Towards Fully Probabilistic Cooperative Decision Making

Miroslav Kárný and Zohreh Alizadeh[⋆]

The Czech Academy of Sciences, Institute of Information Theory and Automation
POB 18, 182 08 Prague 8, Czech Republic, {school,za}@utia.cas.cz,
http://www.utia.cz/AS

Abstract. Modern prescriptive decision theories try to support the dynamic decision making (DM) in incompletely-known, stochastic, and complex environments. Distributed solutions single out as the only universal and scalable way to cope with DM complexity and with limited DM resources. They require a solid cooperation scheme, which harmonises disparate aims and abilities of involved agents (human decision makers, DM realising devices and their mixed groups). The paper outlines a distributed fully probabilistic DM. Its flat structuring enables a fully-scalable cooperative DM of adaptive and wise selfish agents. The paper elaborates the cooperation based on sharing and processing agents' aims in the way, which negligibly increases agents' deliberation effort, while preserving advantages of distributed DM. Simulation results indicate the strength of the approach and confirm the possibility of using an agent-specific feedback for controlling its cooperation.

Keywords: Decision Making · Cooperation · Fully Probabilistic Design · Bayesian Learning.

1 Introduction

A decision making theory supports agents to select actions, which aim to influence the closed decision loop, which couples the agent with its environment. DM is a complex process requiring a selection of relevant variables, adequate technical and theoretical tools, knowledge, aim elicitation, etc. Repeatedly-applicable DM procedures rely on a computer support, which needs a quantification of DM elements and, primarily, algorithmic solutions of all DM steps. Any of permanently-evolving DM theories designs a strategy (policy [25], decision function [38]), which maps the agent's knowledge and aims on actions. The optimal design selects the strategy, which meets agent's aims in the best way under the faced circumstances. An excessive DM complexity is tackled here.

Evolution singled out *distributed* DM as the only universal and fully scalable way of coping with DM complexity and limited resources of individual agents

[⋆] Supported by GAČR, grant 16-09848S. This work was strongly influenced by the feedback provided to us by Dr. Tatiana V. Guy. We appreciate her insight and help.

[4]. *Cooperation* of involved agents decides on the success or failure in reaching individual or collective aims [1,22,27,34,35,40]. It must not recur the problems, which make distributed DM inevitable. This limits applicability of theory of Bayesian games [10] and excludes presence of a cooperation-controlling mediator, who has to deal with a quite aggregated knowledge and a small number of actions. Thus, a personalised machine support of selfish (aim oriented) agents, dynamically acting in the changing environment containing other agents, is needed. To our best knowledge, no general support of this type exists. This paper contributes to its creation by inspecting cooperation within the discussed scenario. It relies on theory of fully probabilistic design of decision strategies (FPD, [9,15,36]). FPD is a proper extension [19] of prevailing Bayesian DM [7,29,39]. The paper deals with the cooperation, which assumes that the involved agents use FPD and are wise enough to cooperate to the degree required for achieving their selfish aims.

1.1 Paper Layout

The outline of the considered flat multi-agents system in Sec. 2 provides backbone of the subsequent text. Sec. 3 recalls a single adaptive agent that uses Bayesian learning and the feasible certainty-equivalent version of FPD. Sec. 4 describes the employed cooperation concept while commenting on its position with respect to its direct predecessors. The experimental part, Sec. 5, indicates soundness of the adopted concept. Remarks in Sec. 6 primarily outline the further anticipated research.

1.2 Notions and Notation

A simple DM task is considered in order to focus on the central cooperation problem. It is close to Markov decision processes [25] dealing with finite numbers of actions and of fully observable states. Throughout:

- The set of ys with $|\boldsymbol{y}| < \infty$ values y_j is denote $\boldsymbol{y} = \{y_j\}_{j=1}^{|\boldsymbol{y}|}$.
- The same symbol marks a random variable, its realisation, and its possible value. San serif fonts mark mappings. Mnemonic symbols are preferred.
- *Probability mass functions* (pmf) are implicitly conditioned on the known initial state $x_0 \in \boldsymbol{x}$.
- The *observable state* x_t of the modelled stochastic environment evolves in *discrete time* $t \in \boldsymbol{t}$. The evolution is influenced by optional *actions* $a \in \boldsymbol{a}$. A value $a_t \in \boldsymbol{a}$ of the action a is selected by the agent at time $t \in \boldsymbol{t}$.
- The closed decision loop, formed by the agent and its environment, operates on the *behaviour* $b = (x_t, a_t)_{t \in \boldsymbol{t}} \in \boldsymbol{b}$, i.e. on the collection of states and actions up to the decision *horizon* $|\boldsymbol{t}| < \infty$.
- The random behaviour $b \in \boldsymbol{b}$ is described by a joint pmf

$$\mathsf{c_s}(b) \equiv \mathsf{c_s}(b|x_0) = \mathsf{c_s}(x_{|\boldsymbol{t}|}, a_{|\boldsymbol{t}|}, x_{|\boldsymbol{t}|-1}, a_{|\boldsymbol{t}|-1}, \ldots, x_1, a_1|x_0).$$

It is the complete c*losed-loop model*. Chain rule for pmfs [28] factorises it

$$\begin{aligned}\mathsf{c_s}(b) &= \prod_{t\in \boldsymbol{t}} \mathsf{c_s}(x_t|a_t, x_{t-1}, a_{t-1}, \dots, a_1, x_0)\mathsf{c_s}(a_t|x_{t-1}, a_{t-1}, \dots, x_1, a_1, x_0)\\ &= \prod_{t\in \boldsymbol{t}} \mathsf{m}(x_t|a_t, x_{t-1})\mathsf{r}_t(a_t|x_{t-1}), \quad b = (x_t, a_t)_{t\in \boldsymbol{t}}. \end{aligned} \tag{1}$$

The mnemonically renamed factors after the second equality in (1): (a) exemplify the adopted assumption that the Markov environment and agent are considered; (b) focus us on time-invariant environments; (c) recognize that the first generic factor is the *environment* m*odel*, describing the probability of transiting to the state x_t from the state x_{t-1} when the action a_t is applied; (d) interpret the second generic factor as the *decision* r*ule*, which assigns the probability of selecting the action a_t, when knowing the state x_{t-1}.

