Czech Technical University in Prague
Faculty of Nuclear Sciences and Physical Engineering
Department of Mathematics

Cooperation Methods in Bayesian Decision Making with Multiple Participants

Jan Kracík

Ph.D. Thesis February, 2009

Acknowledgements

I am thankful to Miroslav Kárný for being my supervisor for this thesis.

This work was supported by grants GAAV 1ET100750401, GAČR 102/08/0567, and MŠMT 1M0572.

Symbols and Notation

The following symbols and notation are generally used throughout the thesis. The notation used for particular quantities, sets, constants, etc., is introduced in the text. In the list, the symbol x stands for an arbitrary random or nonrandom quantity.

$\propto$	proportionality sign
$\equiv$	equality by definition
x'	transposition of vector (matrix) x
x^*	range of x
$\mathring{x}$	cardinality of set x^*
x_i, x_j	i-th and j-th entry of vector x, respectively
x_t, x_τ	x at time t and τ, respectively; time indices are exclusively denoted by t and τ (with additional indices eventually), which makes them distinguishable from vector entries
$x^{t_1:t_2}$	data records from time t_1 till time t_2, i.e., $x^{t_1:t_2} = (x_{t_1}, x_{t_1+1}, \ldots, x_{t_2})$
$f(\cdot), g(\cdot), h(\cdot)$	probability density functions (pdfs) of the random quantity specified by the argument
$f(\cdot\|\cdot), g(\cdot\|\cdot), h(\cdot\|\cdot)$	conditional pdfs
$\mathcal{F}(x)$	set of all pdfs of random quantity x
${}^p x$	random quantity related to the p-th participant
${}^p f(\cdot), {}^p f(\cdot\|\cdot)$	pdfs related to the p-th participant
$\mathsf{D}(f(x)\|\|g(x))$	Kullback-Leibler (KL) divergence of pdf $f(x)$ from pdf $g(x)$
$\mathsf{K}(f(x), g(x))$	Kerridge inaccuracy of pdfs $f(x)$ and $g(x)$
$F << G$	probability distribution F is absolutely continuous with respect to probability distribution G
$\mathsf{E}[x], \mathsf{E}[x\|y]$	expectation of x and conditional expectation of x given y; the corresponding probability distribution always follows from the context
$\mathbb{R}$	set of all real numbers
$\mathbb{R}^n$	set of all real vectors of dimension n
$\mathbb{R}^{m,n}$	set of all real matrices of dimension m, n
$\mathbb{N}$	set of all natural numbers
M^c	complement of set M
$\|M\|$	cardinality of set M
$I_M(x)$	characteristic function of set M
$\det A$	determinant of square matrix A
$k := l$	set value l to variable k – used in descriptions of algorithms
$\delta(\cdot)$	Dirac delta function
$Di_\alpha(x)$	Dirichlet pdf of random quantity x with parameter α
$\mathcal{N}_x(\mu, \Sigma)$	Gaussian pdf of a random quantity x with mean μ and covariance matrix Σ
$\Gamma(x)$	gamma function

General Conventions

- Random quantities and their realizations are not distinguished by the notation.
- Probability density functions are distinguished by their parameters, i.e., $f(x), f(y)$ are different pdfs of different random quantities.
- Random quantities are supposed to be vector ones, if not stated otherwise.
- Probability distributions of continuous random quantities are supposed to be absolutely continuous and have pdfs, if not stated otherwise.
- For simplicity, the term pdf is used also for discrete quantities. In such cases it should be automatically understood as a probability mass function.
- Integral $\int \cdot \, \mathrm{d}x$ is automatically understood as a definite one over the set x^*.
- Vectors are supposed to be column ones, if not stated otherwise.
- Whenever vector entries are distinguished by numerical indices, they are supposed to be ordered so that their indices form an increasing sequence. This agreement enables unambiguous splitting and merging of (sub)vectors in algorithms. An analogical agreement holds for matrices as well.
- The notation $\overbrace{=}^{(x.y)}$, $\overbrace{\geq}^{(x.y)}$, etc., means that the relation under the curly bracket follows from relation $(x.y)$.

Abbreviations

pdf	probability density function
FPD	fully probabilistic design
ML	maximum likelihood

Chapter 1

Introduction

This work has its origin in the GAAV project 1ET100750401 BADDYR – Bayesian Adaptive Distributed Decision Making – solved in the Department of Adaptive Systems in the Institute of Information Theory and Automation in years 2004-2007. The objective of the BADDYR project was to develop a theoretical and algorithmic background for a distributed decision making with multiple Bayesian decision makers. The results of the BADDYR project are further developed in the GAČR project 102/08/0567 – Fully probabilistic design of dynamic decision strategies. This work contributes to these projects by a design of methods for communication of the decision makers.

1.1 Motivation

Decision making is a process in which an entity, called a decision maker, selects one from at least two actions. In order to make such process rational in a common sense, it is necessary to select the action with respect to its possible consequences. Whenever these consequences are supposed to be affected by phenomenons that are unknown to the decision maker at the time of making the decision, we speak about a decision making under uncertainty. The fact that in the real world almost any decision is accompanied by certain amount of uncertainty makes this area extremely important for practical applications.

The Department of Adaptive Systems belongs to the research groups that are focused on the decision making under uncertainty. The widely used Bayesian theory serves here as a main framework for treating the uncertainty [29], together with a specific method for a design of decision strategies, called the fully probabilistic design [27].

