Chapter 3
Automated Preference Elicitation for Decision Making

Miroslav Kárný

Abstract. In the contemporary complex world decisions are made by an imperfect participant devoting limited deliberation resources to any decision-making task. A normative decision-making (DM) theory should provide support systems allowing such a participant to make rational decisions in spite of the limited resources. Efficiency of the support systems depends on the interfaces enabling a participant to benefit from the support while exploiting the gradually accumulating knowledge about DM environment and respecting incomplete, possibly changing, participant's DM preferences. The insufficiently elaborated preference elicitation makes even the best DM supports of a limited use. This chapter proposes a methodology of automatic eliciting of a quantitative DM preference description, discusses the options made and sketches open research problems. The proposed elicitation serves to fully probabilistic design, which includes a standard Bayesian decision making.

Keywords: Bayesian decision making, fully probabilistic design, DM preference elicitation, support of imperfect participants.

3.1 Introduction

This chapter concerns of an imperfect participant[1], which solves a real-life decision-making problem under uncertainty, which is worth of its optimising effort. The topic has arisen from the recognition that a real participant often cannot benefit from sophisticated normative DM theories due to an excessive deliberation effort needed

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

[1] A participant is also known as user, decision maker, agent. A participant can be human, an artificial object or a group of both. We refer the participant as "it".

T.V. Guy et al. (Eds.): *Decision Making and Imperfection*, SCI 474, pp. 65–99.
DOI: 10.1007/978-3-642-36406-8_3

for their mastering and for feeding them by DM elements[2] they need. This observation has stimulated a long-term research, which aims to equip a participant with automatic tools (intelligent interfaces) mapping its knowledge, DM preferences and constraints on DM elements while respecting its *imperfection*, i.e. ability to devote only a limited deliberation effort to a particular DM task. The research as well as this chapter concentrates on the Bayesian DM theory because of its exceptional, axiomatically justified, role in DM under uncertainty, e.g. [49].

The adopted *concept of the ultimate solution* considers creating an automated supporting system, which covers the complete design and use of a decision-generating DM strategy. It has to preserve the theoretically reachable DM quality and free the participant's cognitive resources to tasks specific to its application domain. This concept induces:

Requirement 1: The supporting system uses a consistent and complete DM theory.

Requirement 2: To model the environment[3] the supporting system fully exploits both participant's knowledge and information brought by the data observed during the use of the DM strategy.

Requirement 3: The supporting system respects participant's DM preferences and refines their description by the information gained from the observed data.

This chapter represents a further step to the ultimate solution. It complements the results of the chapter [28] devoted to DM of imperfect participants. The tools needed for a conceptual solution are based on a generalised Bayesian DM, called *fully probabilistic design* (FPD) [23, 29], see also Section 3.2. The FPD minimises Kullback-Leibler divergence [38] of the optimised, strategy-dependent probabilistic model of the closed DM loop on its ideal counterpart. which describes the desired behaviour of the closed decision loop. The design replaces the maximisation of an expected utility over a set of admissible decision strategies [4] for the FPD densely extending all standard Bayesian DM formulations [31]. The richness, intuitive plausibility, practical advantages and axiomatic basis of the FPD motivate its acceptance as a unified theoretical DM basis, which meets Requirement 1.

Requirement 2 concerns the description of the environment with which the participant interacts during DM. Traditionally, its construction splits into a structural and semi-quantitative modelling of the environment and *knowledge elicitation* understood as a quantitative description of unknown variables entering the environment model. Both activities transform domain-specific knowledge into the model-related part of DM elements.

The environment modelling is an extremely wide field that exploits first principles (white and grey box models, e.g. [6, 22]), application-field traditions, e.g. [9],

[2] This is a common label for all formal, qualitatively and quantitatively specified, objects needed for an exploitation of the selected normative DM theory.

[3] The environment, also called system, is an open part of the World considered by the participant and with which it interacts within the solved DM task.

universal approximation (black box models, e.g. [17, 50]) and their combinations. An automated mapping of these models on probabilistic DM elements of the FPD is the expected service of the supporting DM system. The tools summarised in [28] are conjectured to be sufficient to this purpose. Knowledge elicitation in the mentioned narrow sense is well surveyed in [14, 45] and automated versions related to this chapter are in [24, 25]. The ordinary Bayesian framework [4, 47] adds the required ability to learn from the observed data.

Requirement 3 reflects the fact that a feasible and effective solution of *preference elicitation* problem decides on the efficiency of any intelligent system supporting DM. This extracting the information about the participant's DM preferences has been recognised as a vital problem and repeatedly addressed within artificial intelligence, game theory, operation research. Many sophisticated approaches have been proposed [10, 11, 13, 16], often in connection with applied sciences like economy, social science, clinical decision making, transportation, see, for instance, [21, 41].

Various assumptions on the structure of DM preferences have been adopted in order to ensure feasibility and practical applicability of the resulting decision support. Complexity of the elicitation problem has yet prevented to get a satisfactory widely-applicable solution. For instance, the conversion of DM preferences on individual observable decision-quality-reflecting *attributes* into the overall DM preferences is often done by assuming their additive independence [33]. The DM preferences on attributes are dependent in majority of applications and the enforced independence assumption significantly worsens the elicitation results[4]. This example indicates a deeper drawback of the standard Bayesian DM, namely, the lack of unambiguous rules how to combine low-dimensional description of DM preferences into a global one.

The inability of the participant to completely specify its DM preferences is another problem faced. In this common case, the DM preferences should be learned from either domain-specific information (technological requirements and knowledge, physical laws, etc.) or the observed data.

Eliciting the needed information itself is an inherently difficult task, which success depends on experience and skills of an elicitation expert. The process of eliciting of the domain-specific information is difficult, time-consuming and error-prone activity[5]. Domain experts provide subjective opinions, typically expressed in different forms. Their processing requires a significant cognitive and computational effort of the elicitation expert. Even if the cost of this effort[6] is negligible, the elicitation

[4] The assumption can be weakened by a introducing a conditional preferential independence, [8].

[5] 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. The high number of parameters makes this solution unfeasible and enforces attempts to decrease the number to recover feasibility.

[6] 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, a decision quality is considered), see for instance [7].

result is always limited by the expert's imperfection, i.e. his/her inability to devote an unlimited deliberation effort to eliciting. Unlike the imperfection of experts providing the domain-specific information, the imperfection of elicitation experts can be eliminated by preparing a feasible *automated* support of the preference elicitation, that does not rely on any elicitation expert.

The dynamic decision making strengthes the dependence of the DM quality on the preference elicitation. Typically, the participant acting within a dynamically changing environment with evolving parameters gradually changes its DM preferences. The change may depend on the expected future behaviour or other circumstances. The overall task is getting even harder when the participant dynamically interacts with other imperfect participants within a common environment. When DM preferences evolve, their observed-data-based learning becomes vital.

The formal disparity of modelling language (probabilities) and the DM preference description (utilities) makes Bayesian learning of DM preferences difficult. It needs a non-trivial "measurement" of participant's satisfaction of the decision results, which often puts an extreme deliberation load on the participant. Moreover, the degree of satisfaction must be related to conditions under which it has been reached. This requires a non-standard and non-trivial modelling. Even, if these learning obstacles are overcome, the space of possible behaviour is mostly larger than that the observed data cover. Then, the initial DM preferences for the remaining part of the behaviour should be properly assigned and exploration has to care about making the DM preference description more precise. Altogether, a weak support of the preference elicitation (neglecting of Requirement 3) is a significant gap to be filled. Within the adopted FPD, an ideal probability density (pd[7]) is to be elicited[8]. The ideal pd describes the closed-loop behaviour, when the participant's DM strategy is an optimal one and the FPD searches for the optimal randomised strategy minimising the divergence from the current closed-loop description to the ideal one.

Strengthening the support with respect to Requirement 3 forms the core of this chapter. The focus on the preference elicitation for the FPD brings immediate methodological advantages. For instance, the common probabilistic language for knowledge and DM preference descriptions simplifies an automated elicitation as the ideal pd provides a standardised form of quantitatively expressed DM preferences. Moreover, the raw elicitation results reflect inevitably incomplete, competitive or complementing opinions with respect to the same collection of DM preference-expressing multivariate attributes. Due to their automated mapping on probabilities, their logically consistent merging is possible with the tools described in [28]. Besides, domain experts having domain-specific information are often

[7] Radon-Nikodým derivative [48] of the strategy-dependent measure describing closed DM loop with respect to a dominating, strategy-independent measure. The use of this notion helps us to keep a unified notation that covers cases with mixed – discrete and continuous valued – variables.

[8] Let us stress that no standard Bayesian DM is omitted due to the discussed fact that the FPD densely covers all standard Bayesian DM tasks.

unable to provide their opinion on a part of behaviour due to either limited knowledge of the phenomena behind or the indifference towards the possible instances of behaviour. Then, the DM preference description has to be extended to the part of behaviour not "covered" by the domain-specific information. This extension is necessary as the search for the optimal strategy heavily depends on the full DM preference description. It is again conceptually enabled by the tools from [28]. The usual Bayesian learning is applicable whenever the DM preferences are related to the observed data [27].

In summary, the chapter concerns a construction of a probabilistic description of the participant's DM preferences based on the available information. Decision making under uncertainty is considered from the perspective of an imperfect participant. It solves a DM task with respect to its environment and indirectly provides a finite description of the DM preferences in a non-unique way[9] and leaves uncertainty about the DM preferences on a part of closed-loop behaviour. To design an optimal strategy, the participant employs the fully probabilistic design of DM strategies [23, 29] whose DM elements are probability densities used for the environment modelling, DM preference description and description of the observed data.

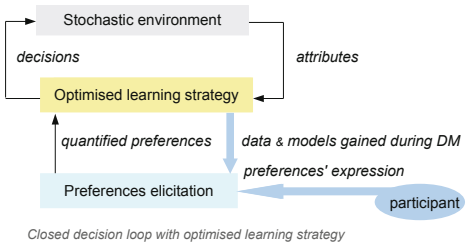
The explanations prefer discussion of the solution aspects over seemingly definite results. After a brief summary of common tools Section 3.2, they start with a problem formalisation that includes the basic adopted assumptions, Section 3.3. The conceptual solution summarised in Section 3.4 serves as a guide in the subsequent extensive discussion of its steps in Section 3.5. Section 3.6 provides illustrative simulations and Section 3.7 contains concluding remarks.

Concept of the proposed preference elicitation is reflected in Figure[10] 3.1. The usual decision loop formed by a stochastic environment and a decision strategy complemented by a preference elicitation block is expanded to the proposed solution. The considered strategy consists of the standard Bayesian learning of the environment model and of a standard fully probabilistic design (FPD). Its explicit structuring reveals the need of the ideal closed-loop model of the desired closed-loop behaviour. The designed strategy makes the closed decision loop closest to this ideal model, which is generated by the elicitation block as follows. The observed data is censored[11] to data, which contains an information about the optimal strategy and serves for its Bayesian learning. The already learnt environment model is combined with the gained model of the optimal strategy into the model of the DM loop closed by it. Within the set of closed-loop models, which comply with the participant's DM preferences and are believed to be reachable, the ideal closed-loop model is selected as the nearest one to the learnt optimal closed-loop model.

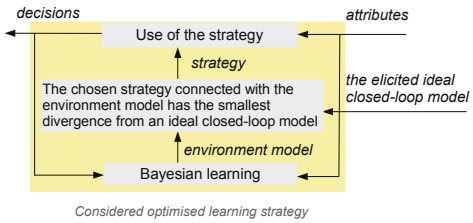
[9] Even, when we identify instances of behaviour that cannot be preferentially distinguished.

[10] A block performing the inevitable knowledge elicitation is suppressed to let the reader focus on the proposed preference elicitation.

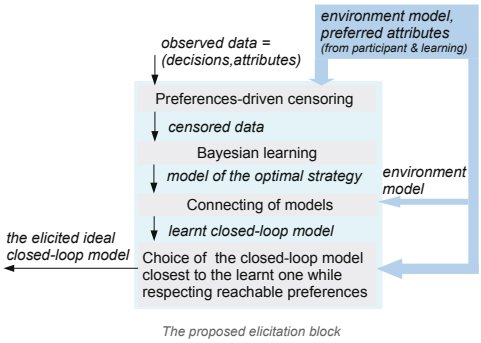
[11] Such data processing is also called filtering. This term is avoided as it has also another meaning.



(a)



(b)



(c)

Fig. 3.1 The figure 3.1a displays a closed decision loop with an optimised learning strategy. The figure 3.1b expands the considered optimised learning strategy that uses the FPD and Bayesian learning. The figure 3.1c shows the *proposed elicitation block*. The observed data is censored to reveal information about an unknown optimal strategy. The Bayesian learning on the censored data provides the model of the optimal strategy, which together with the learnt environment model describes the optimally closed loop. The elicitation block selects the ideal closed-loop model as the model, which: (i) complies with participant's DM preferences; (ii) is reachable by an available strategy; (iii) is the nearest one to the model of the optimal closed loop.

Notation

General conventions

$\boldsymbol{x}$	is a set of x-values having cardinality $\lvert\boldsymbol{x}\rvert$
$d \in \boldsymbol{d}, \boldsymbol{d} \neq \emptyset$	are decisions taken from a finite-dimensional set $\boldsymbol{d}$
$a_i \in \boldsymbol{a}_i$, $i \in \boldsymbol{i} = \{1, \ldots, \lvert\boldsymbol{i}\rvert\}$	are attribute entries in finite-dimensional sets $\boldsymbol{a}_i$
$a \in \boldsymbol{a}$	is a collection of all attributes in the set $\boldsymbol{a} = \mathsf{X}_{i \in \boldsymbol{i}} \boldsymbol{a}_i$,
X	denotes the Cartesian product
$\boldsymbol{\alpha} \subsetneq \boldsymbol{a}, \boldsymbol{\alpha} \neq \emptyset$	is the set of the most desirable attribute values specified entry-wise $\boldsymbol{\alpha} = \mathsf{X}_{i \in \boldsymbol{i}} \boldsymbol{\alpha}_i$
$t \in \boldsymbol{t} = \{1, \ldots, \lvert\boldsymbol{t}\rvert\}$	is discrete time
$(x_t)_{t \in \boldsymbol{t}}$	is a sequence of x_t indexed by discrete time $t \in \boldsymbol{t}$.