- The optional, generally randomised, decision s*trategy* is the collection of decision rules $\mathsf{s} = (\mathsf{r}_t)_{t\in \boldsymbol{t}}$. The optimising DM selects the o*ptimal* s*trategy* $\mathsf{s_o}$. The optimality is defined with respect to agent's decision preferences, which are here quantified by the i*deal* c*losed-loop model*

$$\mathsf{c_i}(b) = \prod_{t\in \boldsymbol{t}} \mathsf{m_i}(x_t|a_t, x_{t-1})\mathsf{r_i}(a_t|x_{t-1}), \quad b = (x_t, a_t)_{t\in \boldsymbol{t}}. \tag{2}$$

It is the product of i*deal environment* m*odels* $\mathsf{m_i}$ and i*deal decision* r*ules* $\mathsf{r_i}$. Both are time invariant for simplicity. The ideal closed-loop model assigns high values to preferred behaviours and low values to unwanted ones. The use of this ideal pmf is in Sec. 3.

2 Flat Multi-Agents Systems

This section outlines the adopted concept of agents' interactions. The wish to support selfish imperfect agents motivates it. The adjective selfish implies that the agent follows its "personal" aims while the adjective "imperfect" labels the agent's limited knowledge, limited observation, evaluation, and acting abilities. Such an agent acts within the environment containing other imperfect selfish agents, which directly or indirectly influence the agent's degree of success or failure in reaching its personal aim. The considered wise but still imperfect agent takes it into account and makes public a part of information it deals with. This allows other agents to modify their strategy so that mutual inevitable clash is diminished and consequently, the considered interacting imperfect agents get chance to reach their individual unchanged aims in a better way.

The assumed common universal strategy-design methodology (FPD) and the common language (probabilistic descriptions of both environment and aims) allows to process the shared information without a special mediating or even facilitating agent, which would become bottleneck as it is always imperfect in the discussed sense.

The imperfection of each agent implies that it can reach only information provided by a small number of other agents, its recognisable neighbours. This makes the intended processing of the shared information feasible and the whole considered multi-agent interaction fully scalable.

It is worthy to add the following comments to the above outline.

– Aims of individual agents can be close and even identical. Thus, the explicit cooperation is well possible in this scheme.
– Individual agents can be created by a group of cooperating agents with its mediator, facilitator or leader up to the point where the joint resources suffice. Thus, the inspected multi-agent scenario may support all traditional multi-agent systems.
– The considered scheme imitates how complex societies act.
– Naturally, individual agents may publish misleading information and locally exploiting it may deteriorate quality of the strategy, which respects them. The agent that uses such an information can recognise this effect in a longer run and assign small weight (trust) to such an adversary agent.

The subsequent text considers such a flat scheme and describes firstly the considered type of agent and then the exploitation of the limited shared information, namely, shared description of neighbours' aims. Fixed trust to a neighbour is assumed at this research stage.

3 Single Agent Using FPD

This section focuses on single agent. It provides the FPD-optimal strategy along with its certainty-equivalent, receding-horizon approximation [23].

3.1 Formulation and Solution of Fully Probabilistic Design

An agent influences the closed-loop behaviour $b \in \boldsymbol{b}$ by selecting its randomised strategy s. Its choice shapes the closed-loop model (1). Ex post, Bayesian DM [29] evaluates the behaviour desirability (as seen by the agent) via a real-valued *loss function* $\mathsf{L}(b),\ b \in \boldsymbol{b}$. A priori, the quality of the strategy is evaluated via the expected loss, which is the generalised moment of the closed-loop model

$$\mathsf{E}_{\mathsf{s}}[\mathsf{L}] = \sum_{b \in \boldsymbol{b}} \mathsf{c}_{\mathsf{s}}(b)\mathsf{L}(b). \tag{3}$$

The Bayesian optimal strategy minimises the expected loss (3). FPD generalises this set up and uses the *ideal closed-loop model* c_{i} (2) instead of the loss function L. FPD selects the optimal strategy, which makes the closed-loop model closest to the given ideal closed-loop model. FPD axiomatisation [19] implies that Kullback-Leibler divergence (KLD, [21])

$$\mathsf{D}(\mathsf{c}_{\mathsf{s}}||\mathsf{c}_{\mathsf{i}}) = \mathsf{E}_{\mathsf{s}}\left[\ln\left(\frac{\mathsf{c}_{\mathsf{s}}}{\mathsf{c}_{\mathsf{i}}}\right)\right] = \sum_{b \in \boldsymbol{b}} \mathsf{c}_{\mathsf{s}}(b) \ln\left(\frac{\mathsf{c}_{\mathsf{s}}(b)}{\mathsf{c}_{\mathsf{i}}(b)}\right) \tag{4}$$

is the adequate proximity measure[1].

The universal loss (4) depends on the optimised strategy, unlike L (3). Thus, FPD defines *FPD-*o*ptimal* s*trategy* s_o as

$$\mathsf{s}_\mathsf{o} \in \operatorname{Arg}\min_{\mathsf{s}\in\boldsymbol{\mathsf{s}}} \mathsf{D}(\mathsf{c}_\mathsf{s}||\mathsf{c}_\mathsf{i}). \tag{5}$$

Algorithm 1 explicitly provides the FPD-optimal strategy s_o (5). It consists of the o*ptimal decision* r*ules* $(\mathsf{r}_{\mathsf{o};t})_{t\in\boldsymbol{t}}$ and exploits the environment model (1) and the factorised ideal closed-loop model (2). The proof of its optimality is, e.g., in [37]. Its presented form prepares the receding-horizon approximation of FPD.

Algorithm 1 Design of FPD-Optimal Decision Strategy

```
Inputs: Dimensions $|\boldsymbol{a}|$, $|\boldsymbol{x}|$; initial $\underline{\tau}$ and terminal $\bar{\tau}$ time moments of the design
        Environment model $\mathsf{m}$                              % belief description
        Factorised ideal closed model $\mathsf{c}_\mathsf{i} = \mathsf{m}_\mathsf{i}\mathsf{r}_\mathsf{i}$       % preference description
Evaluations:
Initialise $\mathsf{g}(x) = 1,\ \forall x \in \boldsymbol{x}$,
for $\tau = \bar{\tau}$ to $\underline{\tau}$ do
  for $x \in \boldsymbol{x}$ do
    for $a \in \boldsymbol{a}$ do
      $\mathsf{d}(a,x) = \sum_{\tilde{x}\in\boldsymbol{x}} \mathsf{m}(\tilde{x}|a,x) \ln\left(\frac{\mathsf{m}(\tilde{x}|a,x)}{\mathsf{m}_\mathsf{i}(\tilde{x}|a,x)\mathsf{g}(\tilde{x})}\right)$
    end for
    $\mathsf{g}(x) = \sum_{\tilde{a}\in\boldsymbol{a}} \mathsf{r}_\mathsf{i}(\tilde{a}|x)\exp(-\mathsf{d}(\tilde{a},x))$     % $-\ln(g(x))$ is the *value function*

    for $a \in \boldsymbol{a}$ do
      $\mathsf{r}_{\mathsf{o};\tau}(a|x) = \frac{\mathsf{r}_\mathsf{i}(a|x)\exp(-\mathsf{d}(a,x))}{\mathsf{g}(x)}$
    end for
  end for
end for
Outputs: FPD-optimal strategy $\mathsf{s}_\mathsf{o} = (\mathsf{r}_{\mathsf{o};\tau})_{\tau=\underline{\tau}}^{|\boldsymbol{\tau}|}$,   % $\mathsf{s}_\mathsf{o}$ is optimal iff $\underline{\tau} = 1,\ \bar{\tau} = |\boldsymbol{t}|$
```