The Bayesian theory offers a normative approach to the decision making under uncertainty, which guarantees a well specified kind of rationality of the decision procedure [10], [25]. The characteristic features of the Bayesian theory are that the uncertainty is reflected by probabilistic modeling of a system, and that the objectives of the decision making are expressed in terms of a utility of the consequences. The optimal decision is then selected as the one which maximizes the expected utility. The use of probabilistic models is also important from a pragmatic point of view as it allows to employ a large scale of theoretical results and algorithmic solutions from the probability theory. The fully probabilistic design represents an alternative approach to the design of decision strategies based on maximization of the expected utility. Its main asset is that it has an explicit solution [30], [49].

The Bayesian theory and fully probabilistic design have been successfully applied in various fields, including industrial applications [19], traffic control [44], nuclear medicine [24], e-democracy [31], and financial markets [48]. A universal applicability of the adopted framework has been verified, e.g., in the EU project IST-1999-12058 ProDaCTool, the aim of which was to develop a domain independent decision-support system for complex industrial processes [29].

However, in practice the applicability of the Bayesian decision making is restricted by an extent of the given task. It is partially caused by the computational complexity, that grows quickly with increasing dimension of the task, and can easily exceed limited computational resources of a decision maker. There is also another, rather conceptual, reason: in large tasks it is often difficult to express prior information and objectives in a proper form. Prior information and objective specification are frequently available

piecewise, i.e., separately for different parts of the system. Moreover, these parts can arbitrarily overlap. Then, it can easily happen that for some part of the system there are several inconsistent prior information pieces available. Such a situation can arise naturally, e.g., as a consequence of the fact that the prior information represents a partial, and thus imprecise, knowledge. The objectives can naturally differ whenever there are several parties interested in the decision making.

1.2 Problem Formulation

The need for a feasible solution of large tasks led to the idea of a distributed decision making with multiple Bayesian decision makers. The staring point of the approach followed in the BADDYR project is that a group of Bayesian decision makers, here called participants, acts in a given system, whereas

- each participant deals with only a part of the system;
- the parts treated by individual participants may arbitrarily overlap;
- prior knowledge as well as objectives of individual participants need not be consistent in any way;
- the decision strategies of individual participants are required to be designed by the participants themselves, i.e., no common mediator is considered.

The main assets of this concept, in comparison with the decision making based on a single decision maker, are:

- The overall computational complexity can be much lower as it depends approximately linearly on a number of participants, whereas the decision making tasks solved by the individual participants can be relatively small.
- The freedom in the prior knowledge and objective setting allows to specify them piecewise, i.e., individually for each participant without a necessity to state them for the complete system.

Moreover, the individual participants can employ much of the solutions previously developed for a single decision maker. A downside of the adopted approach is that local models used by the individual participants cannot reflect all the complex relations that could be modelled by a single decision maker dealing with the complete system. On that account, any distributed solution is necessarily just an approximation of a centralized decision making. However, a rigorous treatment of the addressed problem is an extremely hard task as it requires a solution of problems related to multiplicity of objectives and prior knowledge and inference of distributed procedures. Furthermore, it inherits all technical difficulties arising in the Bayesian decision making with a single participant, such as intractability of the treated probability distributions. This tends us towards a more pragmatic approach based on the following idea.

A distributed solution is trivially reached if the participants act independently as if they were the only decision makers in their parts of the system. Nevertheless, in such case the group decision making can easily become noneffective due to differences in objectives and prior knowledge. It is caused by the fact that the participants acquire information about objectives, knowledge, and decision strategies of the other participants only indirectly through their actions and their consequences. Apparently, this process is slow and thus inefficient. It is expected that the group decision making becomes more efficient if the participants are provided with cooperation means which allow them to communicate directly and to exploit the information acquired in this way. As a distributed solution is aimed at, the communication cannot be realized centrally via a common facilitator, but must be performed directly between the participants. Moreover, the participants are not supposed to deal with the complete system, thus the communication is meaningful only between participants which deal with, at least partially, overlapping parts of the system. A pair of such participants is called neighbours. A design of suitable cooperation methods was considered as an important part of the BADDYR project, and forms the primary aim of this thesis. However, during the work it came out that the proposed "soft" formulation of the multiple participant decision making is not sufficient for the addressed task and that a deeper insight into the problem must be acquired. On that account, a secondary aim of the work is to provide a more detailed analysis of the multiple participant decision making itself.

It should be stressed that our approach is to be taken rather as a survey of possible extensions of a single participant towards the multiple participant decision making. Although we strive for a normative nature of our results, we do not intend to build a theory of the multiple participant decision making from the start.

1.3 State of the Art

There is a rich bibliography related to various aspects of the Bayesian decision making. The theoretical background can be found, e.g., in [8], [10], [16], or [25]. Somewhat more application-oriented publication is, e.g., [29]. This book also stands as a basis of our treatment of the dynamic Bayesian decision making. A universally accepted theory of the Bayesian decision making with multiple participants is, however, missing. On the other hand, particular problems encountered in this area are widely addressed.

The most significant class of such problems arise in connection with combining multiple, potentially contradicting, uncertainty assessments expressed in a form of probability distributions. In the literature, it is often referred to as an aggregation of expert opinions, opinion pooling, or consensus emergence. An extensive annotated bibliography on this topic can be found, e.g., in [22]. Mostly, the result of the pooling process is to be expressed in a form of a single probability distribution. This is also our primary preference.

A significant part of works in this field leans on the Bayesian theory. Many of them, e.g., [12], [13], [15], [21], [41], follow the approach introduced in [43]. The core idea here is that the experts' probability distributions are to be taken as data and processed by a decision maker in a standard Bayesian way. The aggregated opinion is represented by a posterior probability distribution. Note that it is commonly assumed that all experts' distributions are processed by a central decision maker. Although we have claimed that any centralized decision making is undesirable in our approach, it turns out that such a decision maker must be considered as a halfway house towards the distributed solution. This makes the works assuming a central decision maker relevant also for our approach.