Probability densities

$\mathsf{g}(\cdot), \mathsf{h}(\cdot)$	are probability densities (pds): Radon-Nikodým derivatives with respect to a dominating measure denoted $\mathrm{d}\cdot$
$\mathsf{M}_t(a\lvert d)$, $\mathsf{M}(a\lvert d, \Theta)$	are the environment model and its parametric version with an unknown parameter $\Theta \in \boldsymbol{\Theta}$
$\mathsf{F}_t(\Theta)$, $t \in \boldsymbol{t} \cup \{0\}$	is the pd quantifying knowledge available at time t about the unknown parameter Θ of the environment model
$\mathsf{S}_t(d)$	describes the randomised decision strategy to be selected
$\mathsf{s}_t(d)$, $\mathsf{s}(d\lvert\theta)$	are the model of the optimal strategy and its parametric version with an unknown parameter $\theta \in \boldsymbol{\theta}$
$\mathsf{f}_t(\theta)$, $t \in \boldsymbol{t} \cup \{0\}$	is the pd quantifying knowledge available at time t about the unknown parameter $\theta \in \boldsymbol{\theta}$ of the optimal strategy
$\mathsf{I}_t(a, d)$	is the ideal pd quantifying the elicited participant's DM preferences
$\mathsf{P}_t(a, d)$	is the pd modelling the decision loop with the optimal strategy
$\mathsf{M}_t(a_i\lvert d)$, $\mathsf{I}_t(a_i)$, $i \in \boldsymbol{i}$	are marginal pds of a_i derived from pds $\mathsf{M}_t(a\lvert d)$ and $\mathsf{I}_t(a, d)$

Convention on time indexing

$\mathsf{F}_{t-1}(\Theta)$, $\mathsf{f}_{t-1}(\theta)$	quantify knowledge accumulated before time t
$\mathsf{M}_t(a\lvert d)$, $\mathsf{s}_t(d)$, $\mathsf{S}_t(d)$	serve to the tth DM task and exploit
$\mathsf{I}_t(a, d)$, $\mathsf{P}_t(a, d)$	the knowledge accumulated before time t.

Frequent symbols

$\vec{d} \in \boldsymbol{d}$	is a decision leading to a high probability of the set $\boldsymbol{\alpha}$
$\mathbb{D}(\mathsf{h}\lVert\mathsf{g})$	is the Kullback-Leibler divergence (KLD, [38]) of a pd h from a pd g
$\mathbb{E}[\cdot]$	denotes expectation
V	is a sufficient statistic of the exponential family (EF), which becomes the occurrence table in Markov-chain case
$\phi \in [0, 1]$	is a forgetting factor
$\propto$	denotes an equality up to normalisation.

3.2 Preliminaries

Introduction repeatedly refers to the tools summarised in [28]. Here, we briefly recall its sub-selection used within this chapter.

1. The *Kullback-Leibler divergence* (KLD, [38]) $\mathbb{D}(\mathsf{g}||\mathsf{h})$ of a pd g from a pd h, both defined on a set $\boldsymbol{x}$ and determined by a dominating strategy-independent measure $\mathrm{d}x$, is defined by the formula

$$\mathbb{D}(\mathsf{g}||\mathsf{h}) = \int_{\boldsymbol{x}} \mathsf{g}(x) \ln\left(\frac{\mathsf{g}(x)}{\mathsf{h}(x)}\right) \mathrm{d}x. \tag{3.1}$$

   The KLD is a convex functional in the pd g, which reaches its smallest zero value iff $\mathsf{g} = \mathsf{h}$ $\mathrm{d}x$-almost everywhere.
   $\mathbb{D}(\mathsf{g}||\mathsf{h}) \neq \mathbb{D}(\mathsf{h}||\mathsf{g})$ and a correct pd should be used as its first argument when measuring (di)similarity of pds by the KLD. A pd is called *correct*[12] if it fully exploits the knowledge about the random variable it models. Its existence is assumed.
2. Under the commonly met conditions [5], the *optimal Bayesian approximation* $\mathsf{h}^o \in \mathbf{h}$ of a correct pd g by a pd $\mathsf{h} \in \mathbf{h}$ should be defined

$$\mathsf{h}^o \in \mathrm{Arg}\min_{\mathsf{h}\in\mathbf{h}} \mathbb{D}(\mathsf{g}||\mathsf{h}). \tag{3.2}$$

3. The *minimum KLD principle* [28, 51] recommends to select a pd $\mathsf{h}^e \in \mathbf{h}$

$$\mathsf{h}^e \in \mathrm{Arg}\min_{\mathsf{h}\in\mathbf{h}} \mathbb{D}(\mathsf{h}||\mathsf{g}) \tag{3.3}$$

   as an extension of the available information about the correct pd h. The assumed available information consists of a given set $\mathbf{h}$ and of a rough (typically flat) estimate g of the pd h.
   The minimum KLD principle provides such an extension of the available information that the pd h^e deviates from its estimate g only to the degree enforced by the constraint $\mathsf{h} \in \mathbf{h}$. It reduces to the celebrated maximum entropy principle [20] for the uniform pd g.
   The paper [51] axiomatically justifies the minimum KLD principle for sets $\mathbf{h}$ delimited by values of h moments. The generalisation in [28] admits a richer collection of the sets $\mathbf{h}$. For instance, the set $\mathbf{h}$ can be of the form

$$\mathbf{h} = \left\{\mathsf{h} : \mathbb{D}\left(\mathsf{h}||\hat{\mathsf{h}}\right) \leq k < \infty\right\}, \tag{3.4}$$

   determined by a given pd $\hat{\mathsf{h}}$ and by a positive constant k.
   For the set (3.4), the pd h^e (3.3) can be found by using the Kuhn-Tucker optimality conditions [35]. The solution reads

$$\mathsf{h}^e \propto \hat{\mathsf{h}}^{\phi}\mathsf{g}^{1-\phi},\ \phi \in [0,1], \tag{3.5}$$

[12] This is an operational notion unlike often used adjectives "true" or "underlying".

where $\propto$ denotes an equality up to a normalisation factor and ϕ is to be chosen to respect the constraint (3.4). The solution formally coincides with the so-called stabilised forgetting [37] and ϕ is referred as *forgetting factor*.

3.3 Problem Formalisation

The considered participant repeatedly solves a sequence of static[13] DM tasks indexed by *(discrete) time* $t \in \boldsymbol{t} = \{1, 2, \ldots, |\boldsymbol{t}|\}$. DM concerns a stochastic, incompletely-known, time-invariant static environment. The decision d influencing the environment is selected from a finite-dimensional set $\boldsymbol{d}$. The participant judges DM quality according to a multivariate attribute $a \in \boldsymbol{a}$, which is a participant-specified image of the observed environment response to the applied decision. The attribute has $|\boldsymbol{i}| < \infty$, possibly vectorial, entries a_i. Thus, $a = (a_i)_{i=1}^{|\boldsymbol{i}|}$, $a_i \in \boldsymbol{a}_i$, $i \in \boldsymbol{i} = \{1, \ldots, |\boldsymbol{i}|\}$, and $\boldsymbol{a}$ is the Cartesian product $\boldsymbol{a} = \mathsf{X}_{i \in \boldsymbol{i}} \boldsymbol{a}_i$.

The solution of a sequence of static DM tasks consists of the choice and use of an *admissible randomised strategy*, which is formed by a sequence $(\mathsf{S}_t)_{t \in \boldsymbol{t}}$ of the randomised causal mappings

$$\mathsf{S}_t \in \mathbf{S}_t \subset \{\text{knowledge available at time } (t-1) \to d_t \in \boldsymbol{d}\},\ t \in \boldsymbol{t}. \tag{3.6}$$

We accept the following basic non-restrictive assumptions.

Agreement 1 (Knowledge Available). *The knowledge available at time* $(t-1)$ *(3.6),* $t \in \boldsymbol{t}$*, includes*

- *the* data observed *up to time* $(t-1)$ *inclusive, i.e. decisions made* $(d_1, \ldots, d_{t-1})$ *and the corresponding realisations of attributes* $(a_1, \ldots, a_{t-1})$*;*
- *a time-invariant* parametric environment model $\mathsf{M}(a|d, \Theta) > 0$*, which is a conditional pd known up to a finite-dimensional parameter* $\Theta \in \boldsymbol{\Theta}$*;*
- *a* prior pd $\mathsf{F}_0(\Theta) > 0$ *on the unknown parameter* $\Theta \in \boldsymbol{\Theta}$*.*

The standard Bayesian learning and prediction [47] require availability of the knowledge described in Agreement 1 in order to provide the predictive pds $(\mathsf{M}_t(a|d))_{t \in \boldsymbol{t}}$. They model the environment in the way needed for the design of the admissible strategy $(\mathsf{S}_t)_{t \in \boldsymbol{t}}$.

Agreement 2 (Optimality in the FPD Sense). *The following optimal strategy* $(\mathsf{S}_t^o)_{t \in \boldsymbol{t}}$ *in the FPD sense [31] is chosen*

$$(\mathsf{S}_t^o)_{t \in \boldsymbol{t}} \in \operatorname{Arg} \min_{(\mathsf{S}_t \in \mathbf{S}_t)_{t \in \boldsymbol{t}}} \frac{1}{|\boldsymbol{t}|} \sum_{t \in \boldsymbol{t}} \mathbb{E}[\mathbb{D}(\mathsf{M}_t \mathsf{S}_t \| \mathsf{I}_t)], \tag{3.7}$$

[13] The restriction to the static case allows us to avoid technical details making understanding of the conceptual solution difficult. All results are extendable to the dynamic DM with a mixed (discrete and continuous) observed data and considered but unobserved internal variables.

where the participant's DM preferences in tth DM task are quantified by an ideal pd $\mathsf{I}_t(a,d)$ *assigning a high probability to the desirable pairs* $(a,d) \in (\boldsymbol{a},\boldsymbol{d})$ *and a low probability to undesirable ones. The expectation* $\mathbb{E}[\bullet]$ *is taken over conditions of the individual summands in (3.7)*[14].

The strategy $(\mathsf{S}_t^o)_{t\in\boldsymbol{t}}$ minimises an average Kullback-Leibler divergence $\mathbb{D}(\mathsf{M}_t\mathsf{S}_t||\mathsf{I}_t)$ of the strategy-dependent closed-loop model $\mathsf{M}_t(a|d)\mathsf{S}_t(d)$ from the participant's DM preferences-expressing ideal pd $\mathsf{I}_t(a,d)$.

Assumption 1 (Preference Specification). *The participant provides the time-invariant sets* $\boldsymbol{\alpha}_i,\ i \in \boldsymbol{i}$, *of the most desirable values of individual attribute entries* a_i

$$\boldsymbol{\alpha}_i \subset \boldsymbol{a}_i,\ \boldsymbol{\alpha}_i \neq \emptyset,\ i \in \boldsymbol{i}. \tag{3.8}$$

These sets define the set of the most desirable attributes' values $\boldsymbol{\alpha}$

$$\boldsymbol{\alpha} = \mathsf{X}_{i\in\boldsymbol{i}}\boldsymbol{\alpha}_i \subsetneq \boldsymbol{a},\ \boldsymbol{\alpha} \neq \emptyset. \tag{3.9}$$

The participant can also assign *importance weights* $w \in \boldsymbol{w} = \{w = (w_1,\ldots,w_{|\boldsymbol{i}|}),\ w_i \geq 0,\ \sum_{i\in\boldsymbol{i}} w_i = 1\}$[15] to particular attribute entries but the availability of w is rarely realistic.

Generally, the participant may specify a number of not necessarily embedded sets $\boldsymbol{\alpha}^\mu,\ \mu \in \boldsymbol{\mu} = \{1,\ldots,|\boldsymbol{\mu}|\},\ |\boldsymbol{\mu}| > 1$, of the desirable attribute values with the desirability decreasing with μ. The participant may also specify similar information about the possible decisions. The chosen version of the partially specified DM preferences suffices for presenting an essence of the proposed approach.

Preferences are elicited under the following non-standard assumption.

Assumption 2 (Modelling of the Unknown Optimal Strategy). *A parametric model* $\mathsf{s}(d|\theta)$ *of an* unknown optimal randomised strategy *and a prior pd* $\mathsf{f}_0(\theta)$ *of an unknown finite-dimensional* $\theta \in \boldsymbol{\theta}$ *parameterising this model are available.*

The feasibility of Assumption 2 follows from the time-invariance of the parametric model of the environment and from the assumed invariance of the (partially specified) participant's DM preferences[16]. Neither the environment model nor the complete DM preferences are known and the parameter. The only source of knowledge is observed closed-loop data. Therefore the model of the optimal strategy can be learnt from it during application of non-optimal strategy. Having this non-standard learning problem solved, the standard Bayesian prediction [47] provides the model of the optimal strategy as the predictive pd $\mathsf{s}_t(d)$. The chain rule [47] for pds and the already learnt environment model $\mathsf{M}_t(a|d)$ imply the availability of the closed-loop model with the estimated optimal strategy $\mathsf{s}_t(d)$

[14] The considered KLD measures divergence between the conditional pds. The environment model $\mathsf{M}_t(a|d)$, the optimised mapping $\mathsf{S}_t(a)$ as well as the ideal pd $\mathsf{I}_t(a,d)$ depend on the random knowledge available at time $(t-1)$, see Agreement 1.

[15] The set is referred as probabilistic simplex.

[16] The proposed preference elicitation with time-invariant sets $\boldsymbol{\alpha}_i$ can be extended to time-varying cases.

$$\mathsf{P}_t(a,d) = \mathsf{M}_t(a|d)\mathsf{s}_t(d),\ t \in \boldsymbol{t}. \tag{3.10}$$

Problem Formulation. *Under Assumptions 1, 2 describing the available information about the environment and partially specified DM preferences*[17]*, design a well-justified automated construction of the ideal pds* $(\mathsf{I}_t)_{t\in\boldsymbol{t}}$ quantifying the given participant's DM preferences.

The ideal-pds construction is a novel part in the following formalised description of the closed loop depicted in Figure 3.1

$$\overbrace{\left\{\begin{array}{l}\boldsymbol{\alpha} \subsetneqq \boldsymbol{a}, \boldsymbol{d}, \overbrace{a_1,\dots,a_{t-1},d_1,\dots,d_{t-1}}^{\text{observed data}} \\ \boldsymbol{\Theta}, \mathsf{M}(a|d,\Theta), \mathsf{F}_{t-1}(\Theta), \boldsymbol{\theta}, \mathsf{s}(d|\theta), \mathsf{f}_{t-1}(\theta)\end{array}\right\}}^{\text{given DM elements}} \Rightarrow \left\{\begin{array}{l}\mathsf{M}_t(a|d) \\ \mathsf{s}_t(d) \\ \mathsf{I}_t(a,d)\end{array}\right\} \Rightarrow \mathsf{S}_t^o$$
$$\Rightarrow\ d_t \in \boldsymbol{d} \quad \overset{\text{environment}}{\Longrightarrow} \quad a_t \in \boldsymbol{a} \ \Rightarrow \left\{\begin{array}{l}\mathsf{F}_t(\Theta) \\ \mathsf{f}_t(\theta)\end{array}\right\},\ t \in \boldsymbol{t}. \tag{3.11}$$

3.4 Conceptual Solution of Preference Elicitation

The proposed preference elicitation and the optimised learning strategy form a unity described by the following conceptual algorithm. Section 3.5 contains discussion providing details on the options made.

