms*. Springer, London, 2005.

[6] W. B. Powell. *Approximate Dynamic Programming*. Wiley-Interscience, 2007.

[7] J. Zeman. Estimating of Bellman function via suboptimal strategies. In *2010 IEEE Int. Conference on Systems, Man and Cybernetics*. IEEE, 2010.

Springer
the language of science

springer.com



INTELLIGENT SYSTEMS REFERENCE LIBRARY
Volume 28

Tatiana Valentine Guy
Miroslav Kárný
David H. Wolpert (Eds.)

Decision Making with Imperfect Decision Makers

Springer

2011, 2012, XIV, 198 p. 50 illus., 39 in color.

Printed book

Hardcover

- **99,95 € | £90.00 | $129.00**
- ***106,95 € (D) | 109,95 € (A) | SFr. 133.50**

eBook

Available from libraries offering Springer's eBook Collection, or for individual purchase via online bookstores.

A free preview is available on SpringerLink.

- **springer.com/ebooks**

MyCopy

Printed eBook exclusively available to patrons whose library offers Springer's eBook Collection.***

- **€ | $ 24.95**
- **springer.com/mycopy**

T.V. Guy, Institute of Information Theory and Automation, Praha Czech Republic; **M. Kárný**, Institute of Information Theory and Automation, Praha, Czech Republic; **D.H. Wolpert**, NASA, Moffett Field, CA, USA (Eds.)

Decision Making with Imperfect Decision Makers

- Why and how can imperfection be coped with in real life
- Proposes possible ways to approaches suitable to addressing design of decision strategies for and by imperfect designers and decision makers
- Edited outcome of best contributions to a NIPS'2010 workshop: "Decision Making with Imperfect Decision Makers"
- Written by leading experts in the field

Prescriptive Bayesian decision making has reached a high level of maturity and is well-supported algorithmically. However, experimental data shows that real decision makers choose such Bayes-optimal decisions surprisingly infrequently, often making decisions that are badly sub-optimal. So prevalent is such imperfect decision-making that it should be accepted as an inherent feature of real decision makers living within interacting societies.
To date such societies have been investigated from an economic and gametheoretic perspective, and even to a degree from a physics perspective. However, little research has been done from the perspective of computer science and associated disciplines like machine learning, information theory and neuroscience. This book is a major contribution to such research.
Some of the particular topics addressed include:
• How should we formalise rational decision making of a single imperfect decision maker?
• Does the answer change for a system of imperfect decision makers?
• Can we extend existing prescriptive theories for perfect decision makers to make them useful for imperfect ones?
• How can we exploit the relation of these problems to the control under varying and uncertain resources constraints as well as to the problem of the computational decision making?
• What can we learn from natural, engineered, and social systems to help us address these issues?

Order online at springer.com ► or for the Americas call (toll free) 1-800-SPRINGER ► or email us at: orders-ny@springer.com. ► For outside the Americas call +49 (0) 6221-345-4301 ► or email us at: orders-hd-individuals@springer.com.

The first € price and the £ and $ price are net prices, subject to local VAT. Prices indicated with * include VAT for books; the €(D) includes 7% for Germany, the €(A) includes 10% for Austria. Prices indicated with ** include VAT for electronic products; 19% for Germany, 20% for Austria. All prices exclusive of carriage charges. Prices and other details are subject to change without notice. All errors and omissions excepted.

*** Regional restrictions apply.