A key element of all Bayesian methods is the decision maker's likelihood function for experts' opinions. In spite of its significance, a choice of a suitable likelihood is addressed very shallowly in general. It is typically simply assumed that the decision maker has a proper knowledge on the credibility of individual information sources, see, e.g., [15], [26], [43].

On the other hand, a frequently discussed issue is an impact of dependencies among experts' opinions on the resulting distribution [11], [13], [21]. Normal probability distributions are often employed for modeling the dependencies [14], [51]. Other approaches are based on models using t-distributions [41] or copulas [26]. However, most of these models are rather ad-hoc and their suitability can be verified only empirically in particular applications. Moreover, parameters of these dependency models are again generally supposed to be assigned by the decision maker according to its knowledge on the information sources.

Our approach to combining participants' uncertainty assessments does not directly follow any of the above mentioned works. It is partially due to the discussed weak points and partially due to the rather specific conditions: The participants can possibly employ different parametric models and thus their knowledge is expressed by probability distributions of different random quantities. Moreover, the considered central decision maker is just ancillary. On that account, a prior knowledge employed on its level should be minimal, or ideally (but unrealistically) empty. To our knowledge, these issues are not satisfactorily treated by the existing solutions.

Another problem related to the multiple participant decision making arises from the multiplicity of objectives. However, contrary to the participants' knowledge, the objectives of individual decision makers can differ naturally due to their different priorities. Thus, no other possible sources of inconsistency need to be necessarily considered in this case, and the common objectives can be established as some kind of compromise among the participants' ones. For combining of participants' objectives the results presented in [32] are partially employed.

1.4 Aim of the Work

As stated above, in this work we consider a group of Bayesian decision makers (participants) acting in some system. Each participant is supposed to deal with a part of the system, whereas these parts may arbitrarily overlap. Prior knowledge as well as objectives of the participants need not be a priori consistent in any way.

The aim of this work is:

- to provide a survey of a possible extension of the single participant towards the multiple participant decision making;
- to design practically feasible procedures that allows the neighbours to communicate information on their objectives and knowledge about the system, and to exploit them to enhance a quality of the decision making.

1.5 Structure of the Work

Chapter 2 summarizes the theoretical background used in this work. The attention is payed especially to the dynamic Bayesian decision making and the fully probabilistic design, including commonly used approximate methods.

Chapter 3 provides an introduction to the multiple participant decision making. Due to the complexity of the addressed problem, it is not aspired to give its rigorous analysis. Instead, several issues concerning this area are discussed and illustrated by a case study. The results acquired in this chapter motivate the design of the cooperation methods in Chapters 4 and 5.

In Chapter 4, a method which allows a group of participants to find common objectives is presented. This method is specially designed for participants employing the fully probabilistic design.

Chapter 5 describes a method for knowledge sharing between neighbours employing different probabilistic models of their parts of the system.

In the appendix, definitions and properties of binary relation, which are used in the thesis, are summarized. Furthermore, it includes a short discussion on the fully probabilistic design, which is referred in Chapter 3.

Chapter 2

Theoretical Background

This chapter contains a brief overview of the most important theoretical means used in the thesis. Section 2.1 recalls elementary operations with pdfs. In Section 2.2, properties of the Kullback-Leibler divergence and the Kerridge inaccuracy, which are used for quantification of discrepancy of pdfs, are mentioned. In Section 2.3, basic elements of the Bayesian decision making are summarized. Finally, Section 2.4 is focused on a special method for the design of decision strategies referred to as the fully probabilistic design.

2.1 Basic Calculus with Pdfs

Consider a joint pdf $f(x_1, x_2, x_3)$ of (possibly multivariate) random quantities x_1, x_2, x_3. The following manipulations with pdfs are frequently used in Bayesian analysis. For simplicity, we assume here that the right-hand sides of the relations below are well defined.

Conditioning	$f(x_1\|x_2,x_3) = \frac{f(x_1,x_2\|x_3)}{f(x_2\|x_3)}$
Chain rule	$f(x_1,x_2\|x_3) = f(x_1\|x_2,x_3)f(x_2\|x_3)$
Marginalization	$f(x_2\|x_3) = \int f(x_1,x_2\|x_3)\,\mathrm{d}x_1$
Bayes rule	$f(x_1\|x_2,x_3) = \frac{f(x_2\|x_1,x_3)f(x_1\|x_3)}{f(x_2\|x_3)} = \frac{f(x_2\|x_1,x_3)f(x_1\|x_3)}{\int f(x_2\|x_1,x_3)f(x_1\|x_3)\mathrm{d}x_1}$

2.2 Discrepancy of Pdfs

For quantification of discrepancy of a pair of pdfs the Kullback-Leibler divergence and the Kerridge inaccuracy are widely used in this work.

2.2.1 Kullback-Leibler Divergence

The Kullback-Leibler divergence [39] is a member of a class of so called f-divergences [40] known from the information theory. For a pair of pdfs $f(x), g(x)$ corresponding to distributions F and G respectively, the Kullback-Leibler divergence is defined by

$$\mathsf{D}\left(f(x)||g(x)\right) = \begin{cases} \int f(x)\ln\frac{f(x)}{g(x)}\mathrm{d}x & \text{for } F << G \\ \\ +\infty & \text{otherwise,} \end{cases} \tag{2.1}$$

where $F << G$ denotes absolute continuity of F with respect to G, and the integrand is defined using the convention $0\ln 0 = 0$. Basic properties of the Kullback Leibler divergence are as follows ($\mathcal{F}(x)$ denotes a set of all pdfs of x).