3.2 Certainty-Equivalent Receding-Horizon FPD

The cooperation concept, inspected in Secs. 4, 5, assumes that the environment model is obtained by learning it. The model candidates are parameterised by time-invariant pmf values $\mathsf{m}(\tilde{x}|a,x,\theta) = \theta(\tilde{x}|a,x)$. The finite-dimensional parameter $\theta = (\theta(\tilde{x}|a,x))_{\tilde{x},x\in\boldsymbol{x},a\in\boldsymbol{a}}$ is unknown to the applied strategy. Thus, the strategy meets natural conditions of control [28] and the Bayesian learning can be used in the closed decision loop. The considered parametric environment model belongs to exponential family [3]. As such, it possesses self-reproducing Dirichlet's prior. Its finite-dimensional sufficient statistic is the occurrence array

[1] The axiomatisation [19] also shows that any Bayesian DM formulation can be approximated by an FPD formulation to an arbitrary precision.

$\mathsf{v} = (\mathsf{v}(\tilde{x}|a,x))_{\tilde{x},x\in\boldsymbol{x},a\in\boldsymbol{a}}$, $\mathsf{v} > 0$, [14]. The environment model corresponding to the knowledge accessible by the agent is the predictor

$$\mathsf{m}(\tilde{x}|a,x,\mathsf{v}) = \frac{\mathsf{v}(\tilde{x}|a,x)}{\sum_{\tilde{x}\in\boldsymbol{x}} \mathsf{v}(\tilde{x}|a,x)} = \hat{\theta}_{\mathsf{v}}(\tilde{x}|a,x), \quad \tilde{x}, x \in \boldsymbol{x}, a \in \boldsymbol{a}.$$

The pair $\mathcal{X} = (x, \mathsf{v})$ becomes the observable hyper-state so that Algorithm 1 is optimal with $\mathcal{X}$ replacing x. Practically, it is mostly infeasible as the time-varying value functions in Algorithm 1 depend on the hyper-state $\mathcal{X}$ of a huge dimension. At time $t \in \boldsymbol{t}$, feasibility is recovered by the standard certainty-equivalent approximation with receding horizon $h \leq |\boldsymbol{t}| - t$, see e.g. [23,24], and extensive references there. It replaces the unknown parameter θ in $\mathsf{m}(\tilde{x}|a,x,\theta)$ by the current point estimate $\hat{\theta}_{\mathsf{v}_{t-1}}$. $\mathsf{m}(\tilde{x}|a,x,\hat{\theta}_{\mathsf{v}_{t-1}})$ serves as the environment model during the strategy design, which runs backward for $\tau = \bar{\tau} = t+h-1, \ldots, \underline{\tau} = t$ till $t+h-1$, $h \leq |\boldsymbol{t}| - t + 1$. After applying the action a_t and observing the state x_t, the occurrence array v_{t-1} is updated, $\mathsf{v}_t(x_t|a_t, x_{t-1}) = \mathsf{v}_{t-1}(x_t|a_t, x_{t-1}) + 1$, and the procedure repeats, see Algorithm 2.

Algorithm 2 On-Line Certainty-Equivalent Receding-Horizon FPD

```
Inputs: Dimensions $|\boldsymbol{a}|$, $|\boldsymbol{x}|$, receding horizon $h$
        Initial state $x_0$, occurrence array $\mathsf{v}_0 > 0$          % prior belief description
        Factorised ideal closed model $\mathsf{c}_{\mathsf{i}} = \mathsf{m}_{\mathsf{i}}\mathsf{r}_{\mathsf{i}}$        % preference description
Evaluations:
for real time $t = 1$ to $|\boldsymbol{t}|$ do
    Get environment-model estimate $\mathsf{m}(\tilde{x}|a,x) = \frac{\mathsf{v}_{t-1}(\tilde{x}|a,x)}{\sum_{\tilde{x}\in\boldsymbol{x}} \mathsf{v}_{t-1}(\tilde{x}|a,x)}$, $\forall \tilde{x}, x \in \boldsymbol{x}$, $a \in \boldsymbol{a}$
    Get $(\mathsf{r}_{\mathsf{o};\tau})_{\tau=t}^{t+h-1}$=Algorithm 1$(|\boldsymbol{a}|, |\boldsymbol{x}|, \underline{\tau} = t, \bar{\tau} = \min(t+h-1, |\boldsymbol{t}|), \mathsf{m}, \mathsf{m}_{\mathsf{i}}, \mathsf{r}_{\mathsf{i}})$
    Sample action $a_t \sim \mathsf{r}_{\mathsf{o};t}(a|x_{t-1})$
 Closed-loop outputs: Applied action $a_t$, state $x_t$ observed on the environment
    Learn by updating the occurrence array $\mathsf{v}(x_t|a_t, x_{t-1}) = \mathsf{v}(x_t|a_t, x_{t-1}) + 1$
end for
```

Remarks

- Both algorithms have versions for continuous state and action spaces [15]. They are feasible for linear Gaussian models and their finite mixtures, [14,28].
- Algorithm 2 is presented in its rudimentary version. For instance, its computational complexity can be significantly decreased by iterations-spread-in-time strategy [14]. Its design omits the reset of g in Algorithm 1 and allows the use of the receding horizon close to one.
- The (approximately) optimal randomised FPD strategy is explorative. It is adaptive when employing forgetting [20], ideally, data-dependent as in [12].
- Undiscussed automated knowledge [6] and preference [5] elicitation would make Algorithm 2 (relatively) universal for single-agent DM. For a range of DM tasks, it is implementable into cheap portable devices. This makes the cooperation discussed in Sec. 4 realistically applicable.