Algorithm 1

1. *Delimit the time-invariant DM elements listed in Agreements 1, 2 and Assumption 2*:
   a. Specify the set $\boldsymbol{d}$ of available decisions and the set $\boldsymbol{a} = \mathsf{X}_{i\in\boldsymbol{i}}\boldsymbol{a}_i$ of the multivariate attributes $a = a = (a_i)_{i=1}^{|\boldsymbol{i}|}$.
   b. Select the parametric models of the environment $\mathsf{M}(a|d,\Theta)$ and of the optimal strategy $\mathsf{s}(d|\theta)$.
   c. Specify the prior pds $\mathsf{F}_0(\Theta)$, $\mathsf{f}_0(\theta)$ of the parameters $\Theta \in \boldsymbol{\Theta}$, $\theta \in \boldsymbol{\theta}$.

Further on, the algorithm runs for the increasing time $t \in \boldsymbol{t}$.

2. *Specify DM preferences on attributes* $(a_i)_{i\in\boldsymbol{i}}$ *via the sets* $\boldsymbol{\alpha}_i$*, (3.8), as required by Assumption 1. If possible, specify their relative importance by assigning them weights* w_i *in probabilistic simplex* $\boldsymbol{w}$ *or set* $w_i = 1/|\boldsymbol{i}|$*,* $i \in \boldsymbol{i}$*. Change the obsolete pd* $\mathsf{f}_{t-1}(\theta)$ *from the previous time if the participant's DM preferences have been changed at this time instance*.

   This step completes specification of DM elements coinciding with the collection of formal objects in the first curly brackets before the rightwards double arrow, see (3.11).

[17] The partial specification of the DM preferences via Assumptions 1, 2 is much easier than a direct specification of the DM-aims-expressing ideal pds.

3. *Evaluate predictive pds, [47],* $\mathsf{M}_t(a|d)$, $\mathsf{s}_t(d)$ *serving as the environment model and the optimal-strategy model.*

   The models $\mathsf{M}_t(a|d)$, $\mathsf{s}_t(d)$, serve for a design of S_t (3.6) generating the decision d_t. Thus, they can exploit data measured up to and including time $t-1$, cf. Agreement 1.
4. *Select a decision* $\bar{d} = \bar{d}(w)$ *(it depends on the weights w assigned)*

$$\bar{d} = \bar{d}(w) \in \mathrm{Arg}\max_{d\in\boldsymbol{d}} \sum_{i\in\boldsymbol{i}} w_i \int_{\boldsymbol{\alpha}_i} \mathsf{M}_t(a_i|d)\,\mathrm{d}a_i \tag{3.12}$$

   and define the set I_t *of the reachable ideal pds expressing the participant's DM preferences*

$$\mathsf{I}_t = \left\{\mathsf{I}_t(a,d):\ \mathsf{I}_t(a_i) = \mathsf{M}_t\left(a_i|\bar{d}\right),\ \forall a_i \in \boldsymbol{a}_i,\ i\in\boldsymbol{i}\right\}. \tag{3.13}$$

   The decision $\bar{d} \in \boldsymbol{d}$ provides the Pareto-optimised probabilities[18]

$$\left(\int_{\boldsymbol{\alpha}_1} \mathsf{M}_t(a_1|\bar{d})\,\mathrm{d}a_1, \ldots, \int_{\boldsymbol{\alpha}_{|\boldsymbol{i}|}} \mathsf{M}_t(a_{|\boldsymbol{i}|}|\bar{d})\,\mathrm{d}a_{|\boldsymbol{i}|}\right)$$

   of the desirable attribute-entries sets (3.8). The weight w with constant entries $w_i = 1/|\boldsymbol{i}|$ can be replaced by the weight w^o maximising the probability of the set $\boldsymbol{\alpha} = \mathsf{X}_{i\in\boldsymbol{i}}\boldsymbol{\alpha}_i$ of the most desirable attribute values

$$w^o \in \mathrm{Arg}\max_{w\in\boldsymbol{w}} \int_{\boldsymbol{\alpha}} \mathsf{M}_t(a|\bar{d}(w))\,\mathrm{d}a,$$

   see (3.12).
5. *Extend the partial specification* $\mathsf{I}_t \in \mathbf{I}_t$*, (3.13) to the pd* $\mathsf{I}_t^e(a,d)$ *via the following application of the minimum KLD principle*

$$\mathsf{I}_t^e \in \mathrm{Arg}\min_{\mathsf{I}_t\in\mathbf{I}_t} \mathbb{D}(\mathsf{I}_t\|\mathsf{P}_t) \quad \textit{with } \mathsf{P}_t(a,d) = \mathsf{M}_t(a|d)\mathsf{s}_t(d). \tag{3.14}$$

   The set $\mathbf{I}_t$ created in Step 4 reflects the participant's DM preferences. The extension to the ideal pd $\mathsf{I}_t^e(a,d)$ supposes that s_t is a good guess of the optimal strategy.

   This step finishes the specification of the mapping marked by the first rightwards double arrow in (3.11).
6. *Perform the FPD (3.7) with the environment model* $\mathsf{M}_t(a|d)$ *and the ideal pd* $\mathsf{I}_t^e(a,d)$*. Then generate* d_t *according to the mapping* S_t^o *optimal in the FPD sense (3.7), apply it, and observe* a_t*.*

   Enriching the knowledge available makes the solved DM task dynamic one even for the time-invariant parametric environment model. The dynamics is enhanced by the dependence of the used ideal pd on data and time. The unfeasible

[18] A vector function, dependent on a decision, becomes Pareto-optimised if an improvement of any of its entries leads to a deterioration of another one, [46].

optimal design arising for this complex dynamic DM task has to be solved approximately and the approximation should be revised at each time moment.

This step finishes the specification of the mappings symbolised by the second and marked by the third rightwards double arrow in (3.11).

7. *Update the pd* $\mathsf{F}_{t-1}(\Theta) \rightarrow^{(a_t,d_t)} \mathsf{F}_t(\Theta)$ *in the Bayesian way, i.e. enrich the knowledge about the parameter* Θ *of the environment model* $\mathsf{M}(a|d,\Theta)$ by a_t, d_t.
This step is inevitable even when making decisions without the preference elicitation. The updating may include forgetting [37] if the parameter Θ varies.
8. *Update the information about the parameter of the model of the optimal strategy, i.e. update* $\mathsf{f}_{t-1}(\theta) \rightarrow^{(a_t,d_t)} \mathsf{f}_t(\theta)$ *according to the following weighted version of the Bayes rule*[19]

$$\mathsf{f}_t(\theta) \propto \mathsf{s}^{\phi_t}(d|\theta)\mathsf{f}_{t-1}(\theta), \quad \phi_t = \int_{\boldsymbol{\alpha}} \mathsf{M}_{t+1}(a|d=d_t)\,\mathrm{d}a. \tag{3.15}$$

This data censoring is inevitable for learning the optimal strategy.

The step finishes the specification of the mapping expressed by the last rightwards double arrow in (3.11).

9. *Increase time t and go to Step 2 or to Step 3, if the DM preferences have not been changed.*

3.5 Individual Steps of the Conceptual Solution

This section provides details and discussion of the solution steps. The following sections correspond to the individual steps of the conceptual solution summarised in Section 3.4. The third digit of a section coincides with the corresponding number of the discussed step. Steps 2, 4, 5, 6 and 8 are the key ones, the remaining are presented for completeness.

The general solution is specialised to the important parametric models from the exponential family [2] used in the majority of practically feasible solutions.

3.5.1 Specification of Decision Making Elements

This step transforms a real-life DM task formulated in domain-oriented terms into the formal one.

ad 1a Specification of the sets of available decisions $\boldsymbol{d}$ and observable attributes $\boldsymbol{a}$
Often, these sets are uniquely determined by the domain-specific conditions of the solved DM task. Possible ambiguities can be resolved by Bayesian testing of hypothesis, e.g. [32], about informativeness of the prospective attributes and about influence of the available decisions.

[19] The environment model $\mathsf{M}_{t+1}(a|d=d_t)$ used in (3.15) exploits data measured up to and including time t and will also serve for the choice of d_{t+1}.

ad 1b Choice of the parametric models $\mathsf{M}(a|d,\Theta)$, $\Theta\in\boldsymbol{\Theta}$, $\mathsf{s}(d|\theta)$, $\theta\in\boldsymbol{\theta}$

A full modelling art, leading to grey- or black-box models, can be applied here. The modelling mostly provides deterministic but approximate models, which should be extended to the needed probabilistic models via the minimum KLD principle, Section 3.2.

Illustrative simulations, Section 3.6, use zero-order Markov chains that relate discrete-valued attributes and decisions. Markov chains belong to a dynamic *exponential family* (EF)[20], [2, 26],

$$\mathsf{M}(a_t|a_1,\ldots,a_{t-1},d_1,\ldots,d_t,\Theta)=\mathsf{M}(a_t|d_t,\Theta)=\exp\langle\mathsf{B}(a_t,d_t),\mathsf{C}(\Theta)\rangle\,, \tag{3.16}$$

determined by a scalar product $\langle\mathsf{B},\mathsf{C}\rangle$ of compatible values of multivariate functions $\mathsf{B}(a,d)$[21] and $\mathsf{C}(\Theta)$.

The following formula provides the considered most general parametrisation of Markov-chain models and its correspondence with the EF

$$\begin{aligned}\mathsf{M}(a_t|d_t,\Theta)&=\Theta(a_t|d_t)=\exp\left(\sum_{a\in\boldsymbol{a},d\in\boldsymbol{d}}\delta(a,a_t)\delta(d,d_t)\ln(\Theta(a|d))\right)\\&=\exp\langle\mathsf{B}(a_t,d_t),\mathsf{C}(\Theta)\rangle\,,\ \Theta\in\boldsymbol{\Theta},\\ \mathsf{s}(d_t|\theta)&=\exp\left(\sum_{d\in\boldsymbol{d}}\delta(d,d_t)\ln(\theta(d))\right)=\exp\langle\mathsf{b}(d_t),\mathsf{c}(\theta)\rangle\,,\ \theta\in\boldsymbol{\theta},\end{aligned} \tag{3.17}$$

where b,c have the same meaning as B,C in (3.16). $\boldsymbol{\Theta},\boldsymbol{\theta}$ are in appropriate probabilistic simplex sets and *Kronecker delta* $\delta(x,\tilde{x})$ equals to one for $x=\tilde{x}$ and it is zero otherwise.

ad 1c Specification of prior pds $\mathsf{F}_0(\Theta)$, $\mathsf{f}_0(\theta)$, $\Theta\in\boldsymbol{\Theta}$, $\theta\in\boldsymbol{\theta}$

The specification of the prior pds is known as knowledge elicitation. A relatively complete elaboration of the automated knowledge elicitation (focused on the dynamic EF) is in [24]. Comparing to a general case, the knowledge elicitation problem within the EF is much simpler as the EF admits a finite-dimensional sufficient statistic [34] and possesses a conjugate prior pd [4]. It has the following self-reproducing functional form

$$\mathsf{F}_0(\Theta)\propto\exp\langle V_0,\mathsf{C}(\Theta)\rangle\chi_{\boldsymbol{\Theta}}(\Theta),\ \ \mathsf{f}_0(\Theta)\propto\exp\langle v_0,\mathsf{c}(\theta)\rangle\chi_{\boldsymbol{\theta}}(\theta) \tag{3.18}$$

where an *indicator function* $\chi_{\boldsymbol{x}}(x)$ equals one for $x\in\boldsymbol{x}$ and is zero otherwise. In the EF, the knowledge elicitation reduces to a selection of finite-dimensional tables V_0,v_0 parameterising these prior pds.

[20] Models from the EF dominate in practice.

[21] The dependence only on d_t formalises the assumed static case. In a fully dynamic case, the function B acts on values of a finite-dimensional function of the observed data and of the current decision. Its dimension is fixed even for growing number of collected data and its values can be updated recursively.

For Markov chains, the conjugate priors are Dirichlet pds, cf. (3.17). Their functional forms (chosen for $t = 0$) are preserved for all $t \in \boldsymbol{t}$

$$\begin{aligned} \mathsf{F}_t(\Theta) &= \prod_{d\in\boldsymbol{d}} \frac{\prod_{a\in\boldsymbol{a}} \Theta(a|d)^{V_t(a|d)-1}}{\beta(V_t(\cdot|d))} \chi_{\boldsymbol{\theta}}(\Theta) = \prod_{d\in\boldsymbol{d}} \mathsf{Di}_{\Theta(\cdot|d)}(V_t(\cdot|d)) \\ V_t(\cdot|d) &= (V_t(a_1|d),\ldots,V_t(a_{|\boldsymbol{i}|}|d)),\ \ \Theta(\cdot|d) = (\Theta(a_1|d),\ldots,\Theta(a_{|\boldsymbol{i}|}|d)),\ a_i \in \boldsymbol{a}_i, \\ \mathsf{f}_t(\theta) &= \mathsf{Di}_\theta(v_t),\ V_0(a|d) > 0,\ v_0(d) > 0 \ \text{ on } \boldsymbol{a},\boldsymbol{d}. \end{aligned} \tag{3.19}$$

The used multivariate beta function

$$\beta(x) = \frac{\prod_{l\in\boldsymbol{l}} \Gamma(x_l)}{\Gamma(\sum_{l\in\boldsymbol{l}} x_l)}$$

is defined for a positive $|\boldsymbol{l}|$-vector x. $\Gamma(\cdot)$ is Euler gamma function [1]. V_0 can be interpreted as an *occurrence table*: $V_0(a|d)$ means the (thought) number of occurrences of the value a following the decision d. v_0 has a similar interpretation.