- nonnegativity: $\forall f(x), g(x) \in \mathcal{F}(x),\ \mathsf{D}\left(f(x)||g(x)\right) \geq 0$ (2.2)
- asymetry: $\exists f(x), g(x) \in \mathcal{F}(x),\ \mathsf{D}\left(f(x)||g(x)\right) \neq \mathsf{D}\left(g(x)||f(x)\right)$
- $\forall f(x), g(x) \in \mathcal{F}(x),\ \mathsf{D}\left(f(x)||g(x)\right) = 0$ iff $f = g$ (2.3)
- convexity in both arguments: $\forall f(x), g(x), h(x) \in \mathcal{F}(x), \forall \alpha \in [0,1]$,

$$\begin{aligned} \mathsf{D}\left(\alpha f(x) + (1-\alpha)h(x)||g(x)\right) &\leq \alpha \mathsf{D}\left(f(x)||g(x)\right) + (1-\alpha)\mathsf{D}\left(h(x)||g(x)\right) \\ \mathsf{D}\left(f(x)||\alpha g(x) + (1-\alpha)h(x)\right) &\leq \alpha \mathsf{D}\left(f(x)||g(x)\right) + (1-\alpha)\mathsf{D}\left(f(x)||h(x)\right) \end{aligned} \tag{2.4}$$

For a joint pdf $f(x,y)$ and a conditional pdf $g(y|x)$, a conditional version of the Kullback-Leibler divergence $\mathsf{D}\left(f(x,y)||g(y|x)\right)$ is defined by

$$\mathsf{D}\left(f(x,y)||g(y|x)\right) = \int f(x)\mathsf{D}\left(f(y|x)||g(y|x)\right)\mathrm{d}x, \tag{2.5}$$

where, for a fixed x, $\mathsf{D}\left(f(y|x)||g(y|x)\right)$ is understood as a Kullback-Leibler divergence of a pair of pdfs from $\mathcal{F}(y)$.

Using (2.5), we get for a pair of joint pdfs $f(x,y), g(x,y)$

$$\mathsf{D}\left(f(x,y)||g(x,y)\right) = \int f(x)f(y|x)\left(\ln\frac{f(x)}{g(x)} + \ln\frac{f(y|x)}{g(y|x)}\right)\mathrm{d}x\mathrm{d}y = \mathsf{D}\left(f(x)||g(x)\right) + \mathsf{D}\left(f(x,y)||g(y|x)\right). \tag{2.6}$$

From (2.6) we get immediately these properties of the Kullback-Leibler divergence:

$$\mathsf{D}\left(f(x)||g(x)\right) = \mathsf{D}\left(f(x,y)||g(x)f(y|x)\right), \tag{2.7}$$

$$\mathsf{D}\left(f(x)||g(x)\right) \leq \mathsf{D}\left(f(x,y)||g(x,y)\right). \tag{2.8}$$

Proposition 2.2.1 *For arbitrary pdfs $f(x), g(x) \in \mathcal{F}(x)$ and a measurable set $M \subset x^*$, let $a = \int_M f(x)\mathrm{d}x, b = \int_M g(x)\mathrm{d}x$. Then*

$$\mathsf{D}\left(f(x)||g(x)\right) \geq a\ln\frac{a}{b} + (1-a)\ln\frac{1-a}{1-b}, \tag{2.9}$$

using the conventions $0\ln 0 = 0, 0\ln\frac{0}{0} = 0, \ln\frac{c}{0} = +\infty$ for $c > 0$.

Proof: Suppose that $a, b \in (0,1)$. Then

$$\begin{aligned} \mathsf{D}\left(f(x)||g(x)\right) &= \int_M f(x)\ln\frac{f(x)}{g(x)}\mathrm{d}x + \int_{M^C} f(x)\ln\frac{f(x)}{g(x)}\mathrm{d}x \\ &= a\int_M \frac{f(x)}{a}\left(\ln\frac{\frac{f(x)}{a}}{\frac{g(x)}{b}} + \ln\frac{a}{b}\right)\mathrm{d}x + (1-a)\int_{M^C}\frac{f(x)}{1-a}\left(\ln\frac{\frac{f(x)}{1-a}}{\frac{g(x)}{1-b}} + \ln\frac{1-a}{1-b}\right)\mathrm{d}x \\ &= a\ln\frac{a}{b} + (1-a)\ln\frac{1-a}{1-b} + \mathsf{D}\left(\frac{1}{a}f(x)I_M(x)\middle|\middle|\frac{1}{b}g(x)I_M(x)\right) \\ &\quad + \mathsf{D}\left(\frac{1}{1-a}f(x)I_{M^C}(x)\middle|\middle|\frac{1}{1-b}g(x)I_{M^C}(x)\right) \\ &\geq a\ln\frac{a}{b} + (1-a)\ln\frac{1-a}{1-b}. \end{aligned}$$

Verification of (2.9) for $a \in \{0,1\}$ or $b \in \{0,1\}$ is trivial. □

Note that a proposition analogous to (2.2.1) can be stated for any finite partition of x^*; for more details see [40].

2.2.2 Kerridge Inaccuracy

The Kerridge inaccuracy [35] does not belong among f-divergences; nevertheless, for its tight relation to the Kullback-Leibler divergence it is often employed in this work. We use the following definition of the Kerridge inaccuracy.