4 Multiple Agents Sharing Ideal Closed-Loop Models

An agent mostly acts in the environment populated by other active agents. The agent should model them and respect their influence [10]. Such Bayesian games soon reach scalability limits as the learning and the strategy design become infeasible due to the quickly growing complexity of the handled DM elements.

The agent may ignore other agents and take them as non-modelled part of its environment. This feasible way may often lead to unfavourable results and calls for feasible countermeasures. Conceptually, the agent is to share an information with its *neighbours*. These are agents with whom its behaviour overlaps. Such sharing enables automated cooperation, negotiation [40] and conditions a conflict resolution. Quest for scalability admits only the information-sharing schemes working without a mediating center, i.e. a *flat cooperation structure*.

Agents exploiting FPD use the joint probabilistic ontology, which describes both their beliefs about environment and their DM preferences. This both enables generic flat cooperation schemes [16] and decreases the information sharing to a combination of probabilistic distributions, a classical pooling problem [8]. Among various possibilities, supra-Bayesian pooling fits the FPD framework. Its lack of a complete algorithmic solution is counteracted in [2,30,31,32]. These solutions are *impartial* with respect to the involved agents. They have led to a tuning-knob-free solution [13], which may serve as a “universal” impartial pooling solver.

Tests of impartial solutions were relatively successful. However, except the specialised case [18], they focused on static DM tasks. Also, the universality is not for free. The proposed solutions do not differentiate importance, strength and other specific properties of interacting agents. The decoupling of the processing of the shared information from the ultimate DM aim is the price paid for universality.

This criticism motivates the research whose basic steps are presented here.

4.1 Cooperation Circumstances

Opening the way towards filling the differentiation-gap left by predecessors [30,31,32] is the main paper aim. To focus on it, a simple, but well-generalisable, flat cooperation is treated. It concerns FPD-using agents in a common environment. Superscript ^{k}, $k \in \boldsymbol{k}$, marks *DM elements of the kth agent*: the behaviour $b^k \in \boldsymbol{b}^k$, the environment model m^k, the ideal closed-loop model c_i^k and its factors, i.e. the ideal environment model m_i^k, and the ideal decision rule r_i^k.

Inspected agents are neighbours of an agent. It means that its behaviour overlaps with behaviours of neighbours and the agent is aware of existence of common variables in them. In the considered case, the environment state $x \in \boldsymbol{x}$ is the commonly accessible behaviour part. The kth agent generates its optimised actions $a^k \in \boldsymbol{a}^k$. Others may at most observe it. This enhances the DM quality but it is unconsidered as it makes no conceptual difference. We focus on yet-untested sharing of information about ideal models, i.e. about neighbours’ preferences.

The inspected cooperation concerns *wise agents* who are willing to broadcast parts of pmfs (here ideal pfms) they use. Each agent utilises the information broadcasted by neighbours for the modification of its closed-loop ideal model employed in the strategy design. The agents remain *selfish*, and do not change their ideal closed-loop models according which they evaluate improvements achieved due to the information sharing.

4.2 Question Related to Pooling for FPD

The main questions encountered in pooling pmfs for FPD are:

1. How to pool the shared pmfs?
2. How to cope with the fact that behaviour sets of neighbours differ?
3. How to present the results to agents and how they should use them?
4. How to tune optional pooling parameters in order to support individual agents in approaching their disparate DM aims?

Questions 1, 2, 3 are mostly answered by predecessors, see below. Sec. 5 reflects the search for an insight indispensable for answering the open question 4.

Answer to question 1 follows [30,31,32]. The *pooled* ideal closed-loop model offered to kth agent $\tilde{\mathsf{c}}_{\mathsf{i}}^{k}$ is to be a convex combination of the processed pmfs

$$\tilde{\mathsf{c}}_{\mathsf{i}}^{k} = \sum_{j \in \boldsymbol{k}} \lambda_j^k \mathsf{c}_{\mathsf{i}}^{j}, \quad \lambda_j^k \geq 0, \;\; \sum_{j \in \boldsymbol{k}} \lambda_j^k = 1, \;\; \forall k \in \boldsymbol{k}. \tag{6}$$

This excludes, for instance, the popular geometric pooling [8]. The referred papers select the weights $(\lambda_j^k)_{j \in \boldsymbol{k}}$ uniquely using the involved prior pmf and the impartiality requirement. When relaxing the latter, the weights become optional and allow the reflection of the ultimate pooling aim: the support of agent's DM.

Answer to question 2: The combination (6) is meaningful iff all agents operate on the same behaviour $\boldsymbol{b} = \boldsymbol{b}^k$, $k \in \boldsymbol{k}$, i.e. iff the involved agents know and model all variables entering the neighbours' behaviours. This is definitely unrealistic. In the inspected case, this would imply to know and model the behaviour on the super-set $\boldsymbol{b}$ of behaviours treated by all neighbours

$$\boldsymbol{b} = \big(x_t, (a_t^j)_{j \in \boldsymbol{k}}\big)_{t \in \boldsymbol{t}}. \tag{7}$$

Thus, the pooling (6) can be applied iff the shared pmfs are *extended on* $\boldsymbol{b}$.

The original neighbours' pmfs could be interpreted as marginal pmfs of the constructed extensions. The extensions are, however, not unique as proved in connection with the copula theory [26]. Even more importantly, the combined pmfs are generically incompatible. Then no extension having them as marginal pmfs exists. It is well seen on the considered pooling of the ideal closed-loop models. In this case, the cooperation is to counteract the fact that selfish agents have different preferences with respect to the common environment state. This reflects that an agent wants to reach its specific closed-loop behaviour by assigning the highest probability to it by the personally-chosen ideal closed-loop model.