3.5.2 Specification of Preferences and Their Changes

The domain-specific description of the participant's DM preferences via the set $(\boldsymbol{\alpha}_i)_{i\in\boldsymbol{i}}$ of the most desirable attribute values (3.8) is a relatively straightforward task. The considered entry-wise specification respects limits of the human being, who can rarely go beyond pair-wise comparison. The DM preferences specified by (3.8) mean

$$\boldsymbol{\alpha}_i \text{ is the more desirable than } (\boldsymbol{a}_i \setminus \boldsymbol{\alpha}_i),\ \ i \in \boldsymbol{i}, \tag{3.20}$$

where $\boldsymbol{\alpha}_i \subsetneq \boldsymbol{a}_i$ is a set of desirable values of the ith attribute entries and $\boldsymbol{a}_i \setminus \boldsymbol{\alpha}_i$ is its complement to $\boldsymbol{a}_i$. The specification of the DM preferences (3.20) is mostly straightforward. However, the participant can change them at some time t, for instance, by changing the selection of the attribute entries non-trivially constrained by the set $\boldsymbol{\alpha} = \mathsf{X}_{i\in\boldsymbol{i}}\boldsymbol{\alpha}_i$. Let us discuss how to cope with a DM preference change from $\boldsymbol{\alpha}$ to $\tilde{\boldsymbol{\alpha}} \neq \boldsymbol{\alpha}$.

The discussed change modifies the set of candidates of the optimal strategy and makes the learnt pd $\mathsf{f}_{t-1}(\theta)$ inadequate. It is possible to construct a new pd $\tilde{\mathsf{f}}_{t-1}(\theta)$ from scratch if the observed data is stored. It requires a sufficient supply of deliberation resources for performing a completely new censoring of the observed data, see Section 3.5.8, respecting the new DM preferences given by the set $\tilde{\boldsymbol{\alpha}} \subsetneq \boldsymbol{a}$.

The situation is more complex if the pd $\mathsf{f}_{t-1}(\theta)$, $\theta \in \boldsymbol{\theta}$, reflecting the obsolete DM preferences is solely stored. Then, the prior pd $\mathsf{f}_0(\theta)$ is the only safe guess of the parameter $\theta \in \boldsymbol{\theta}$, which should describe a new optimal strategy. We hope that the divergence of the correct pd $\tilde{\mathsf{f}}_{t-1}(\theta)$ (describing the strategy optimal with respect to $\tilde{\boldsymbol{\alpha}}$) on $\mathsf{f}_{t-1}(\theta)$ is bounded. This motivates the choice of the pd $\tilde{\mathsf{f}}_{t-1}(\theta)$ via the following version of the minimum KLD principle

$$\tilde{\mathsf{f}}_{t-1}(\cdot) = \mathsf{f}_{t-1}(\cdot|\phi) \in \operatorname{Arg}\min_{\tilde{\mathsf{f}}\in\tilde{\mathbf{f}}_{t-1}} \mathbb{D}(\tilde{\mathsf{f}}||\mathsf{f}_0),\ \ \tilde{\mathbf{f}}_{t-1} = \left\{\text{pds } \tilde{\mathsf{f}}(\cdot) \text{ on } \boldsymbol{\theta},\ \mathbb{D}(\tilde{\mathsf{f}}||\mathsf{f}_{t-1}) \leq k\right\} \tag{3.21}$$

for some $k \geq 0$. The solution of (3.21), recalled in Section 3.2, provides the following rule for tracking of DM preference changes.

The change of the most desirable set $\boldsymbol{\alpha}$ to $\tilde{\boldsymbol{\alpha}} \neq \boldsymbol{\alpha}$ is respected by the change [22]

$$\mathsf{f}_{t-1}(\theta) \;\; \rightarrow \;\; \mathsf{f}_{t-1}(\theta|\phi) \propto \mathsf{f}_{t-1}^{\phi}(\theta)\mathsf{f}_{0}^{1-\phi}(\theta). \tag{3.22}$$

The adequate forgetting factor $\phi \in [0,1]$ is unknown due to the lack of the knowledge of k, which depends on the sets $\boldsymbol{\alpha}$ and $\tilde{\boldsymbol{\alpha}}$ in too complex way. Each specific choice of the forgetting factor ϕ provides a model

$$\mathsf{s}_t(d|\phi) = \int_{\boldsymbol{\theta}} \mathsf{s}(d|\theta)\mathsf{f}_{t-1}(\theta|\phi)\,\mathrm{d}\theta \tag{3.23}$$

of the optimal strategy. Each model (3.23) determines the probability of a new set $\tilde{\boldsymbol{\alpha}}$ of the most desirable attribute values, which should be maximised by the optimal strategy. This leads to the choice of the best forgetting factor ϕ^o as the maximiser of this probability

$$\phi^o \in \mathrm{Arg}\max_{\phi\in[0,1]} \int_{\tilde{\boldsymbol{\alpha}}} \mathsf{M}_t(a|d)\mathsf{s}_t(d|\phi)\,\mathrm{d}d. \tag{3.24}$$

Qualitatively this solution is plausible due to a smooth dependence of $\mathsf{f}_{t-1}(\theta|\phi)$ (3.22) on the forgetting factor ϕ. The pd $\mathsf{f}_{t-1}(\theta|\phi)$ has also desirable limit versions, which for: i) $\phi \approx 0$ describe that the optimal strategies corresponding to $\boldsymbol{\alpha}$ and $\tilde{\boldsymbol{\alpha}}$ are unrelated; ii) $\phi \approx 1$ express that the DM preference change of $\boldsymbol{\alpha}$ to $\tilde{\boldsymbol{\alpha}}$ has a negligible influence on the optimal strategy.

Quantitative experience with this tracking of the DM preference changes is still limited but no conceptual problems are expected. Unlike other options within the overall solution, the choice (3.24) is a bit of ad-hoc nature.

3.5.3 Evaluation of Environment and Optimal Strategy Models

The admissible strategies can at most use the knowledge available, Agreement 1. They cannot use correct values of parameters Θ, θ, i.e. they have to work with predictive pds serving as the environment model $\mathsf{M}_t(a|d)$ and the model of the optimal strategy $\mathsf{s}_t(d)$ [23]. If the DM preferences have changed, $\mathsf{s}_t(d)$ should be replaced by the pd $\mathsf{s}_t(d|\phi)$ reflecting this change (3.23) with the best forgetting factor $\phi = \phi^o$ (3.24)

[22] For pds $\mathsf{f}_{t-1}(\theta)$, $\mathsf{f}_0(\theta)$ conjugated to an EF member, the pd $\mathsf{f}_{t-1}(\theta|\phi)$ (3.22) is also conjugated to it.

[23] Let us recall that all DM tasks work with the same time-invariant parametric models of the environment and the optimal strategy $\mathsf{M}(a|d,\Theta)$ and $\mathsf{s}(d|\theta)$. The predictive pds $\mathsf{M}_t(a|d)$, $\mathsf{s}_t(d)$, serving the tth DM task, exploit the knowledge accumulated before time t quantified by $\mathsf{F}_{t-1}(\Theta)$ and $\mathsf{f}_{t-1}(\theta)$.

$$\mathsf{M}_t(a|d) = \int_{\boldsymbol{\Theta}} \mathsf{M}(a|d,\Theta)\mathsf{F}_{t-1}(\Theta)\,\mathrm{d}\Theta > 0 \ \text{ on } \ (\boldsymbol{a},\boldsymbol{d}) \tag{3.25}$$

$$\mathsf{s}_t(d) = \int_{\boldsymbol{\theta}} \mathsf{s}(d|\theta)\mathsf{f}_{t-1}(\theta)\,\mathrm{d}\theta \ \text{ or } \ \mathsf{s}_t(d|\phi) = \frac{\int_{\boldsymbol{\theta}} \mathsf{s}(d|\theta)\mathsf{f}_{t-1}^{\phi}(\theta)\mathsf{f}_0^{1-\phi}(\theta)\,\mathrm{d}\theta}{\int_{\boldsymbol{\theta}} \mathsf{f}_{t-1}^{\phi}(\theta)\mathsf{f}_0^{1-\phi}(\theta)\,\mathrm{d}\theta}.$$

The formulae (3.25) can be made more specific for the exponential family (3.16) and for the corresponding conjugate pds (3.18). The self-reproducing property of the pd

$$\mathsf{F}_{t-1}(\Theta) = \mathsf{F}(\Theta|V_{t-1}) \propto \exp\langle V_{t-1}, \mathsf{C}(\Theta)\rangle \tag{3.26}$$

conjugated to a parametric environment model $\mathsf{M}(a|d,\Theta) = \exp\langle \mathsf{B}(a,d), \mathsf{C}(\Theta)\rangle$ and the parametric model of the optimal strategy $\mathsf{s}(d|\theta) = \exp\langle \mathsf{b}(d), \mathsf{c}(\theta)\rangle$ imply

$$\begin{aligned} \mathsf{M}_t(a|d) &= \frac{\mathsf{J}(V_{t-1} + \mathsf{B}(a,d))}{\mathsf{J}(V_{t-1})}, \ \ \mathsf{J}(V) = \int_{\boldsymbol{\Theta}} \exp\langle V, \mathsf{C}(\Theta)\rangle \mathsf{F}(\Theta|V)\,\mathrm{d}\Theta \\ \mathsf{s}_t(d) &= \frac{\mathsf{j}(v_{t-1} + \mathsf{b}(d))}{\mathsf{j}(v_{t-1})}, \qquad \mathsf{j}(v) = \int_{\boldsymbol{\theta}} \exp\langle v, \mathsf{c}(\theta)\rangle \mathsf{f}(\theta|v)\,\mathrm{d}\theta. \end{aligned} \tag{3.27}$$

The stabilised forgetting (3.5), suitable also for tracking of the varying parameter of the environment model, replaces the sufficient statistics V_{t-1}, v_{t-1} by the convex combinations

$$\tilde{V}_{t-1} = \phi V_{t-1} + (1-\phi)V_0, \ \ \tilde{v}_{t-1} = \phi v_{t-1} + (1-\phi)v_0.$$

For Markov chains and the parametrisation (3.17), the conjugate Dirichlet pds (3.19)

$$\mathsf{F}_{t-1}(\Theta) = \mathsf{F}(\Theta|V_{t-1}) = \prod_{d\in\boldsymbol{d}} \mathsf{Di}_{\Theta(\cdot|d)}(V_{t-1}(\cdot|d))$$

and

$$\mathsf{f}_{t-1}(\theta) = \mathsf{f}(\theta|v_{t-1}) = \mathsf{Di}_{\theta}(v_{t-1})$$

reproduce. They depend on the occurrence tables V_{t-1}, v_{t-1}, which sufficiently compress the observed (censored) data. The corresponding explicit forms of predictive pds serving as the environment model and the model of the optimal strategy are

$$\mathsf{M}_t(a|d) = \frac{V_{t-1}(a|d)}{\sum_{\tilde{a}\in\boldsymbol{a}} V_{t-1}(\tilde{a}|d)}, \ a\in\boldsymbol{a}, \ \ \mathsf{s}_t(d) = \frac{v_{t-1}(d)}{\sum_{\tilde{d}\in\boldsymbol{d}} v_{t-1}(\tilde{d})}, \ d\in\boldsymbol{d}. \tag{3.28}$$

Up to the important influence of initial values of occurrence tables, these formulae coincide with the relative frequency of occurrences of the realised configurations of a specific attribute a after making a specific decision d and the relative frequency of occurrences of the decision value d. The formulas (3.28) follow from the known property $\Gamma(x+1) = x\Gamma(x)$ [1], which also implies that the environment model coincides with conditional expectations $\hat{\Theta}_{t-1}(a|d) = \mathbb{E}[\Theta(a|d)|V_{t-1}]$ of $\Theta(a|d)$

$$\mathsf{M}_t(a|d) = \hat{\Theta}_{t-1}(a|d) = \frac{V_{t-1}(a|d)}{\sum_{\tilde{a}\in\boldsymbol{a}} V_{t-1}(\tilde{a}|d)},\ a\in\boldsymbol{a}, d\in\boldsymbol{d}. \tag{3.29}$$

It suffices to store the point estimates $\hat{\Theta}_{t-1}(a|d)$ of the unknown parameter $\Theta(a|d)$ and the *imprecision vector* κ_{t-1} with entries

$$\kappa_{t-1}(d) = \frac{1}{\sum_{a\in\boldsymbol{a}} V_{t-1}(a|d)},\ d\in\boldsymbol{d}, \tag{3.30}$$

instead of V_{t-1} as the following transformations are bijective

$$V \leftrightarrow (\hat{\Theta},\kappa) \text{ and } v \leftrightarrow (\hat{\theta},\tau) = \left(\frac{v_d}{\sum_{d\in\boldsymbol{d}} v_d}, \frac{1}{\sum_{d\in\boldsymbol{d}} v_d}\right). \tag{3.31}$$

These transformations of sufficient statistics suit for the approximate design of the optimal strategy, see Section 3.5.6.

3.5.4 Expressing of Preferences by Ideal Probability Densities

Within the set of pds on $(\boldsymbol{a},\boldsymbol{d})$, we need to select an ideal pd, which respects the participant's partially specified DM preferences. We use the following way.

Find pds assigning the highest probability to the most desirable attribute values according to the environment model when using a proper decision $\bar{d}\in\boldsymbol{d}$. Then choose the required singleton among them via the minimum KLD principle.

This verbal description delimits the set of ideal-pd candidates both with respect to their functional form and numerical values but does not determine them uniquely. Here, we discuss considered variants of their determination and reasons, which led us to the final choice presented at the end of this section.

Throughout several subsequent sections the time index t is fixed and suppressed as uninformative.

Independent Choice of Marginal Probability Densities of the Ideal Probability Density

Let us consider the ith attribute entry a_i and find the decision $\bar{d}^i\in\boldsymbol{d}$

$$\bar{d}^i \in \operatorname{Arg}\max_{d\in\boldsymbol{d}} \int_{\boldsymbol{\alpha}_i} \mathsf{M}(a_i|d)\,\mathrm{d}a_i,\ i\in\boldsymbol{i}.$$

The maximised probability of the set $\boldsymbol{\alpha}_i$ of the most desirable values of the ith attribute is given by the ith marginal pd $\mathsf{M}(a_i|d)$ of the joint pd $\mathsf{M}(a|d)$. The decision $\bar{d}^i$ guarantees the highest probability of having the ith attribute entry in the set $\boldsymbol{\alpha}_i$ at the price that a full decision effort concentrates on it. This motivates to consider a set of ideal pds I respecting the participant's DM preferences (3.8) as the pds $\mathsf{I}(a,d)$ having the marginal pds

$$\mathsf{I}(a_i) = \mathsf{M}\left(a_i|\bar{d}^i\right),\ \ i\in\boldsymbol{i}. \tag{3.32}$$

Originally, we have focused on this option, which respects entry-wise specification of the DM preferences (3.8). It also allows a simple change of the number of attribute entries that actively delimit the set of the most desirable attribute values. Moreover, this specification of I uses the marginal pds $\mathsf{M}(a_i|d)$ of the pd $\mathsf{M}(a|d)$, which are more reliable than the full environment model $\mathsf{M}(a|d)$. Indeed, the learning of mutual dependencies of respective attribute entries is data-intensive as the number of "dependencies" of $|\boldsymbol{i}|$-dimensional discrete-valued vector a grows very quickly.