For a pair of absolutely continuous distributions F and G having pdfs $f(x)$ and $g(x)$ respectively, the Kerridge inaccuracy is defined by

$$\mathsf{K}\left(f(x), g(x)\right) = \begin{cases} \int f(x) \ln \frac{1}{g(x)} \mathrm{d}x & \text{for } F << G \\ \\ +\infty & \text{otherwise} \end{cases}, \tag{2.10}$$

where the integral is defined using the convention $0 \ln 0 = 0$. For pdfs $f(x), g(x)$ of a discrete quantity x (i.e., probability mass functions) the Kerridge inaccuracy is defined analogously with sum instead of integral. In this text, the Kerridge inaccuracy is used also in a more general sense: let $r_{x_1,\ldots,x_n}(x)$ be an empirical pdf, i.e., $r_{x_1,\ldots,x_n}(x) = \frac{1}{n}\sum_{i=1}^{n} \delta(x - x_i)$, for some $n \in \mathbb{N}$ and $x_1, \ldots, x_n \in x^*$, where $\delta(x)$ denotes the Dirac delta function. The Kerridge inaccuracy of $r_{x_1,\ldots,x_n}(x)$ and an arbitrary pdf $g(x)$ corresponding to an absolutely continuous distribution is to be understood as

$$\mathsf{K}\left(r_{x_1,\ldots,x_n}(x), g(x)\right) \equiv \int \left(\frac{1}{n}\sum_{i=1}^{n} \delta(x - x_i)\right) \ln \frac{1}{g(x)} \mathrm{d}x = -\frac{1}{n} \ln \left(\prod_{i=1}^{n} g(x_i)\right). \tag{2.11}$$

Notice, that for a parametric pdf, say $g(x|\Theta)$, $\mathsf{K}\left(r_{x_1,\ldots,x_n}(x), g(x|\Theta)\right)$ is, up to a constant factor $-\frac{1}{n}$, a log-likelihood of the parameter Θ.

Basic properties of the Kerridge inaccuracy are:

- asymmetry: $\exists f(x), g(x) \in \mathcal{F}(x),\ \mathsf{K}\left(f(x), g(x)\right) \neq \mathsf{K}\left(g(x), f(x)\right)$,
- linearity in the first argument: $\forall f(x), g(x), h(x) \in \mathcal{F}(x), \forall \alpha \in [0, 1]$,

$$\mathsf{K}\left(\alpha f(x) + (1-\alpha)h(x), g(x)\right) = \alpha \mathsf{K}\left(f(x), g(x)\right) + (1-\alpha)\mathsf{K}\left(h(x), g(x)\right), \tag{2.12}$$

- convexity in the second argument: $\forall f(x), g(x), h(x) \in \mathcal{F}(x), \forall \alpha \in [0, 1]$,

$$\mathsf{K}\left(f(x), \alpha g(x) + (1-\alpha)h(x)\right) \leq \alpha \mathsf{K}\left(f(x), g(x)\right) + (1-\alpha)\mathsf{K}\left(f(x), h(x)\right),$$

- $\forall f(x), g(x) \in \mathcal{F}(x)$,

$$\mathsf{K}\left(f(x), f(x)\right) \leq \mathsf{K}\left(f(x), g(x)\right) \text{with equality iff } f = g. \tag{2.13}$$

For a joint pdf $f(x, y)$ and a conditional pdf $g(y|x)$, we define a conditional Kerridge inaccuracy

$$\mathsf{K}\left(f(x, y), g(y|x)\right) = \int f(x) \mathsf{K}\left(f(y|x), g(y|x)\right) \mathrm{d}x, \tag{2.14}$$

where, for fixed x, $\mathsf{K}\left(f(y|x), g(y|x)\right)$ is taken as a Kerridge inaccuracy of pdfs from $\mathcal{F}(y)$. Using (2.14), we get for a pair of joint pdf $f(x, y), g(x, y)$,

$$\mathsf{K}\left(f(x, y), g(x, y)\right) = \int f(x) f(y|x) \left(\ln \frac{1}{g(x)} + \ln \frac{1}{g(y|x)}\right) = \mathsf{K}\left(f(x), g(x)\right) + \mathsf{K}\left(f(x, y), g(y|x)\right). \tag{2.15}$$

Note that the Kerridge inaccuracy and the Kullback-Leibler divergence are related through the differential entropy (or entropy in case of discrete random quantities)

$$\mathsf{H}(f(x)) \equiv \mathsf{K}\left(f(x), f(x)\right) = -\int f(x) \ln f(x) \mathrm{d}x \tag{2.16}$$

by the equality

$$\mathsf{K}\left(f(x), g(x)\right) = \mathsf{D}\left(f(x) || g(x)\right) + \mathsf{H}(f(x)), \tag{2.17}$$

if both sides of (2.17) exist.

2.3 Bayesian Decision Making

Decision making task arises whenever an individual, referred to as a decision maker, dealing with some part of the world, denoted as a system, has to choose among at least two different actions. A criterion according to which the decision maker decides which action to choose is typically based on some assessing of consequences of the considered actions. The decision making then reduces to some kind of optimization. Selection of the best action becomes far more complicated whenever the individual is not able to determine exactly the consequences of the considered actions. In this case we speak about a decision making under uncertainty. The uncertainty can be caused by decision maker's incomplete knowledge about the system or by a random behaviour of the system itself. In fact, what is commonly taken as a random behaviour is typically just a consequence of a lack of knowledge.

Considering uncertainty, the decision maker faces a much wider problem than a pure optimization. Mainly, it is necessary to utilize some framework for treating the uncertainty. In the project, to which this thesis partially contributes, the Bayesian theory is employed for this purpose.