It implies that the extension is to be approached as a search for a compromise. The search is a supporting decision task with the extended pmfs being the optional actions, cf. [17,13]. In the considered case, the solution reduces to application of the maximum entropy principle [33] to time-invariant factors of the ideal closed-loop models. When requiring the preservation of the kth closed-loop model and individual agents' strategies, the gained e*xtension* $\mathsf{e}_{\mathsf{i}}^{k}$, $k \in \boldsymbol{k}$, ignores unknown influence of neighbours' actions on the kth state as well as their unknown dependence. The resulting kth extension reads

$$\mathsf{e}_{\mathsf{i}}^{k}(\tilde{x},(a^j)_{j\in\boldsymbol{k}}|x) = \mathsf{m}_{\mathsf{i}}^{k}(\tilde{x}|a^k,x)\prod_{j\in\boldsymbol{k}}\mathsf{r}_{\mathsf{i}}^{j}(a^j|x), \quad k \in \boldsymbol{k}. \tag{8}$$

Answer to question 3: The use of (6) to extensions (8) gives the pooled closed-loop ideal model on the super-set $\boldsymbol{b}$ (7) of the behaviour sets $\boldsymbol{b}^k$. The kth agent is uninterested and even unaware of actions a^j, $j \in \boldsymbol{k} \setminus \{k\}$, complementing $\boldsymbol{b}^k$ to $\boldsymbol{b}$. Thus, it makes sense to present this agent only the relevant marginal pmf $\tilde{\mathsf{c}}_{\mathsf{i}}^{k}$ of the pooled closed-loop ideal model. The result offered to kth agent is

$$\begin{aligned}\tilde{\mathsf{c}}_{\mathsf{i}}^{k}(\tilde{x},a^k|x) &= \Big[\lambda_k^k\mathsf{m}_{\mathsf{i}}^{k}(\tilde{x}|a^k,x) + \sum_{j\in\boldsymbol{k}\setminus\{k\}}\lambda_j^k\mathsf{f}_{\mathsf{i}}^{j}(\tilde{x}|x)\Big]\mathsf{r}_{\mathsf{i}}^{k}(a^k|x), \quad \text{where}\\ \mathsf{f}_{\mathsf{i}}^{j}(\tilde{x}|x) &= \sum_{a^j\in\boldsymbol{a}^j}\mathsf{m}_{\mathsf{i}}^{j}(\tilde{x}|a^j,x)\mathsf{r}_{\mathsf{i}}^{j}(a^j|x), \quad \tilde{x},x\in\boldsymbol{x}, \quad a^k\in\boldsymbol{a}^k.\end{aligned} \tag{9}$$

The *wise* agent k should use the ideal pmf $\tilde{\mathsf{c}}_{\mathsf{i}}^{k}$ (9) for *designing* its strategy.

Towards answering question 4 The algorithmic choice of the weight λ_j^k, which kth agent assigns to jth neighbour, is yet unsolved. The solution direction is, however, obvious. As said, the kth agent uses the pooled ideal closed-loop model when designing its approximation of the FPD-optimal strategy. It has at disposal its original ideal. Thus, it can evaluate the action quality, after using the designed action and after observing the realised environment state. This enables to relate the weights $(\lambda_j^k)_{j\in\boldsymbol{k}}$ to the reached DM quality and to design an additional feedback generating better weights for the subsequent design round.

A systematic design of the mentioned feedback is an important auxiliary DM task. Its solution needs a model relating the optional weights $(\lambda_j^k)_{j\in\boldsymbol{k}}$ to the observable DM quality, which is quantified by the reached value of the original closed-loop ideal model $\mathsf{c}_{\mathsf{i}}^{k}$. The extensive experiments, whose samples are in Sec. 5, primarily serve to the accumulation of experience needed for a feasible modelling of the relation of the weights to the truly reached DM quality.

Remarks

- The limited resources of an agent are helpful and make the solution scalable as the real agent has a small number $|\boldsymbol{k}|$ of recognised neighbours.
- The weights λ_j^k, $j \in \boldsymbol{k}$, are private for and specifically selected by kth agent.
- Pmf $\mathsf{f}_{\mathsf{i}}^{j}$ (9) is an action-independent, ideal f*orecaster* offered by jth agent.
- Adaptive learning is inevitable as the agent uses, almost by definition, a simplified model of its environment containing other active agents, [11].

- The pooled ideal closed-loop model should be modified at each real time moment. This allows an adaptation to the varying set of neighbours, their changing ideal forecasters, as well as (foreseen) data-dependent changes of the λ-weights, driven by the selfish preferences and built in personal $\mathsf{c}_{\mathsf{i}}^{j}$.
- Agents' selfishness implies that equilibrium, if reachable, will be of Nash's type. An analysis will be possible after operationally resolving question 4.
- The cooperation via sharing environment models is algorithmically identical and desirable [18]. It should be used jointly with the discussed one.
- Omissions of the mentioned ways to improve "cheaply" agent's performance is driven by the wish to preserve the presentation simplicity.

4.3 Algorithmic Summary

This part summarises the proposed fully-scalable cooperation of wise, selfish, FPD-using agents. It shows that the computational costs paid by an agent for this cooperation are small. Broadcasting the information about shared closed-loop ideals is the probably most demanding operation. It only needs to broadcast the ideal forecasters $(\mathsf{f}^{j})_{j\in\boldsymbol{k}}$ (9), possibly less often than the agents act.

Each agent $k \in \boldsymbol{k}$ acts according to Algorithm 3, which modifies its ideal closed-loop model $\mathsf{c}_{\mathsf{i}}^{k}$ to the pooled ideal $\tilde{\mathsf{c}}_{\mathsf{i}}^{k}$ (9). Otherwise, it coincides with Algorithm 2. The boxed $\boxed{\text{text}}$ in Algorithm 3 stresses the made changes.

```
Algorithm 3 FPD by Wise Selfish Cooperating Agent
Inputs: Agent's identifier $\underline{k} \in \boldsymbol{k}$, dimensions $|\boldsymbol{a}^k|$, $|\boldsymbol{x}|$, receding horizon $h^k$
        Initial state $x_0$, occurrence array $\mathsf{v}_0^k > 0$                  % prior belief description
        Factorised ideal closed model $\mathsf{c}_{\mathsf{i}}^k = \mathsf{m}_{\mathsf{i}}^k \mathsf{r}_{\mathsf{i}}^k$        % preference description
        The neighbours' ideal forecasters $\boxed{(\mathsf{f}_{\mathsf{i}}^{j})_{j\in\boldsymbol{k}\setminus\{k\}}}$   % of the state evolution
        The cooperation weights $\boxed{(\lambda_j^k)_{j\in\boldsymbol{k}}}$              % $\sum_{j\in\boldsymbol{k}} \lambda_j^k = 1,\ \lambda_j^k \geq 0$
Evaluations:
Get the pooled ideal $\boxed{\tilde{\mathsf{c}}_{\mathsf{i}}^k = \lambda_k^k \mathsf{m}_{\mathsf{i}}^k \mathsf{r}_{\mathsf{i}}^k + \sum_{j\in\boldsymbol{k}\setminus\{k\}} \lambda_j^k \mathsf{f}_{\mathsf{i}}^{j}}$   % cooperation
Outputs: $(a_{t^k}^k, x_{t^k})_{t^k\in\boldsymbol{t}^{\boldsymbol{k}}}$ =Algorithm 2$(|\boldsymbol{a}^k|, |\boldsymbol{x}|, h^k, x_0, \mathsf{v}_0^k, \underline{\tilde{\mathsf{c}}_{\mathsf{i}}^k})$
```

Remarks

- The agent identifier ^{k} delimits, which non-marginalised closed-loop ideal is used. Importantly, it stresses that all DM elements are fully under the agent's control, except of the environment state and external forecasters.
- The agent may work in a fully asynchronous mode and use a "personal" real time $t^k \in \boldsymbol{t}^{\boldsymbol{k}}$. This makes the advocated cooperation way quite flexible.