A closer look, however, reveals that such ideal pds are unrealistic as generally $\bar{d}^i \neq \bar{d}^j$ for $i \neq j$. Then, the ideal pd $\mathsf{I}(a)$ with the marginal pds (3.32) cannot be attained. This feature made us to abandon this option.

Joint Choice of the Marginal Probability Densities of the Ideal PD

The weakness of the independent choice has guided us to consider the following option. A single decision $\bar{d} \in \boldsymbol{d}$ is found, which maximises the probability of reaching the set of the most desirable attribute values $\boldsymbol{\alpha}$

$$\bar{d} \in \operatorname{Arg}\max_{d \in \boldsymbol{d}} \int_{\boldsymbol{\alpha}} \mathsf{M}(a|d)\,\mathrm{d}a. \tag{3.33}$$

It serves for constraining the set of ideal pds I to those having the marginal pds

$$\mathsf{I}(a_i) = \mathsf{M}\left(a_i|\bar{d}\right) = \int_{\boldsymbol{a}_{\setminus i}} \mathsf{M}\left(a|\bar{d}\right)\mathrm{d}a_{\setminus i}, \tag{3.34}$$

where subscript $_{\setminus i}$ indexes a vector created from a by omitting the ith entry

$$a_{\setminus i} = (a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_{|\boldsymbol{i}|}).$$

This variant eliminates drawback of the independent choice at the price of using joint pd in (3.33). Otherwise, it seemingly meets all requirements on the DM preferences-expressing ideal pds. It may, however, be rather bad with respect to the entry-wise specified DM preferences (3.8). Indeed, the marginalisation (3.34) may provide an ideal pd with marginal probabilities of the sets $\boldsymbol{\alpha}_i$ (i.e. of the most desirable attribute values) are strictly smaller than their complements

$$\int_{\boldsymbol{\alpha}_i} \mathsf{I}(a_i)\,\mathrm{d}a_i < \int_{\boldsymbol{a}_i \setminus \boldsymbol{\alpha}_i} \mathsf{I}(a_i)\,\mathrm{d}a_i,\ \forall i \in \boldsymbol{i}. \tag{3.35}$$

This contradicts the participant's wish (3.20). The following example shows that this danger is real one.

Example 1

Let us consider a two-dimensional attribute $a = (a_1, a_2)$ and a scalar decision $d = d_1$ with the sets of possible values $\boldsymbol{a}_1 = \boldsymbol{a}_2 = \boldsymbol{d} = \{1, 2\}$. The environment model $\mathsf{M}(a|d)$ is a table, explicitly given in Table 3.1 together with the marginal pds $\mathsf{M}_1(a_1|d), \mathsf{M}_2(a_2|d)$.

Table 3.1 The left table describes the discussed environment model $\mathsf{M}(a|d)$, the right one provides its marginal pds $\mathsf{M}(a_i|d)$. The parameters in this table have to meet constraints $\sigma_i,\rho_i,\zeta_i > 0$, $1-\sigma_i-\rho_i-\zeta_i > 0$, $i \in \boldsymbol{i} = \{1,2\}$ guaranteeing that $\mathsf{M}(a|d)$ is a conditional pd.

	a=(1,1)	a=(1,2)	a=(2,1)	a=(2,2)
d=1	σ_1	ρ_1	ζ_1	$1-\sigma_1-\rho_1-\zeta_1$
d=2	σ_2	ρ_2	ζ_2	$1-\sigma_2-\rho_2-\zeta_2$

	a_1=1	a_1=2	a_2=1	a_2=2
d=1	$\sigma_1+\rho_1$	$1-\sigma_1-\rho_1$	$\sigma_1+\zeta_1$	$1-\sigma_1-\zeta_1$
d=2	$\sigma_2+\rho_2$	$1-\sigma_2-\rho_2$	$\sigma_2+\zeta_2$	$1-\sigma_2-\zeta_2$

For an entry-wise-specified set of the most desirable attribute values $\boldsymbol{\alpha}_1 = \{1\}$, $\boldsymbol{\alpha}_2 = \{1\}$, we get the set $\boldsymbol{\alpha}$ of the most desirable attribute values as the singleton $\boldsymbol{\alpha} = \{(1,1)\}$. Considering $\sigma_1 > \sigma_2$ in Table 3.1, the decision $\bar{d} = 1$ maximises $\int_{\boldsymbol{\alpha}} \mathsf{M}(a|d)\,\mathrm{d}a = \mathsf{M}(1,1)|d)$. There is an infinite amount of the parameter values σ_i,ρ_i,ζ_i, $i \in \boldsymbol{i}$, for which $\mathsf{M}\left(a|\bar{d}\right)$ has the marginal pds (3.34) with the adverse property (3.35). A possible choice of this type is in Table 3.2.

Table 3.2 The left table contains specific numerical values of the discussed environment model $\mathsf{M}(a|d)$, the right one provides its marginal pds $\mathsf{M}(a_i|d)$, $i \in \boldsymbol{i} = \{1,2\}$

	a=(1,1)	a=(1,2)	a=(2,1)	a=(2,2)
d=1	0.40	0.05	0.05	0.50
d=2	0.30	0.30	0.30	0.10

	a_1=1	a_1=2	a_2=1	a_2=2
d=1	0.45	0.55	0.45	0.55
d=2	0.60	0.40	0.60	0.40

Table 3.2 shows that the decision $\bar{d} = 1$, maximising $\int_{\boldsymbol{\alpha}} \mathsf{M}(a|d)\,\mathrm{d}a = \mathsf{M}((1,1)|d)$, gives the marginal pds $\mathsf{M}\left(a_i = 1|\bar{d} = 1\right) = 0.45 < 0.55 = \mathsf{M}\left(a_i = 2|\bar{d} = 1\right)$. The other decision $d^o = 2$ leads to $\mathsf{M}(a_i = 1|d^o = 2) = 0.6 > 0.4 = \mathsf{M}(a_i = 2|d^o = 2)$, both for $i = 1,2$. This property disqualifies the choice (3.33), (3.34).

Pareto-Optimal Marginal Probability Densities of the Ideal PD

The adverse property (3.35) means that the solution discussed in the previous section can be dominated in Pareto sense [46]: the marginal pds $\mathsf{I}(a_i)$ (3.34) may lead to probabilities $\int_{\boldsymbol{\alpha}_i} \mathsf{I}(a_i)\,\mathrm{d}a_i$, $i \in \boldsymbol{i}$, which are smaller than those achievable by other decision $d^o \in \boldsymbol{d}$ used in (3.34) instead of $\bar{d}$. This makes us to search directly for a non-dominated, Pareto optimal, solution reachable by a $\bar{d} \in \boldsymbol{d}$.

Taking an $|\boldsymbol{i}|$-dimensional vector w of arbitrary positive probabilistic weights $w \in \boldsymbol{w} = \{w = (w_1,\ldots,w_{|\boldsymbol{i}|}),\ w_i > 0,\ \sum_{i\in\boldsymbol{i}} w_i = 1\}$ and defining the w-dependent decision

$$\bar{d} \in \mathrm{Arg}\max_{d\in\boldsymbol{d}} \sum_{i\in\boldsymbol{i}} w_i \int_{\boldsymbol{\alpha}_i} \mathsf{M}(a_i|d)\,\mathrm{d}a_i \ \text{ and } \mathsf{I}(a_i) = \mathsf{M}\left(a_i|\bar{d}\right),\ i \in \boldsymbol{i}, \tag{3.36}$$

ensures the found solution be non-dominated.

Indeed, let us assume that there is another $d^o \in \boldsymbol{d}$ such that $\int_{\boldsymbol{\alpha}_i} \mathsf{M}(a_i|d^o)\,\mathrm{d}a_i \geq \int_{\boldsymbol{\alpha}_i} \mathsf{M}(a_i|\bar{d})\,\mathrm{d}a_i$, $\forall i \in \boldsymbol{i}$, with some inequality being strict. Multiplying these inequalities by the positive weights w_i and summing them over $i \in \boldsymbol{i}$ we get the sharp inequality contradicting the definition of $\bar{d}$ as the maximiser in (3.36).

The possible weights w: i) are either determined by the participant if it is able to distinguish the importance of individual attribute entries; ii) or are fixed to the constant $w_i = 1/|\boldsymbol{i}|$ if the participant is indifferent with respect to individual attribute entries; iii) or can be chosen as maximiser of the probability of a set of the most desirable attribute values $\boldsymbol{\alpha} = \mathsf{X}_{i\in\boldsymbol{i}}\boldsymbol{\alpha}_i$ by selecting

$$w^o \in \mathrm{Arg}\max_{w\in\boldsymbol{w}} \mathsf{M}\left(\mathsf{X}_{i\in\boldsymbol{i}}\boldsymbol{\alpha}_i|\bar{d}\right), \text{ with } w\text{-dependent } \bar{d} \text{ given by (3.36).} \tag{3.37}$$

The primary option i) is rarely realistic and the option iii) is to be better than ii).

In summary, the most adequate set of ideal pds respecting the participant's DM preferences (3.8) in a reachable way reads

$$\mathsf{I} = \left\{\mathsf{I}(a,d) > 0 \text{ on } (\boldsymbol{a},\boldsymbol{d}) : \mathsf{I}(a_i) = \mathsf{M}\left(a_i|\bar{d}\right), \text{ with } \bar{d} \text{ given by (3.36)}\right\}. \tag{3.38}$$

Note that in the example of the previous section the optimisation (3.37) is unnecessary. General case has not been analysed yet but no problems are foreseen.

A natural question arises: Why the decision $\bar{d}$ is not directly used as the optimal one? The negative answer follows primarily from heavily exploitation-oriented nature of $\bar{d}$. It does not care about exploration, which is vital for a gradual improvement of the used strategy, which depends on the improvement of the environment and optimal strategy models. In the considered static case, the known fact that the repetitive use of $\bar{d}$ may completely destroy learning and consequently decision making [39] manifests extremely strongly. The example of such an adverse behaviour presented in Section 3.6 appeared without any special effort.

Moreover in dynamic DM, the use of this myopic[24] strategy leads to an inefficient behaviour, which even may cause instability of the closed decision loop [30].

3.5.5 Extension of Marginal Probability Densities to the Ideal PD

The set I (3.38) of the ideal pds is determined by linear constraints explicitly expressing the participant's wishes. This set is non-empty as

$$\mathsf{I}(a,d) = \tilde{\mathsf{P}}(d|a)\prod_{i\in\boldsymbol{i}} \mathsf{I}(a_i) \in \mathsf{I}, \text{ with } \mathsf{I}(a_i) = \mathsf{M}\left(a_i|\bar{d}\right),$$

[24] The myopic strategy is mostly looking one-stage-ahead.

where $\tilde{\mathsf{P}}(d|a)$ is an arbitrary pd positive on $\boldsymbol{d}$ for all conditions in $\boldsymbol{a}$. The arbitrariness reflects the fact known from the copula theory [44] that marginal pds do not determine uniquely the joint pd having them. Thus, I contains more members and an additional condition has to be adopted to make the final well-justified choice. The selection should not introduce extra, participant-independent, DM preferences. The minimum KLD principle, Section 3.2, has this property if it selects the ideal pd $\mathsf{I}(a,d)$ from (3.38) as the closest one to the pd $\mathsf{P}(a,d)$ expressing the available knowledge about the closed decision loop with the optimal strategy.

The minimised KLD $\mathbb{D}(\mathsf{I}\|\mathsf{P})$ is a strictly convex functional of the pd I from the convex set I (3.38) due to the assumed positivity of the involved pds $\mathsf{M}(a|d)$ and $\mathsf{P}(a,d)$. Thus, see [18], the constructed I is a unique minimiser of the Lagrangian functional determined by the multipliers $-\ln(\Lambda(a_i))$, $i \in \boldsymbol{i}$,

$$\begin{aligned}
\mathsf{I}^e(a,d) &= \arg\min_{\mathsf{I}\in\mathsf{I}} \int_{(\boldsymbol{a},\boldsymbol{d})} \left[\mathsf{I}(a,d) \ln\left(\frac{\mathsf{I}(a,d)}{\mathsf{P}(a,d)}\right) - \sum_{i\in\boldsymbol{i}} \ln(\Lambda(a_i))\mathsf{I}(a,d) \right] \mathrm{d}a\,\mathrm{d}d \\
&= \arg\min_{\mathsf{I}\in\mathsf{I}} \int_{\boldsymbol{a}} \mathsf{I}(a) \left[\ln\left(\frac{\mathsf{I}(a)}{\mathsf{P}(a)\prod_{i\in\boldsymbol{i}}\Lambda(a_i)}\right) + \int_{\boldsymbol{d}} \mathsf{I}(d|a) \ln\left(\frac{\mathsf{I}(d|a)}{\mathsf{P}(d|a)}\right) \mathrm{d}d \right] \mathrm{d}a \\
&= \frac{\mathsf{P}(d|a)\mathsf{P}(a)\prod_{i\in\boldsymbol{i}}\Lambda(a_i)}{\mathsf{J}(\mathsf{P},\Lambda)}.
\end{aligned} \tag{3.39}$$

The second equality in (3.39) is implied by Fubini theorem [48] and by the chain rule for pds [47]. The result in the third equality follows from the facts that: i) the second term (after the second equality) is a conditional version of the KLD, which reaches its smallest zero value for $\mathsf{I}^e(d|a) = \mathsf{P}(d|a)$ and ii) the first term is the KLD minus logarithm of an I-independent normalising constant $\mathsf{J}(\mathsf{P},\Lambda)$.