2.3.1 Introduction

The Bayesian theory represents a normative approach to the decision making under uncertainty. On the very fundamental level, the Bayesian theory can be build on a formal structure representing basic elements of the decision making – uncertain events, potential actions of a decision maker, their consequences, and preference ordering of actions. These elements are then required to satisfy a set of conditions, in the literature, somewhat inappropriately, referred to as axioms. There is a number of such axiomatic systems varying in details, see, e.g., [10]. However, all of them require the preference ordering on actions to be "rich enough" and consistent in some sense. Their common implication is that the decision making task can be equivalently formulated in terms of belief about uncertain events and preferences on consequences. It also follows from the axioms that the belief about the uncertain events has a structure of a probability measure depending on the selected action, and that the preferences can be expressed in terms of utility assigned by a utility function to the consequences. The preferences on actions are then induced by the expected utility. The optimal actions are the ones for which the expected utility is maximal. The notion of a utility is often substituted by a loss. The utility function is then substituted by a loss function and maximization of the expected utility is replaced by minimization of the expected loss. The loss functions are used also in this work.

2.3.2 Dynamic Bayesian Decision Making

Application of the Bayesian theory to dynamic decision tasks results in a probabilistic framework, a particular case of which is described in this paragraph.

Let the system be described by a sequence of random quantities $(a_t, \Theta_t, \Delta_t)_{t \in t^*}$, where t is a discrete quantity interpreted as time; typically $t^* \equiv \{1, 2, \ldots, \mathring{t}\}$. $\mathring{t} \in \mathbb{N}$ is called a decision horizon and represents the time to which the decision making is performed.

Actions a_t are quantities values of which are selected by the decision maker.

Observations Δ_t represent quantities realizations of which are observed on the system immediately after an action a_t is applied.

Internals Θ_t are unobservable quantities influencing the system behaviour. Internals can be hidden quantities with physical meaning as well as completely abstract quantities, e.g., unknown parameters of otherwise known probability distributions.

It is supposed that at each time t an action a_t is performed at first, then a new internal Θ_t is generated, and finally an observation Δ_t occurs.

The following notation and terminology are related to these quantities:

$x^{t_1:t_2}$ denotes a sequence of random quantities $(x_{t_1}, \ldots, x_{t_2})$, for $t_1, t_2 \in t^*$.

Data $d_t \equiv (a_t, \Delta_t)$ represent random vectors consisting of actions a_t and observations Δ_t at time t.

Trajectory of the system is a sequence of realizations of all involved quantities to the decision horizon, i.e., $(a^{1:\mathring{t}}, \Theta^{1:\mathring{t}}, \Delta^{1:\mathring{t}})$. The notation used is to be understood so that $(a^{1:\mathring{t}}, \Theta^{1:\mathring{t}}, \Delta^{1:\mathring{t}}) \equiv (a_1, \Theta_1, \Delta_1, \ldots, a_{\mathring{t}}, \Theta_{\mathring{t}}, \Delta_{\mathring{t}})$.

Behavior of the system is taken, more or less, in its intuitive meaning. Formally it could be characterized, e.g., as a probability distribution (not necessarily corresponding to the adopted models (2.18), (2.19)) over all possible trajectories; however, a precise definition is not required in the text.

In order to establish a probabilistic description of the system, the decision maker specifies the following pdfs:

Observation model is a collection of conditional pdfs

$$\left(f(\Delta_t|a_t, d^{1:t-1}, \Theta^{1:t})\right)_{t\in t^*},$$

modelling dependence of observations on all preceding quantities. The observation model is supposed to be selected so that, given the present action a_t and past data $d^{1:t-1}$, the observation Δ_t depends only on the present internal Θ_t. Thus, its form reduces to

$$\left(f(\Delta_t|a_t, d^{1:t-1}, \Theta_t)\right)_{t\in t^*}. \tag{2.18}$$

Time-evolution model

$$\left(f(\Theta_t|a_t, d^{1:t-1}, \Theta^{1:t-1})\right)_{t\in t^*}$$

models an evolution of the unknown internals. Similarly as in the case of the observation model, given the present action a_t and the preceding data $d^{1:t-1}$, the internal Θ_t is supposed to depend only on the directly preceding internal Θ_{t-1}. The time-evolution model then reduces to

$$\left(f(\Theta_t|a_t, d^{1:t-1}, \Theta_{t-1})\right)_{t\in t^*} \tag{2.19}$$

Prior pdf

$$f\left(\Theta_0\right) \tag{2.20}$$

describes initial knowledge about the unknown internal Θ_0.

Objectives of the decision making are expressed in terms of preferences on a set of all trajectories. Namely, the decision maker has to select a loss function

$$L : (a^{1:\mathring{t}})^* \times (\Delta^{1:\mathring{t}})^* \times (\Theta^{1:\mathring{t}})^* \to [C, +\infty], \tag{2.21}$$

for some $C \in \mathbb{R}$, assigning to any system trajectory a value in a sense of a negative profit, i.e., the smaller value of the loss function, the more preferred trajectory.

A decision strategy, the design of which is a primary aim of the decision maker, is a sequence of mappings, indexed by time t, assigning a particular value of a_t or a pdf of a_t to the past data $d^{1:t-1}$. In the former case the decision strategy is called deterministic; in the later case it is randomized. In this text we deal mainly with randomized decision strategies, so the label "randomized" is mostly omitted. A decision strategy is represented by a collection of conditional pdfs

$$\left(f(a_t|d^{1:t-1})\right)_{t\in t^*}. \tag{2.22}$$

The individual conditional pdfs $f\left(a_t\middle| d^{1:t-1}\right)$ are called decision rules.