5 Experimental Part

The adopted concept is demonstrated on a simple well-understandable example. It exhibits all features of the general case and illustrates all notions used.

5.1 Simulation Set Up

The considered pair of agents, $k \in \boldsymbol{k} = \{1,2\}$ is interpreted as independent heaters influencing the common room temperature x. The quantised temperature is the observable state $x \in \boldsymbol{x} = \{1,\ldots,20\}$. Agents' actions are $a^k \in \boldsymbol{a}^k = \boldsymbol{a} = \{1,2\} \equiv \{\text{off,on}\}$. The closed-loop behaviours are $b^k = (x_t, a_t^k)_{t\in\boldsymbol{t}}$ up to $|\boldsymbol{t}| = 500$.

The room is the common simulated environment modeled by the transition probabilities $\pi(x_t|a_t^1, a_t^2, x_{t-1}))$, $x_t, x_{t-1} \in \boldsymbol{x}$, $a_t^1, a_t^2 \in \boldsymbol{a}$, $t \in \boldsymbol{t}$. They are obtained via quantisation of the linear Gaussian model with the conditional moments

$$x_t\text{-mean} = 0.65(a_t^1+a_t^2-2)+0.96x_{t-1}-0.02,\ x_t\text{-variance} = 0.25,\ x_0 = 10. \quad (10)$$

The constants in (10) are chosen to: (a) imitate a slow response of the heated room; (b) make the influence of both actions the same; (c) make the highest temperature $x = 20$ reachable when one agent is heating only; d) let the temperature fall to the lowest temperature $x = 1$ if both heaters are permanently off; and (e) let random effect be visible but not excessive.

A cooperation is vital as the *agents differ in* i*deal (desired) room temperatures*

$$x_{\mathsf{i}}^1 = 12, \quad x_{\mathsf{i}}^2 = 15.$$

The agents model their wishes by the ideal environment model, for $a^k =$ "on",

$$\mathsf{m}_{\mathsf{i}}^k(\tilde{x}|a^k = 2, x) = \begin{cases} 0.9 & \text{if } \ \tilde{x} = x_{\mathsf{i}}^k \\ 0.05 & \text{if } \ \tilde{x} = x_{\mathsf{i}}^k - 1 \\ 0.025 & \text{if } \ \tilde{x} = x_{\mathsf{i}}^k + 1 \\ \text{uniform} & \ \text{otherwise} \end{cases} \quad k \in \boldsymbol{k} = \{1,2\}.$$

For the actions $a^k =$ "off" probabilities of $\tilde{x} = x_{\mathsf{i}}^k \pm 1$ are swapped.

The ideal decision rules try to spare energy and prefer the action "off"

$$\mathsf{r}_{\mathsf{i}}^k(a^k = 1 = \text{"off"}|x^k) = \begin{cases} 0.9 & \text{if } \ x^k \geq x_{\mathsf{i}}^k \\ 0.1 & \text{if } \ x^k < x_{\mathsf{i}}^k \end{cases}, \quad k \in \boldsymbol{k} = \{1,2\}.$$

The agents recursively learn Markov models starting from the occurrence arrays

$$\mathsf{v}_0^k = (\text{the model (10) with the gain of the other action set to zero}) \times \nu^k.$$

The optional degrees of freedom $\nu^k > 0$ determine precision of the prior Dirichlet's distribution. The presented results correspond to the choice $\nu^k = 1$, $k \in \boldsymbol{k}$.

For $|\boldsymbol{k}| = 2$, each agent selects single cooperation weight $\lambda^k = \lambda_k^k$. The weights $\lambda_{j\neq k}^k = 1 - \lambda^k$. Simulations run for all pairs (λ^1, λ^2) on the grid $\lambda^k \in \boldsymbol{\lambda} = \{0, 0.1, \ldots, 0.9, 1.0\}$. The option $\lambda^k = 1$ means no cooperation. The kth agent accepts the ideal pmf of its neighbour as its own if $\lambda^k = 0$,

Each agent applies FPD, Algorithm 3, with the receding horizon $h = 5$ and the pooled ideal closed-loop model (6), $\tilde{x}, x \in \boldsymbol{x}$, $a^k \in \boldsymbol{a}^k$, for $\lambda^k \in \boldsymbol{\lambda}$, $k \in \boldsymbol{k} = \{1,2\}$,

$$\tilde{\mathsf{c}}_{\mathsf{i}}^1(\tilde{x}, a^1|x) = \Big[\lambda^1 \mathsf{m}_{\mathsf{i}}^1(\tilde{x}|a^1, x) + (1-\lambda^1) \sum_{a^2\in\boldsymbol{a}^2} \mathsf{m}_{\mathsf{i}}^2(\tilde{x}|a^2, x)\mathsf{r}_{\mathsf{i}}^2(a^2|x)\Big]\mathsf{r}_{\mathsf{i}}^1(a^1|x)$$

$$\tilde{\mathsf{c}}_{\mathsf{i}}^2(\tilde{x}, a^2|x) = \Big[\lambda^2 \mathsf{m}_{\mathsf{i}}^2(\tilde{x}|a^2, x) + (1-\lambda^2) \sum_{a^1 \in \boldsymbol{a}^1} \mathsf{m}_{\mathsf{i}}^1(\tilde{x}|a^1, x)\mathsf{r}_{\mathsf{i}}^1(a^1|x)\Big]\mathsf{r}_{\mathsf{i}}^2(a^2|x).$$

The reached q*uality* is judged via logarithm of the agent's ideal closed-loop model evaluated in the realised behaviour b^k. To make the result comparable with the value of the neighbour, this value is shifted by the absolute maximum

$$\mathsf{q}^k = \ln(\mathsf{c}_{\mathsf{i}}^k(b^k)) - \max_{b^k \in \boldsymbol{b}^k} \ln(\mathsf{c}_{\mathsf{i}}^k(b^k)), \ k \in \boldsymbol{k}. \tag{11}$$

5.2 Commented Results

The results are influenced by the inherent asymmetry of the problem: the agent may contribute to the temperature increase but the decrease depends only on the environment dynamics and on the realisation of random influences.