The Lagrangian multipliers solve, for $a_{\setminus i} = (a_1,\ldots,a_{i-1},a_{i+1},\ldots,a_{|\boldsymbol{i}|})$, $\Lambda_{\setminus i} = (\Lambda_1,\ldots,\Lambda_{i-1},\Lambda_{i+1},\ldots,\Lambda_{|\boldsymbol{i}|})$, $i \in \boldsymbol{i}$,

$$\begin{aligned}
\mathsf{M}\left(a_i|\bar{d}\right) &= \frac{\int_{\boldsymbol{a}_{\setminus i}} \mathsf{P}(a)\prod_{j\in\boldsymbol{i}}\Lambda_j(a_j)\,\mathrm{d}a_{\setminus i}}{\mathsf{J}(\mathsf{P},\Lambda)} \\
&= \frac{\mathsf{P}_i(a_i)\Lambda_i(a_i)}{\mathsf{J}(\mathsf{P},\Lambda)} \int_{\boldsymbol{a}_{\setminus i}} \mathsf{P}(a_{\setminus i}|a_i) \prod_{j\in\boldsymbol{i}\setminus\{i\}} \Lambda_j(a_j)\,\mathrm{d}a_{\setminus i} = \frac{\mathsf{P}_i(a_i)\Lambda_i(a_i)}{\mathsf{J}(\mathsf{P},\Lambda)}\Phi(a_i,\Lambda_{\setminus i}).
\end{aligned} \tag{3.40}$$

By construction, equations (3.40) have a unique solution, which can be found by the successive approximations

$$\frac{{}^k\Lambda_i(a_i)}{\mathsf{J}(\mathsf{P},{}^k\Lambda)} = \frac{\mathsf{M}\left(a_i|\bar{d}\right)}{\mathsf{P}_i(a_i)\Phi(a_i,{}^{k-1}\Lambda_{\setminus i})}, \tag{3.41}$$

where an evaluation of the kth approximation ${}^k\Lambda = \left[{}^k\Lambda_1,\ldots,{}^k\Lambda_{|\boldsymbol{i}|}\right]$, $k = 1,2,\ldots$ starts from an initial positive guess ${}^0\Lambda$ and stops after the conjectured stabilisation. The factor $\mathsf{J}(\mathsf{P},{}^k\Lambda)$ is uniquely determined by the normalisation of the kth approximation of the constructed ideal pd (3.39).

The successive approximations, described by (3.41), do not face problems with division by zero for the assumed $\mathsf{M}(a|d), \mathsf{P}(a,d) > 0 \Rightarrow \mathsf{M}\left(a_i|\bar{d}\right), \mathsf{P}_i(a_i) > 0,\ i \in \boldsymbol{i}$. Numerical experiments strongly support the adopted conjecture about their convergence. If the conjecture is valid then the limit provides the searched solution due to its uniqueness.

3.5.6 Exploring Fully Probabilistic Design

Here, the dynamic exploration aspects of the problem are inspected so that time-dependence is again explicitly expressed. As before, the discussion is preferred against definite results. A practically oriented reader can skip the rest of this section after reading the first section, where the used certainty-equivalent strategy is described.

The repetitive solutions of the same type static DM problems form a single dynamic problem due to the common, recursively-learnt, environment and optimal strategy models. The KLD (scaled by $|\boldsymbol{t}|$) to be minimised over $(\mathsf{S}_t)_{t\in\boldsymbol{t}}$ reads

$$\frac{1}{|\boldsymbol{t}|}\mathbb{D}\left(\prod_{t\in\boldsymbol{t}}\mathsf{M}_t\mathsf{S}_t\middle\|\prod_{t\in\boldsymbol{t}}\mathsf{I}_t\right)=\frac{1}{|\boldsymbol{t}|}\sum_{t\in\boldsymbol{t}}\mathbb{E}\left[\mathbb{D}(\mathsf{M}_t\mathsf{S}_t\|\mathsf{I}_t)\right].$$

The expectation $\mathbb{E}[\bullet]$ is taken over conditions occurring in the individual summands and

$$\mathbb{D}(\mathsf{M}_t\mathsf{S}_t\|\mathsf{I}_t)=\int_{(\boldsymbol{a},\boldsymbol{d})}\mathsf{M}_t(a|d)\mathsf{S}_t(d)\ln\left(\frac{\mathsf{M}_t(a|d)\mathsf{S}_t(d)}{\mathsf{I}_t(a,d)}\right)\mathrm{d}a\,\mathrm{d}d. \tag{3.42}$$

Recall, $\mathsf{M}_t(a|d)=\int_{\boldsymbol{\Theta}}\mathsf{M}(a|d,\Theta)\mathsf{F}_{t-1}(\Theta)\,\mathrm{d}\Theta$ and $\mathsf{I}_t(a,d)$ is an image of $\mathsf{M}_t(a|d)$ and $\mathsf{s}_t(d|\phi^o)=\int_{\boldsymbol{\theta}}\mathsf{s}(d|\theta)\mathsf{f}_{t-1}(\theta|\phi^o)\,\mathrm{d}\theta$, see Sections 3.5.3, 3.5.5.

The formal solution of this FPD [53] has the following form evaluated backward for $t=|\boldsymbol{t}|,|\boldsymbol{t}|-1,\ldots,1$,

$$\begin{aligned}
\mathsf{S}^o_t(d)&=\frac{\mathsf{I}_t(d)\exp[-\omega_t(d)]}{\int_{\boldsymbol{d}}\mathsf{I}_t(d)\exp[-\omega_t(d)]\,\mathrm{d}d}=\frac{\mathsf{I}_t(d)\exp[-\omega_t(d)]}{\gamma(a_1,\ldots,a_{t-1},d_1,\ldots,d_{t-1})}\\
\omega_t(d)&=\int_{\boldsymbol{a}}\mathsf{M}_t(a|d)\ln\left(\frac{\mathsf{M}_t(a|d)}{\gamma(a_1,\ldots,a_{t-1},a,d_1,\ldots,d_{t-1},d)\mathsf{I}_t(a|d)}\right)\mathrm{d}a\\
&\text{starting from } \gamma(a_1,\ldots,a_{|\boldsymbol{t}|},d_1,\ldots,d_{|\boldsymbol{t}|})=1.
\end{aligned} \tag{3.43}$$

The optimal design (3.43) provides a dual strategy [12], which optimally balances exploration and exploitation activities. In the considered static case, the observed data enters the conditions via the learnt models of the environment and the optimal strategy (3.25) determining the used ideal pd $\mathsf{I}_t=\mathsf{I}^e_t$. The data realisation is influenced by the applied strategy, which can be optimal only if the overall effect of exploitation and exploration is properly balanced. This design is mostly infeasible and some approximate-dynamic-programming technique [52] has to be used.

Certainty Equivalence Decision Strategy

The simplest approximate design, labelled as *certainty equivalence*, replaces the unknown parameter by its current point estimate and assumes that the estimate is uninfluenced by the chosen decisions.

This strategy is asymptotically reasonable as the posterior pds $\mathsf{F}_t(\Theta), \mathsf{f}_t(\theta)$ form martingales [42] and under the general conditions (met always for Markov chains) they almost surely converge to singular pds [3, 26]. Consequently, the only source of dependence between successive static DM tasks diminishes as the dependence of γ in (3.43) on data disappears. The certainty-equivalent strategy that neglects this dependence breaks the single task (3.42) into the sequence of DM tasks consisting of independently solved one-stage-ahead looking static problems with the solutions of the form (cf. (3.43))

$$\mathsf{S}_t^o(d) \propto \mathsf{I}_t(d)\exp[-\omega_t(d)],\ \omega_t(d) = \int_{\boldsymbol{a}} \mathsf{M}_t(a|d)\ln\left(\frac{\mathsf{M}_t(a|d)}{\mathsf{I}_t(a|d)}\right)\mathrm{d}a,\ t \in \boldsymbol{t}.$$

The transient learning period is, however, critical as – without an exploration – the posterior pds may concentrate on wrong sets, whenever the conditioning data is not informative enough. Here, one of the advantages of the FPD enters the game. The FPD provides a *randomised* optimal strategy and the sampling of decisions from it adds a well-scaled dither (exploring) noise, which diminishes with a proper rate.

Strictly speaking, these statements are conjectures supported by experiments whose samples are in Section 3.6. To get a guaranteed version of sufficiently exploring, almost optimal, strategy, a more sophisticated approximation (3.43) is needed. It requires tailoring of techniques known from approximate dynamic programming [52] to the FPD. It is possible and relatively simple as the strategy optimal in the FPD sense is described explicitly. A widely applicable construction framework is outlined below.

Failing Cautious Decision Strategy

During the whole development, we have been aware (and experiments confirmed it) that exploration is vital for a good overall closed-loop behaviour. Primarily, this made us avoid a direct use of $\bar{d}$ even in the static DM.

For continuous-valued (a, d), there is an experimental evidence that certainty-equivalent strategy is often exploratory enough even in the standard Bayesian DM. On contrary, in Markov-chain case, there is a proof that deterministic certainty-equivalent strategy may fail with a positive probability [39]. This made us to let one stage-ahead design know that the unknown parameters Θ, θ are in the game. It did not helped as we got cautious strategy [19], which is even less exploring than the certainty equivalent one and worked improperly. The formal mechanism of this failure becomes obvious when noticing that the imprecisions of parameter estimates κ, τ (3.30), (3.31) enter the design only for the design horizon $|\boldsymbol{t}| > 1$.

Approximate Dynamic Programming in FPD

This section touches of a rich area of approximate dynamic programming [52]. It indicates that within the FPD it is possible to obtain algorithmically feasible approximation of the optimally exploring strategy. The approach is just outlined and thus it suffices to focus on the Markov-chain case to this purpose. The corresponding function $\gamma(\cdot)$, determining the function $\omega(\cdot)$ and the optimal strategy, see (3.43), depends on the sufficient statistics. They consist of the point estimates $\hat{\Theta}_{t-1}, \hat{\theta}_{t-1}$ of the parameters Θ, θ and of the corresponding imprecisions κ_{t-1}, τ_{t-1}, see (3.29), (3.30) and (3.31). Note that the use of this form of sufficient statistics is important as the statistics are expected to converge to finite constants (ideally, $\hat{\Theta}_t, \hat{\theta}_t$ to the correct values of the unknown parameters Θ, θ and imprecisions κ_t, τ_t to zeros) unlike the statistics V_t, v_t.

For this open-ended presentation, it suffices to consider the explicit dependence of $\gamma(\cdot)$ and $\omega(\cdot)$ on $\hat{\Theta}$ and κ only. The first step of the optimal design (3.43) for $t = |\boldsymbol{t}|$ with $\gamma(\cdot) = 1$ gets the form

$$\mathsf{S}_t^o(d) = \mathsf{S}^o(d|V_{t-1}) = \mathsf{S}^o(d|\hat{\Theta}_{t-1}, \kappa_{t-1}) = \frac{\mathsf{I}_t(d)\exp[-\omega_t(d, \hat{\Theta}_{t-1}, \kappa_{t-1})]}{\gamma_t(\hat{\Theta}_{t-1}, \kappa_{t-1})}$$

$$\gamma_t(\hat{\Theta}_{t-1}, \kappa_{t-1}) = \sum_{d \in \boldsymbol{d}} \mathsf{I}_t(d)\exp[-\omega_t(d, \hat{\Theta}_{t-1}, \kappa_{t-1})]$$

$$\omega_t(d, \hat{\Theta}_{t-1}, \kappa_{t-1}) = \sum_{a \in \boldsymbol{a}} \hat{\Theta}_{t-1}(a|d) \ln\left(\frac{\hat{\Theta}_{t-1}(a|d)}{\mathsf{I}_t(a|d)}\right).$$

Further design steps can be interpreted as value iterations searching for the optimal stationary strategy. Even for the FPD, they converge for large $|\boldsymbol{t}|$ under the rather general conditions [26]. We care about this stationary phase, drop the time subscript at $\gamma_t(\cdot), \omega_t(\cdot)$, and write down a general step of the design in terms of $\gamma(\cdot)$. It holds (cf. Section 3.5.7)

$$\begin{aligned}
\gamma(\hat{\Theta}_{t-1}, \kappa_{t-1}) = \sum_{d \in \boldsymbol{d}} \Bigg\{ & \mathsf{I}_t(d)\exp\left[-\sum_{\boldsymbol{a}} \hat{\Theta}_{t-1}(a|d)\ln\left(\frac{\hat{\Theta}_{t-1}(a|d)}{\mathsf{I}_t(a|d)}\right)\right] \\
& \times \exp\left[\sum_{a \in \boldsymbol{a}} \hat{\Theta}_{t-1}(a|d)\ln(\gamma(\hat{\Theta}_{t-1} + \Delta_{t-1}(\cdot,\cdot,a,d), \kappa_{t-1}\Omega_{t-1}(\cdot,d)))\right]\Bigg\} \\
\Delta_{t-1}(\tilde{a}, \tilde{d}, a, d) = & \frac{\delta(\tilde{d}, d)\kappa_{t-1}(d)}{1 + \delta(\tilde{d}, d)\kappa_{t-1}(d)}(\delta(\tilde{a}, a) - \hat{\Theta}_{t-1}(a|d)), \;\; \tilde{a}, a \in \boldsymbol{a}, \; \tilde{d}, d \in \boldsymbol{d} \\
\Omega_{t-1}(\tilde{d}, d) = & \frac{1}{1 + \kappa_{t-1}(\tilde{d})\delta(\tilde{d}, d)}, \;\; \tilde{d}, d \in \boldsymbol{d}.
\end{aligned} \tag{3.44}$$

Let us insert into the right-hand side of $\gamma(\cdot)$ in (3.44) an exponential-type approximation of $\gamma(\cdot) > 0$

$$\gamma(\hat{\Theta}, \kappa) \approx \exp\left[\mathrm{tr}(G\hat{\Theta}) + g\kappa\right]. \tag{3.45}$$

The approximation is parameterised by a matrix G (of the size of the transposed $\hat{\Theta}_t$) and by a vector g (of the transposed size of κ_t). It gives a mixture of exponential functions of a similar type on the left-hand side of (3.44). The recovering of the feasible exponential form (3.45) requires an approximation of this mixture by a single exponential function. The key question is what proximity measure should be used. To decide it, it suffices to observe that the optimal strategy $\mathsf{S}^o(d|\hat{\Theta}_{t-1},\kappa_{t-1})$ is the pd of d conditioned on $\hat{\Theta}_{t-1},\kappa_{t-1}$. Thus, $\gamma(\hat{\Theta}_{t-1},\kappa_{t-1})$ can be interpreted as a (non-normalised) marginal pd of $\hat{\Theta}_{t-1},\kappa_{t-1}$, which should be approximated by a feasible pd having the form (3.45). This singles out the approximation in terms of the KLD (3.2). In the considered case, it reduces to a fitting of the moments of $\hat{\Theta}_{t-1},\kappa_{t-1}$ of the left-hands-side mixture by moments of a pd having the exponential form (3.45).