We assume that decision strategies generate actions in dependence on past data $d^{1:t-1}$. More precisely, it is assumed that actions a_t and internal quantities $\Theta^{1:t-1}$ are conditionally independent given the observed data $d^{1:t-1}$, i.e.,

$$f(a_t|d^{1:t-1}, \Theta^{1:t-1}) = f(a_t|d^{1:t-1}). \tag{2.23}$$

This assumption is referred to as natural conditions of control [29]. A direct consequence of (2.23) is a conditional independence of Θ_{t-1} and a_t given data $d^{1:t-1}$

$$f(\Theta_{t-1}|a_t, d^{1:t-1}) = f(\Theta_{t-1}|d^{1:t-1}). \tag{2.24}$$

Bayesian theory establishes [29] that the optimal decision strategy ${}^{O}R \equiv \left({}^{O}f(a_t|d^{1:t-1})\right)_{t\in t^*}$ is a decision strategy minimizing the expected loss

$$\mathsf{E}\left[L(a^{1:\mathring{t}}, \Delta^{1:\mathring{t}}, \Theta^{1:\mathring{t}})\right], \tag{2.25}$$

where the expectation is taken with respect to the joint pdf on the set of all trajectories

$$f(a^{1:\mathring{t}}, \Delta^{1:\mathring{t}}, \Theta^{1:\mathring{t}}) = \int \prod_{t=1}^{\mathring{t}} f(\Delta_t|a_t, d^{1:t-1}, \Theta_t) f(\Theta_t|a_t, d^{1:t-1}, \Theta_{t-1}) f(a_t|d^{1:t-1}) f(\Theta_0)\mathrm{d}\Theta_0. \tag{2.26}$$

$f(a^{1:\mathring{t}}, \Delta^{1:\mathring{t}}, \Theta^{1:\mathring{t}})$ in (2.26) is defined by the observation model (2.18), the time-evolution model (2.19), the searched decision strategy (2.22), and the prior pdf (2.20). The optimal strategy ${}^{O}R$ is typically constructed using a stochastic version of dynamic programming [7].

The optimization procedure can be taken as consisting of two interconnected parts, shortly referred to as learning and design. The objective of learning is to extract information about the unobserved internals Θ_t from past data. In the design phase an optimal decision strategy is constructed using the results provided by learning. In case of a sub-optimal design, which employs various approximate techniques and which is common in applications, learning and design are typically performed successively. On that account, we describe the two phases separately.

Learning

At any time t, all the available knowledge about the unknown internals, combining the initial knowledge and the knowledge acquired from data $d^{1:t}$, is represented by a conditional pdf

$$f(\Theta_t|d^{1:t}) \tag{2.27}$$

called Bayesian filtration [29]. Its evaluation is the objective of learning.

Using observation model (2.18), time evolution model (2.19), prior pdf (2.20), and the conditional independence (2.24), the filtration (2.27) is evaluated in a recursive way consisting of the following steps. The recursion starts for $t = 0$ with the prior pdf $f(\Theta_0)$.

1. *Time updating*

$$\begin{aligned} f(\Theta_t|a_t, d^{1:t-1}) &= \int f(\Theta_t|a_t, d^{1:t-1}, \Theta_{t-1}) f(\Theta_{t-1}|a_t, d^{1:t-1})\mathrm{d}\Theta_{t-1} = \\ &= \int f(\Theta_t|a_t, d^{1:t-1}, \Theta_{t-1}) f(\Theta_{t-1}|d^{1:t-1})\mathrm{d}\Theta_{t-1} \end{aligned} \tag{2.28}$$

2. *Data updating*

$$f(\Theta_t|d^{1:t}) = \frac{f(\Delta_t|a_t, d^{1:t-1}, \Theta_t) f(\Theta_t|a_t, d^{1:t-1})}{f(\Delta_t|a_t, d^{1:t-1})}, \quad \text{where} \tag{2.29}$$

$$f(\Delta_t|a_t, d^{1:t-1}) = \int f(\Delta_t|a_t, d^{1:t-1}, \Theta_t) f(\Theta_t|a_t, d^{1:t-1})\mathrm{d}\Theta_t \tag{2.30}$$

The conditional pdf $f(\Delta_t|a_t, d^{1:t-1})$ in general, i.e., not necessarily resulting from (2.30), is called an outer model of the system. The outer model of the system (2.30) acquired in learning, is referred to as a predictive pdf.

Design

The optimal decision strategy is constructed by a recursive procedure consisting in evaluation of functions $\mathcal{V}_t : (d^{1:t-1})^* \to \mathbb{R}$,

$$\mathcal{V}_t(d^{1:t-1}) = \min_{f(a_t|d^{1:t-1})} \mathsf{E}\left[\mathcal{V}_{t+1}\left(d^{1:t}\right) \middle| d^{1:t-1}\right]. \tag{2.31}$$

The recursion is performed in a backward manner. It starts at $t = \mathring{t}$ and the initial function $\mathcal{V}_{\mathring{t}+1}(d^{1:\mathring{t}})$ is defined by

$$\mathcal{V}_{\mathring{t}+1}(d^{1:\mathring{t}}) = \mathsf{E}\left[L(a^{1:\mathring{t}}, \Delta^{1:\mathring{t}}, \Theta^{1:\mathring{t}})\middle| d^{1:\mathring{t}}\right]. \tag{2.32}$$

For evaluation of (2.32) a conditional pdf $f(\Theta^{1:\mathring{t}}|d^{1:\mathring{t}})$ is needed. For the observation model (2.18), time evolution model (2.19), and under natural conditions of control (2.23), it can be constructed using a recursive relation

$$f(\Theta^{1:t}|d^{1:t}) = \frac{f(\Delta_t|a_t, \Theta_t, d^{1:t-1}) f(\Theta_t|a_t, d^{1:t-1}, \Theta_{t-1})}{f(\Delta_t|a_t, d^{1:t-1})} f(\Theta^{1:t-1}|d^{1:t-1}),$$

where $f(\Delta_t|a_t, d^{1:t-1})$ is given by (2.30). The recursion starts with the prior pdf $f(\Theta_0)$.