Figures 1, 2, and 3 present time courses of the room temperature. Fig. 1 corresponds to the cooperation coefficients $(\lambda^1, \lambda^2) = (0.1, 0.1)$ for which q^1 reaches its highest value. Fig. 2 corresponds to the cooperation coefficients $(\lambda^1, \lambda^2) = (0.8, 0.2)$ for which q^2 reaches its highest value. Fig. 3 corresponds to the combination of cooperation coefficients $(\lambda^1, \lambda^2) = (0.2, 0.3)$ for which the impartially judged joint quality $\mathsf{q}^1 + \mathsf{q}^2$ reaches its maximum.

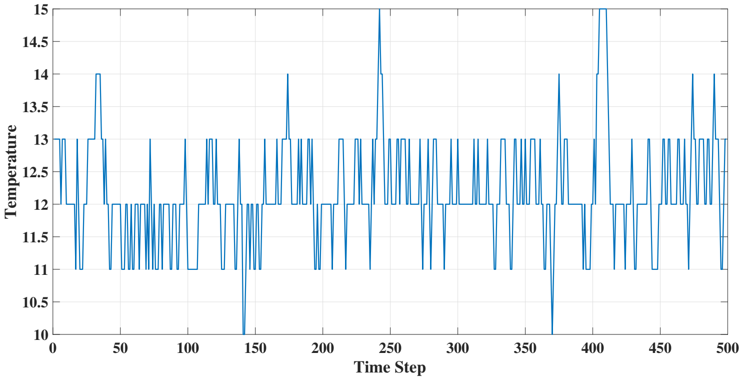

Fig. 1. The best temperature trajectory, maximising q^1 (11), for the agent k=1 wishing the temperature $x_{\mathsf{i}}^1 = 12$. It is reached for the weights $(\lambda^1, \lambda^2) = (0.1, 0.1)$.

The results primarily show that the value of the unchanged ideal closed-loop model in the measured data (11) is indeed a good indicator of the closed-loop quality from the agent's view point. This confirms the chance for a successful data-dependent choice of λ^k. Also, the example: (a) illustrates the theory; (b) confirms that the information sharing influences the achieved closed-loop behaviour; (c) shows asymmetry of the chosen environment; (d) indicates that possible Nash's equilibria could be searched around the maximum of $\sum_{k \in \boldsymbol{k}} \mathsf{q}^k$.

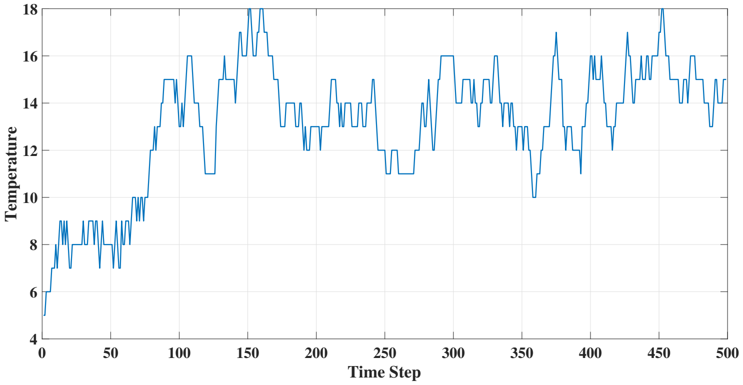

Fig. 2. The best temperature trajectory, maximising q^2 (11), for the agent k=2 wishing the temperature $x_{\mathsf{i}}^2 = 15$. It is reached for the weights $(\lambda^1, \lambda^2) = (0.8, 0.2)$.

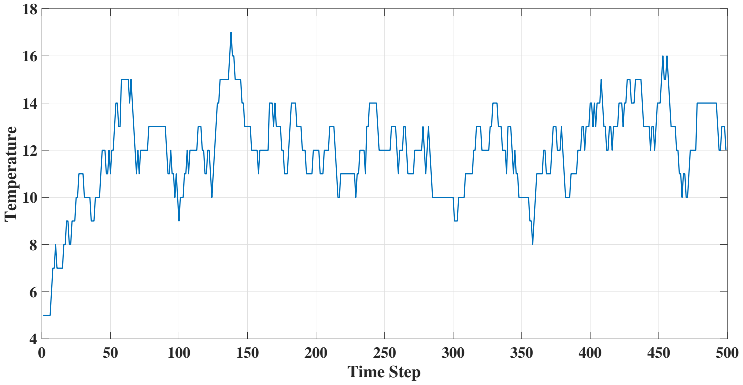

Fig. 3. The temperature trajectory, which is expected to be the best reachable compromise, maximising $\mathsf{q}^1 + \mathsf{q}^2$ (11), between wishes $x_{\mathsf{i}}^1 = 12$ d $x_{\mathsf{i}}^2 = 15$. It is reached for the weights $(\lambda^1, \lambda^2) = (0.2, 0.3)$.

6 Concluding Remarks

This paper inspects a cooperation methodology for a flatly-interacting multiple agents. It is based on sharing of ideal closed-loop models for FPD. It confirms the chance to adapt the cooperation-controlling weights in closed loop according to selfish aims of respective agents. At the same time, the accumulated experience demonstrates that sharing solely some DM elements does not guarantee a high decision quality. Other DM elements, as the learnt environment model, have to be shared. FPD makes it an identical task. Also, other unused possibilities as prior knowledge elicitation, forgetting, exploitation/exploration balance have to be exploited to get a robust practical tool. These measures will be addressed in near future. Good news is that this way is promising and feasible. All induced tasks are solvable at single-agent level with a negligible deliberation overhead caused by acting in multi-participant environment. The paper exemplifies this.