The outlined idea is applicable generally. Algorithmic details for the EF will be elaborated in an independent paper.

3.5.7 Learning of the Environment Model

The environment model serves to the DM task for which the DM preferences are elicited. Thus, it has to be constructed anyway. Within the considered context of repetitive DMs, its standard Bayesian learning is available. It works with an environment model $\mathsf{M}(a|d,\Theta)$ parameterised by a finite-dimensional unknown parameter $\Theta\in\boldsymbol{\Theta}$. The following version of the Bayes rule [4] evolves the pd $\mathsf{F}_{t-1}(\Theta)$ comprising all information about the unknown parameter Θ

$$\mathsf{F}_t(\Theta)=\frac{\mathsf{M}(a_t|d_t,\Theta)\mathsf{F}_{t-1}(\Theta)}{\int_{\boldsymbol{\Theta}}\mathsf{M}(a_t|d_t,\Theta)\mathsf{F}_{t-1}(\Theta)\,\mathrm{d}\Theta}, \tag{3.46}$$

where the pair (a_t,d_t) is realised in the tth DM task. This pd provides the predictive pd (3.25) used as the environment model $\mathsf{M}_{t+1}(a|d)$. It is worth stressing that the applied strategy cancels in the directly applied Bayes rule. Within the considered context, it "naturally" fulfils natural conditions of control (decision making) [47] requiring $\mathsf{S}_t(d|\Theta)=\mathsf{S}_t(d)$ and expressing that the parameter Θ is unknown to the used strategy.

For a parametric model from the EF (3.16) and a conjugate self-reproducing pd $\mathsf{F}_t(\Theta)\propto\exp\langle V_t,\mathsf{C}(\Theta)\rangle$, the functional form of the Bayes rule (3.46) reduces to the updating of the sufficient statistic, cf. (3.27),

$$V_t=V_{t-1}+\mathsf{B}(a_t,d_t),\ V_0 \text{ determines the prior pd and has to make } \mathsf{J}(V_0)<\infty.$$

For the controlled Markov chain used in simulations, V_t is an occurrence table with entries $V_t(a|d)$, $a\in\boldsymbol{a}$, $d\in\boldsymbol{d}$. The specific form of the function $\mathsf{B}(a,d)$ (3.17) provides the updating in the form

$$V_t(a|d)=V_{t-1}(a|d)+\delta(a,a_t)\delta(d,d_t),\ \ V_0(a|d)>0\ \Leftrightarrow\ \mathsf{J}(V_0)<\infty, \tag{3.47}$$

for the realised pair (a_t, d_t) and $\delta(a, \tilde{a})$ denoting Kronecker delta. This recursion transforms to the recursion for the point estimates $\hat{\Theta}_t$ (3.29) and the imprecisions κ_t (3.30)

$$\begin{aligned}\hat{\Theta}_t(a|d) &= \hat{\Theta}_{t-1}(a|d) + \delta(d, d_t)\kappa_t(d)(\delta(a, a_t) - \hat{\Theta}_{t-1}(a|d))\\ \kappa_t(d) &= \frac{\kappa_{t-1}(d)}{1 + \kappa_{t-1}(d)\delta(d, d_t)},\end{aligned}$$

which is used in Section 3.5.6 discussing an approximate dynamic programming.

This learning is conceptually very simple but it is strongly limited by the curse of dimensionality as the involved occurrence tables are mostly too large. Except very short vectors of attributes and decisions with a few possible values, their storing and updating require extremely large memory and, even worse, an extreme number of the observed data. Learning a mixture of low-dimensional approximate environment models relating scalar entries of attributes to scalar entries of the decision, [26, 43], seems to be a systematic viable way around.

Note that if parameter vary either because of physical reasons or due to approximation errors, the pd $\mathsf{F}_t(\Theta)$ differs from a correct pd and the situation discussed in connection with changing DM preferences, Section 3.5.2, recurs. Thus, the parameter changes can be respected when complementing the Bayes rule (3.46) by the stabilised forgetting

$$\begin{aligned}&\mathsf{F}_t(\Theta) \to \mathsf{F}_t^{\phi}(\Theta)(\mathsf{F}_0(\Theta))^{1-\phi_0} \text{ in general case}\\ &V_t \to \phi V_t + (1-\phi)V_0 \text{ for members of the EF.}\end{aligned}$$

In this context, the forgetting factor $\phi \in [0, 1]$ can be learnt in the usual Bayesian way at least on a discrete grid in $[0, 1]$, see e.g. [36, 40].

3.5.8 Learning of the Optimal Strategy

The construction of the ideal pd $\mathsf{I}_t = \mathsf{I}_t^e$ strongly depends on availability of the model $\mathsf{P}_t(a, d)$ of the closed decision loop with the optimal strategy $\mathsf{P}_t(a, d) = \mathsf{M}_t(a|d)\mathsf{s}_t(d)$, see Section 3.5.5. The Bayes rule is used for learning the environment model, Sections 3.5.7, 3.5.3. This rule could be used for learning the optimal strategy if all past decisions were generated by it. This cannot be expected within the important transient learning period. Thus, we have to decide whether a generated decision comes from the (almost) optimal strategy or not: we have to use a censored data.

If the realised attribute falls in the set of the most desirable attribute values $\boldsymbol{\alpha}$ then we have a strong indicator that the used decision is optimal. When relying only on it, we get an unique form of the learning with the strict data censoring

$$\mathsf{f}_t(\theta) = \frac{\mathsf{s}(d_t|\theta)^{\chi_{\boldsymbol{\alpha}}(a_t)}\mathsf{f}_{t-1}(\theta)}{\int_{\boldsymbol{\theta}} \mathsf{s}(d_t|\theta)^{\chi_{\boldsymbol{\alpha}}(a_t)}\mathsf{f}_{t-1}(\theta)\,\mathrm{d}\theta}. \tag{3.48}$$

However, the event $a_t \in \boldsymbol{\alpha}$ may be rare or a random consequence of a bad decision within a particular realisation. Thus, an indicator working with "almost optimality" is needed. It has to allow a learning even for $a_t \notin \boldsymbol{\alpha} \Leftrightarrow \chi_{\boldsymbol{\alpha}}(a_t) = 0$. For its design, it suffices to recognise that no censoring can be errorless. Thus, the pd $\mathsf{f}_{t-1}(\theta)$ is an approximate learning result: even if $a_t \in \boldsymbol{\alpha}$, we are uncertain whether the updated pd, labelled $\tilde{\mathsf{f}}_t(\theta) \propto \mathsf{s}(d_t|\theta)\mathsf{f}_{t-1}(\theta)$ coincides with a correct pd $\mathsf{f}_t(\theta)$. In other words, $\tilde{\mathsf{f}}_t$ only approximates the correct pd f_t. Again, as shown in [5], [28], the KLD is the proper Bayesian expression of their divergence. Thus, $\mathbb{D}(\mathsf{f}_t||\tilde{\mathsf{f}}_t) \leq k_t$ for some $k_t \geq 0$. At the same time, the pd $\mathsf{f}_{t-1}(\theta)$ is the best available guess before processing the realisation (a_t, d_t). The extension of this knowledge is to be done by the minimum KLD principle, Section 3.2, which provides

$$\mathsf{f}_t(\theta) = \frac{\mathsf{s}^{\phi_t}(d_t|\theta)\mathsf{f}_{t-1}(\theta)}{\int_{\boldsymbol{\theta}} \mathsf{s}^{\phi_t}(d_t|\theta)\mathsf{f}_{t-1}(\theta)\,\mathrm{d}\theta}, \ \phi \in [0,1]. \tag{3.49}$$

The formula (3.49) resembles (3.48) with the forgetting factor $\phi_t \in [0,1]$ replacing the value of the indicator function $\chi_{\boldsymbol{\alpha}}(a_t) \in \{0,1\}$. This resemblance helps us to select the forgetting factor, which is unknown due to the unspecified K_t, as a prediction of the indicator-function value

$$\phi_t = \int_{\boldsymbol{\alpha}} \mathsf{M}_{t+1}(a|d_t)\,\mathrm{d}a = \int_{\boldsymbol{\alpha}} \int_{\boldsymbol{\Theta}} \mathsf{M}(a|d_t,\Theta)\mathsf{F}_t(\Theta)\,\mathrm{d}\Theta\,\mathrm{d}a. \tag{3.50}$$

The use of d_t in the condition complies with checking of the (approximate) optimality of this decision sought before updating. Its use in the condition differentiates the formula (3.50) from (3.24), which cares about an "average" degree of optimality of the past decisions. The current experience with the choice (3.50) is positive but still the solution is of ad-hoc type.

3.6 Illustrative Simulations

All solution steps were tested on simulation experiments, which among others allowed cutting off clear cul-de-sacs of the developed solution. Here, we present a simple illustrative example, which can be confronted with intuitive expectations.

3.6.1 Simulation Set Up

The presentation follows the respective steps of Algorithm 1, see Section 3.4.

1. DM elements
   a. One-dimensional decisions $d \in \boldsymbol{d} = \{1,2\}$ and two-dimension observable attributes $(a_1, a_2) \in \mathsf{X}_{i\in\boldsymbol{i}}\boldsymbol{a}_i$, with $\boldsymbol{a}_1 = \{1,2\}$, $\boldsymbol{a}_2 = \{1,2,3\}$ coincide with the sets simulated by the environment, see Table 3.3.
   b. Zero-order Markov chains with the general parameterisations (3.17) are used.

c. The conjugate Dirichlet pds (3.19) are used as priors with $V_0(a|d) = 0.1$ on $(\boldsymbol{a},\boldsymbol{d})$, $v_0 = [4,1]$. The latter choice intentionally prefers a bad strategy. This choice checks learning abilities of the proposed preference elicitation.

2. The sets of the most desirable individual entries are $\boldsymbol{\alpha}_1 = \{1\}$, $\boldsymbol{\alpha}_2 = \{1\}$ giving $\boldsymbol{\alpha} = \{(1,1)\}$. No change of the DM preferences is assumed.
   Further on, the algorithm runs for increasing time $t \in \boldsymbol{t} = \{1,\ldots,100\}$.
3. The predictive pds serving as the environment model and the model of the optimal strategy are evaluated according to formulae (3.28).
4. The decision $\bar{d}$ (3.12) is evaluated for the uniform weights $w_i = 1/|\boldsymbol{i}| = 1/2$ reflecting the indifference with respect to the attribute entries.
5. The marginal ideal pds $\mathsf{I}_t(a_i) = \mathsf{M}_t\left(a_i|\bar{d}\right)$ are extended to the ideal pd $\mathsf{I}_t(a,d) = \mathsf{I}_t^e(a,d)$ as described in Section 3.5.5.
6. The FPD is performed in its certainty-equivalent version, see Section 3.5.6. The decision d_t is sampled from the tested strategy and it is fed into the simulated environment described by the transition probabilities in Table 3.3

Table 3.3 The simulated environment with probabilities of the configurations of attributes a responding to decision d being in respective cells. Under the complete knowledge of these probabilities, the optimal strategy selects $d = 2$.

	a=(1,1)	a=(1,2)	a=(1,3)	a=(2,1)	a=(2,2)	a=(2,3)
d=1	0.20	0.30	0.10	0.10	0.10	0.20
d=2	0.35	0.05	0.05	0.15	0.15	0.25

A fixed seed of random-numbers generator is used in all simulation runs, which makes the results comparable.

7. Bayesian learning of the environment model is performed according to the Bayes rule (3.47) without forgetting.
8. Learning of the optimal strategy runs exactly as proposed in Section 3.5.8.

3.6.2 Simulation Results

The numerical results of experiments show outcomes for: i) the *naive strategy* that exploits directly $\bar{d}_t$ (3.12) as the applied decision; ii) the *optimal strategy* that permanently applies the optimal decision ($d_t = 2$); iii) the *proposed strategy* that samples decisions from the certainty-equivalent result of the FPD.

Table 3.4 summarises relative estimation error of Θ parameterising the environment model

$$(1 - \hat{\Theta}_t(a|d)/\Theta(a|d)) \times 100,\ a \in \boldsymbol{a},\ d \in \boldsymbol{d}, \tag{3.51}$$

where $\hat{\Theta}_t(a|d)$ are point estimates (3.29) of simulated transition probabilities $\Theta(a|d)$ listed in Table 3.3. The occurrences of attribute and decision values for the respective tested strategies are in Table 3.5.

Table 3.4 Relative estimation errors [%] (3.51) after simulating $|\boldsymbol{t}| = 100$ samples for the respective tested strategies. The error concerning the most desirable attribute value is in the first column.

Results for the naive strategy

$d = 1$	0.9	-69.8	11.9	5.7	5.7	10.4
$d = 2$	52.4	-11.1	-233.3	-11.1	-233.3	33.3

Results for the optimal strategy

$d = 1$	16.7	-66.7	44.4	-66.7	-66.7	16.7
$d = 2$	-21.3	5.7	5.7	24.5	-32.1	17.0

Results for the proposed strategy

$d = 1$	23.9	-52.2	13.0	-30.4	-8.7	2.2
$d = 2$	-34.2	9.1	-21.2	49.5	-21.2	21.2

Table 3.5 Occurrences of attribute and decision values among $|\boldsymbol{t}| = 100$ samples for the respective tested strategies. The attribute entries $a_1 = 1$ and $a_2 = 1$ are the most desirable ones.

Results for the naive strategy

a_1	value 1: 55 times	value 2: 45 times	
a_2	value 1: 37 times	value 2: 36 times	value 3: 27 times
d	value 1: 100 times	value 2: 0 times	

Results for the optimal strategy

a_1	value 1: 55 times	value 2: 45 times	
a_2	value 1: 58 times	value 2: 15 times	value 3: 27 times
d	value 1: 0 times	value 2: 100 times	

Results for the proposed strategy

a_1	value 1: 57 times	value 2: 43 times	
a_2	value 1: 50 times	value 2: 22 times	value 3: 28 times
d	value 1: 40 times	value 2: 60 times	

Numerical outcomes of experiments are complemented by figures characterising the run of the proposed strategy. The left-hand side of Figure 3.2 shows the simulated attributes and the right-hand side shows the used decisions. The left-hand side of Figure 3.3 shows the probability (3.50) of the set $\boldsymbol{\alpha} = \{(1,1)\}$ of the most desirable attribute values used in the data censoring. The right-hand side displays tuning of the strategy, which gradually swaps from a bad strategy to the optimal one.

A representative fragment of the performed experiments is reported here. It suffices to add that longer runs (1000 samples and more) exhibited the expected convergence of the parameter estimates and of the proposed strategy.

3.6.3 Discussion of the Simulation Results

Estimation results reflected in Table 3.4 show that values of estimation errors are much less significant than their distribution over the parameter entries. They can fairly be evaluated only with respect to the achieved decision quality.

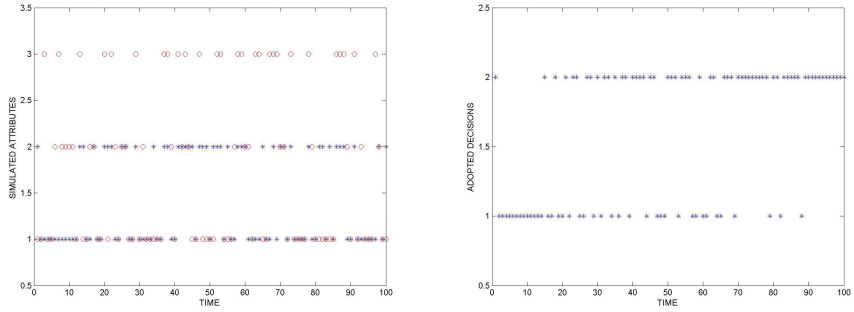


Fig. 3.2 The left figure shows simulated attributes (a_1 stars, a_2 circles). The right figure shows decisions generated by the proposed strategy.

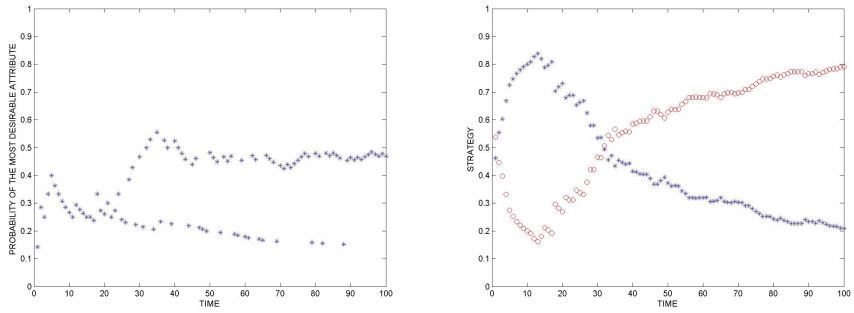


Fig. 3.3 The left figure shows probability $\int_{\boldsymbol{\alpha}} \mathsf{M}_{t+1}(a|d_t)\,\mathrm{d}a$ (3.50) used in the data censoring, Section 3.5.8. The right figure shows the proposed strategy, i.e. the pd given by the values $\mathsf{S}_t(d=1)$ (stars) and $\mathsf{S}_t(d=2) = 1 - \mathsf{S}_t(d=1)$ (circles).

Table 3.5 confirms that the naive strategy can be very bad indeed. It is also visible that the specific simulated environment, described in Table 3.3, makes only the values of the attribute a_2 sensitive to the decision taken: the simulated example is non-trivial in spite of its apparent simplicity. It is also worth noticing that the proposed strategy generates a relatively large portion of exploring decisions. Figure 3.2 shows that with increasing time the non-optimal decision ($d = 1$) is used less and less often than the optimal one ($d = 2$), as desirable. The similar desirable time dependence is visible in Figure 3.3: smaller probabilities (3.50) of the most desirable attribute pair, used for the data censoring (3.49), occur less and less often. The forgetting value stabilises still well below unity, which conforms with a high uncertainty connected with responses of the simulated environment. The same figure shows that after a relatively short learning period the proposed strategy starts to converge to the optimal one (as mentioned long runs confirmed).

3.7 Concluding Remarks

The chapter proposes a methodology of an automated elicitation of DM preference when the set of the most desirable attribute values is specified. This specification is quite common and can be extended to a preferential hierarchy with tunable weights of the DM preferences.

Many important features of the proposed solution are implied by the fact that the constructed ideal pd respects the relation between attributes and decisions as described by the environment model $\mathsf{M}_t(a|d)$. Specifically,

- The proposed ideal pd is *not* fully concentrated on the most desirable attribute value, which reflects that it cannot be reached with certainty.
- The functional form of the ideal pd is determined by the model $\mathsf{M}_t(a|d)$: it is not created in an ad-hoc, model independent, manner unlike utilities [33].
- It is always possible to project the constructed ideal pd into a class of feasible pds by using information criterion justified in [5], [28], Section 3.2, if the constructed ideal pd is too complex for numerical treatment or analysis.
- The environment model $\mathsf{M}_t(a|d)$ as well as the closed-loop model with the optimal strategy $\mathsf{P}_t(a,d) = \mathsf{M}_t(a|d)\mathsf{s}_t(d)$ are sequentially learnt; consequently, the DM preference description given by the ideal pd $\mathsf{I}_t(a,d)$ derived from them, is learned, too.
- The approach can be directly extended to a dynamic DM with a regression-type dependence.
- The involved pds can quantify the joint distributions of discrete-valued and continuous-valued attributes. This simplifies the elicitation of the DM preferences given by categorical as well as numerical attributes.

Suppression or a more firm justification of ad-hoc steps (e.g. choice of forgetting factors in tracking of the changed DM preferences or in data censoring) is the key methodological research direction. At a more practical level, covering other DM preference specifications is the important problem to be addressed. The proposed solution is clearly connected with the DM preference learning presented in [27]. There, an explicit dependence between the environment-model parameter and the parameter of the optimal strategy has been established in a special, practically significant, way. A more detailed and general inspection of this relation is another open problem. The design of specific algorithmic solutions for commonly used environment models is a further topic to be covered.

In spite of the width of the problems hidden behind these open research directions, the proposed methodology appears practically promising.

Acknowledgements. This research and chapter have been strongly influenced by Dr. Tatiana V. Guy who I take as co-author of innovative and interesting parts of this text. I do acknowledge her help very much indeed.

The support of the project GAČR 13-13502S is acknowledged. The support of the project GAČR 102/08/0567 is also acknowledged.

References

1. Abramowitz, M., Stegun, I.: Handbook of Mathematical Functions. Dover Publications, New York (1972)
2. Barndorff-Nielsen, O.: Information and exponential families in statistical theory, New York (1978)
3. Berec, L., Kárný, M.: Identification of reality in Bayesian context. In: Warwick, K., Kárný, M. (eds.) Computer-Intensive Methods in Control and Signal Processing: Curse of Dimensionality, Birkhäuser, Boston, pp. 181–193 (1997)
4. Berger, J.: Statistical Decision Theory and Bayesian Analysis. Springer, New York (1985)
5. Bernardo, J.M.: Expected information as expected utility. The Annals of Statistics 7(3), 686–690 (1979)
6. Bohlin, T.: Interactive System Identification: Prospects and Pitfalls. Springer, New York (1991)
7. Boutilier, C.: A POMDP formulation of preference elicitation problems. In: Biundo, S. (ed.) AAAI-2002 Proc. of the Fifth European Conference on Planning, pp. 239–246. AAAI Press, Durham (2002)
8. Boutilier, C., Brafman, R., Geib, C., Poole, D.: A constraint-based approach to preference elicitation and decision making. In: Proceedings of AAAI Spring Symposium on Qualitative Decision Theory, Stanford, CA, pp. 19–28 (1997)
9. Bowong, S., Dimi, J., Kamgang, J., Mbang, J., Tewa, J.: Survey of recent results of multi-compartments intra-host models of malaria and HIV. Revue ARIMA 9, 85–107 (2008)
10. Chajewska, U., Koller, D.: Utilities as random variables: Density estimation and structure discovery. In: Proceedings of UAI 2000, pp. 63–71 (2000)
11. Cooke, N.: Varieties of knowledge elicitation techniques. International Journal of Human-Computer Studies 41, 801–849 (1994)
12. Feldbaum, A.: Theory of dual control. Autom. Remote Control 21(9) (1960)
13. Gajos, K., Weld, D.: Preference elicitation for interface optimization. In: Proceedings of UIST 2005 (2005)
14. Garthwaite, P., Kadane, J., O´Hagan, A.: Statistical methods for eliciting probability distributions. J. of the American Statistical Association 100(470), 680–700 (2005)
15. Guy, T.V., Kárný, M., Wolpert, D.H. (eds.): Decision Making with Imperfect Decision Makers. ISRL, vol. 28. Springer, Heidelberg (2012)
16. Viappiani, H.P., Boutilier, S.Z., Learning, C.: complex concepts using crowdsourcing: A Bayesian approach. In: Proceedings of the Second Conference on Algorithmic Decision Theory (ADT 2011), Piscataway, NJ (2011)
17. Haykin, S.: Neural Networks: A Comprehensive Foundation. Macmillan, New York (1994)
18. Horst, R., Tuy, H.: Global Optimization, p. 727. Springer (1996)
19. Jacobs, O., Patchell, J.: Caution and probing in stochastic control. Int. J. of Control 16, 189–199 (1972)
20. Jaynes, E.: Information theory and statistical mechanics. Physical Review Series II 106(4), 620–630 (1957)
21. Jimison, H., Fagan, L., Shachter, R., Shortliffe, E.: Patient-specific explanation in models of chronic disease. AI in Medicine 4, 191–205 (1992)
22. Jorgensen, S.B., Hangos, K.M.: Qualitative models as unifying modelling tool for grey box modelling. Int. J. of Adaptive Control and Signal Processing 9(6), 461–562 (1995)
23. Kárný, M.: Towards fully probabilistic control design. Automatica 32(12), 1719–1722 (1996)

24. Kárný, M., Andrýsek, J., Bodini, A., Guy, T., Kracík, J., Nedoma, P., Ruggeri, F.: Fully probabilistic knowledge expression and incorporation. Tech. Rep. 8-10MI, CNR IMATI, Milan (2008)
25. Kárný, M., Andrýsek, J., Bodini, A., Guy, T.V., Kracík, J., Ruggeri, F.: How to exploit external model of data for parameter estimation? Int. J. of Adaptive Control and Signal Processing 20(1), 41–50 (2006)
26. Kárný, M., Böhm, J., Guy, T.V., Jirsa, L., Nagy, I., Nedoma, P., Tesař, L.: Optimized Bayesian Dynamic Advising: Theory and Algorithms. Springer, London (2006)
27. Kárný, M., Guy, T.: Preference elicitation in fully probabilistic design of decision strategies. In: Proc. of the 49th IEEE Conference on Decision and Control. IEEE (2010)
28. Kárný, M., Guy, T.V.: On Support of Imperfect Bayesian Participants. In: Guy, T.V., Kárný, M., Wolpert, D.H. (eds.) Decision Making with Imperfect Decision Makers. ISRL, vol. 28, pp. 29–56. Springer, Heidelberg (2012)
29. Kárný, M., Guy, T.V.: Fully probabilistic control design. Systems & Control Letters 55(4), 259–265 (2006)
30. Kárný, M., Halousková, A., Böhm, J., Kulhavý, R., Nedoma, P.: Design of linear quadratic adaptive control: Theory and algorithms for practice. Kybernetika 21(Supplement to Nos. 3, 4 ,5, 6) (1985)
31. Kárný, M., Kroupa, T.: Axiomatisation of fully probabilistic design. Information Sciences 186(1), 105–113 (2012)
32. Kárný, M., Kulhavý, R.: Structure determination of regression-type models for adaptive prediction and control. In: Spall, J. (ed.) Bayesian Analysis of Time Series and Dynamic Models. ch.12, Marcel Dekker, New York (1988)
33. Keeney, R., Raiffa, H.: Decisions with Multiple Objectives: Preferences and Value Tradeoffs. JohnWiley and Sons Inc. (1976)
34. Koopman, R.: On distributions admitting a sufficient statistic. Tran. of American Mathematical Society 39, 399 (1936)
35. Kuhn, H., Tucker, A.: Nonlinear programming. In: Proc. of the 2nd Berkeley Symposium, pp. 481–492. University of California Press, Berkeley (1951)
36. Kulhavý, R.: Can we preserve the structure of recursive Bayesian estimation in a limited-dimensional implementation? In: Helmke, U., Mennicken, R., Saurer, J. (eds.) Systems and Networks: Mathematical Theory and Applications, vol. I, pp. 251–272. Akademie Verlag, Berlin (1994)
37. Kulhavý, R., Zarrop, M.B.: On a general concept of forgetting. Int. J. of Control 58(4), 905–924 (1993)
38. Kullback, S., Leibler, R.: On information and sufficiency. Annals of Mathematical Statistics 22, 79–87 (1951)
39. Kumar, P.: A survey on some results in stochastic adaptive control. SIAM J. Control and Applications 23, 399–409 (1985)
40. Lainiotis, D.: Partitioned estimation algorithms, i: Nonlinear estimation. Information Sciences 7, 203–235 (1974)
41. Linden, G., Hanks, S., Lesh, N.: Interactive assessment of user preference models: The automated travel assistant. In: Proceedings of User Modelling 1997 (1997)
42. Loeve, M.: Probability Theory. van Nostrand, Princeton (1962), Russian translation, Moscow
43. Nagy, I., Suzdaleva, E., Kárný, M., Mlynářová, T.: Bayesian estimation of dynamic finite mixtures. Int. Journal of Adaptive Control and Signal Processing 25(9), 765–787 (2011)
44. Nelsen, R.: An Introduction to Copulas. Springer, New York (1999)
45. O'Hagan, A., Buck, C.E., Daneshkhah, A., Eiser, J.R., Garthwaite, P.H., Jenkinson, D.J., Oakley, J., Rakow, T.: Uncertain judgement: eliciting experts' probabilities. John Wiley & Sons (2006)

46. Osborne, M., Rubinstein, A.: A course in game theory. MIT Press (1994)
47. Peterka, V.: Bayesian system identification. In: Eykhoff, P. (ed.) Trends and Progress in System Identification, pp. 239–304. Pergamon Press, Oxford (1981)
48. Rao, M.: Measure Theory and Integration. John Wiley, New York (1987)
49. Savage, L.: Foundations of Statistics. Wiley, New York (1954)
50. Shi, R., MacGregor, J.: Modelling of dynamic systems using latent variable and subspace methods. J. of Chemometrics 14(5-6), 423–439 (2000)
51. Shore, J., Johnson, R.: Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. IEEE Tran. on Information Theory 26(1), 26–37 (1980)
52. Si, J., Barto, A., Powell, W., Wunsch, D. (eds.): Handbook of Learning and Approximate Dynamic Programming. Wiley-IEEE Press, Danvers (2004)
53. Šindelář, J., Vajda, I., Kárný, M.: Stochastic control optimal in the Kullback sense. Kybernetika 44(1), 53–60 (2008)