References

1. Aknine, S., Caillou, P., Pinson, S.: Searching Pareto optimal solutions for the problem of forming and restructuring coalitions in multi-agent systems. Group Decision and Negotiation **19**, 7–37 (2010)
2. Azizi, S., Quinn, A.: Hierarchical fully probabilistic design for deliberator-based merging in multiple participant systems. IEEE Transactions on Systems, Man, and Cybernetics **PP**(99), 1–9 (2016). https://doi.org/10.1109/TSMC.2016.2608662
3. Barndorff-Nielsen, O.: Information and Exponential Families in Statistical Theory. Wiley, NY (1978)
4. Bond, A., Gasser, L.: Readings in Distributed Artificial Intelligence. Morgan Kaufmann (2014)
5. Chen, L., Pu, P.: Survey of preference elicitation methods. Tech. Rep. IC/2004/67, HCI Group Ecole Politechnique Federale de Lausanne, Switzerland (2004)
6. Daee, P., Peltola, T., Soare, M., Kaski, S.: Knowledge elicitation via sequential probabilistic inference for high-dimensional prediction. Machine Learning **106**(9), 1599–1620 (2017)
7. DeGroot, M.: Optimal Statistical Decisions. McGraw-Hill, NY (1970)
8. Genest, C., Zidek, J.: Combining probability distributions: A critique and annotated bibliography. Statistical Science **1**(1), 114–148 (1986)
9. Guan, P., Raginsky, M., Willett, R.: Online Markov decision processes with Kullback Leibler control cost. IEEE Trans. on Automatic Control **59**(6), 1423–1438 (2014)
10. Harsanyi, J.: Games with incomplete information played by Bayesian players, I–III. Management Science **50**(12) (2004), supplement
11. Kárný, M.: Adaptive systems: Local approximators? In: Workshop on Adaptive Systems in Control and Signal Processing. pp. 129–134. IFAC, Glasgow (1998)
12. Kárný, M.: Recursive estimation of high-order Markov chains: Approximation by finite mixtures. Infor. Sciences **326**, 188–201 (2016)
13. Kárný, M.: Implementable prescriptive decision making. In: Guy, T., Kárný, M., D., D.R.I., Wolpert, D. (eds.) Proceedings of the NIPS 2016 Workshop on Imperfect Decision Makers. vol. 58, pp. 19–30. JMLR (2017)
14. Kárný, M., Böhm, J., Guy, T.V., Jirsa, L., Nagy, I., Nedoma, P., Tesař, L.: Optimized Bayesian Dynamic Advising: Theory and Algorithms. Springer (2006)
15. Kárný, M., Guy, T.V.: Fully probabilistic control design. Systems & Control Letters **55**(4), 259–265 (2006)
16. Kárný, M., Guy, T.V., Bodini, A., Ruggeri, F.: Cooperation via sharing of probabilistic information. Int. J. of Comput. Intelligence Studies pp. 139–162 (2009)
17. Kárný, M., Guy, T.: On support of imperfect Bayesian participants. In: Guy, T., et al (eds.) Decision Making with Imperfect Decision Makers, vol. 28. Springer, Berlin (2012), Intelligent Systems Reference Library
18. Kárný, M., Herzallah, R.: Scalable harmonization of complex networks with local adaptive controllers. IEEE Trans. on SMC: Systems **47**(3), 394–404 (2017)
19. Kárný, M., Kroupa, T.: Axiomatisation of fully probabilistic design. Infor. Sciences **186**(1), 105–113 (2012)
20. Kulhavý, R., Zarrop, M.B.: On a general concept of forgetting. Int. J. of Control **58**(4), 905–924 (1993)

21. Kullback, S., Leibler, R.: On information and sufficiency. Annals of Mathematical Statistics **22**, 79–87 (1951)
22. Lewicki, R., Weiss, S., Lewin, D.: Models of conflict, negotiation and 3rd party intervention - a review and synthesis. J. of Organ. Behavior **13**(3), 209 – 252 (1992)
23. Mattingley, J., Wang, Y., Boyd, S.: Receding horizon control. IEEE Control Systems Magazine **31**(3), 52 – 65 (2011)
24. Mayne, D.: Model predictive control: Recent developments and future promise. Automatica pp. 2967–2986 (2014)
25. Mine, H., Osaki, S.: Markovian Decision Processes. Elsevier (1970)
26. Nelsen, R.: An Introduction to Copulas. Springer, NY (1999)
27. Nurmi, H.: Resolving group choice paradoxes using probabilistic and fuzzy concepts. Group Decision and Negotiation **10**, 177–198 (2001)
28. Peterka, V.: Bayesian system identification. In: Eykhoff, P. (ed.) Trends and Progress in System Identification, pp. 239–304. Pergamon Press, Oxford (1981)
29. Savage, L.: Foundations of Statistics. Wiley, NY (1954)
30. Sečkárová, V.: Cross-entropy based combination of discrete probability distributions for distributed decision making. Ph.D. thesis, Charles University in Prague, Faculty of Mathematics and Physics, Dept. of Probability and Mathematical Statistics., Prague (2015), submitted in May 2015. Successfully defended on 14.09.2015.
31. Sečkárová, V.: Weighted probabilistic opinion pooling based on cross-entropy. In: Neural Information Processing, 22nd International Conference, ICONIP 2015, November 9-12, 2015, Proceedings, Part II. pp. 623–629. Istanbul, Turkey (2015)
32. Sečkárová, V.: On supra-Bayesian weighted combination of available data determined by Kerridge inaccuracy and entropy. Pliska Stud. Math. Bulgar. **22**, 159–168 (2013)
33. Shore, J., Johnson, R.: Axiomatic derivation of the principle of maximum entropy & the principle of minimum cross-entropy. IEEE Tran. on Inf. Th. **26**(1), 26–37 (1980)
34. Simpson, E.: Combined Decision Making with Multiple Agents. Ph.D. thesis, Hertford College, Dept. of Eng. Sci., University of Oxford (2014)
35. Simpson, E., Roberts, S., Psorakis, I., Smith, A.: Dynamic Bayesian combination of multiple imperfect classifiers. In: Guy, T., Kárný, M., Wolpert, D. (eds.) Decision Making and Imperfection, vol. 28, pp. 1–38. Springer, Berlin (2013), studies in Computation Intelligence
36. Todorov, E.: Linearly-solvable Markov decision problems. In: Schölkopf, B., et al (eds.) Advances in Neural Inf. Processing, pp. 1369 – 1376. MIT Press, NY (2006)
37. Šindelář, J., Vajda, I., Kárný, M.: Stochastic control optimal in the Kullback sense. Kybernetika **44**(1), 53–60 (2008)
38. Wald, A.: Statistical Decision Functions. John Wiley, NY, London (1950)
39. Wallenius, J., Dyer, J., Fishburn, P., Steuer, R., Zionts, S., Deb, K.: Multiple criteria decision making, multiattribute utility theory: Recent accomplishments and what lies ahead. Management Sci. **54**(7), 1336–1349 (2008)
40. Zlotkin, G., Rosenschein, J.: Mechanism design for automated negotiation and its applicatin to task oriented domains. Artificial Intelligence **86**, 195–244 (1996)