The expectation in (2.31) is taken with respect to

$$f\left(d_t\middle| d^{1:t-1}\right) = f\left(\Delta_t\middle| a_t, d^{1:t-1}\right) f\left(a_t\middle| d^{1:t-1}\right),$$

determined by the predictive pdf (2.30) and the particular decision rule $f\left(a_t\middle| d^{1:t-1}\right)$. The optimal strategy ${}^{O}R$ then consists of minimizing arguments in (2.31), i.e.,

$${}^{O}f(a_t|d^{1:t-1}) \in \underset{f(a_t|d^{1:t-1})}{\operatorname{argmin}} \mathsf{E}\left[\mathcal{V}_{t+1}(d^{1:t})\middle| d^{1:t-1}\right]. \tag{2.33}$$

${}^{O}f(a_t|d^{1:t-1})$ in (2.33) are not determined uniquely. Nevertheless, as all minimizers in (2.33) lead to the same expected loss, the optimal decision rules ${}^{O}f(a_t|d^{1:t-1})$ can be selected arbitrarily. Because $\mathsf{E}\left[\mathcal{V}_{t+1}(d^{1:t})\middle| d^{1:t-1}\right]$ in (2.31) depends linearly on $f(a_t|d^{1:t-1})$, it is clear that ${}^{O}f(a_t|d^{1:t-1})$ is any pdf which gives probability 1 to the set

$$\underset{a_t \in a^*}{\arg\min}\, \mathsf{E}\left[\mathcal{V}_{t+1}(d^{1:t})\middle| a_t, d^{1:t-1}\right],$$

and the optimal strategy can be always selected as a non-randomized one.

2.3.3 Practical Aspects

Decision strategies designed according to (2.33) are optimal ones within the Bayesian framework. However, the price paid for the optimality is often too high. Evaluation of functions $\mathcal{V}_t(d^{1:t-1})$, filtrations $f(\Theta_t|d^{1:t})$, decision strategies $f(a_t|d^{1:t-1})$, and others is computationally too expensive. Furthermore, these functions are typically too complex to be represented in a computer. In practice, it is necessary to adopt additional simplifying assumptions and employ some approximate techniques [29]. In what follows, the most common assumptions and techniques simplifying learning and design are shortly mentioned.

Reduced time-evolution model

Internals Θ_t frequently represent unknown parameters (in a common sense) of the observation model (2.18) and are supposed to be time invariant. In this case the time evolution model (2.19) reduces to

$$f(\Theta_t|a_t, d^{1:t-1}, \Theta_{t-1}) = \delta(\Theta_t - \Theta_{t-1}), \tag{2.34}$$

time updating step (2.28) vanishes, and the evaluation of the conditional pdf $f\left(\Theta_t\middle| d^{1:t}\right)$ gets a simple form

$$f(\Theta|d^{1:t}) \propto f(\Delta_t|a_t, d^{1:t-1}, \Theta) f(\Theta|d^{1:t-1}), \tag{2.35}$$

where $\Theta \equiv \Theta_t$, forall $t \in t^*$. $f\left(\Theta\middle| d^{1:t}\right)$ in (2.35) is called a posterior pdf. If not said otherwise, it is assumed in the following text that the internals Θ_t are time invariant. The subscript t of the internals is then omitted.

Models with finite memory

In applications it is usually reasonable to suppose that observations Δ_t do not depend on a complete history of observations but only on a few last data records. In other words, it is supposed that the

observation model (2.18) has a form

$$f(\Delta_t|a_t, d^{1:t-1}, \Theta) = f(\Delta_t|a_t, d^{t-T:t-1}, \Theta), \tag{2.36}$$

for some $T \geq 1$. Furthermore, observations Δ_t often do not depend on all entries of vector $d^{t-T:t-1}$. Then, the observation model can be written in a form

$$f(\Delta_t|a_t, d^{1:t-1}, \Theta) = f(\Delta_t|a_t, \phi_{t-1}, \Theta), \tag{2.37}$$

where ϕ_{t-1}, referred to as a state vector, is a subvector of $d^{t-T:t-1}$, typically with a fixed structure; see Section 3.1. Note, that if an observation model with finite memory is employed, it is necessary to specify starting data $d^{1-T:0}$ needed for construction of initial state vectors $\phi_0 \ldots, \phi_{T-1}$.

Conjugate prior

A typical problem of the Bayesian learning is a form of a posterior pdf, which is often too complex to be treated in a computer. Under assumptions (2.34) and (2.36), this difficulty vanishes if the observation model is from an exponential family, i.e., the model can be expressed as

$$f(\Delta_t|a_t, \phi_{t-1}, \Theta) = A(\Theta) \exp(\langle B(\Psi_t), C(\Theta)\rangle), \tag{2.38}$$

where $\Psi_t' \equiv (\Delta_t', a_t', \phi_{t-1}')$, $A(\cdot)$ is a nonnegative scalar function defined on Θ^*, $B(\cdot)$, and $C(\cdot)$ are vector functions of the same dimension defined on Ψ_t^* and Θ^*, and $\langle\cdot,\cdot\rangle$ is a standard scalar product. For models from an exponential family the prior pdfs can be selected in a conjugate form [8], which is preserved during estimation. Considering such an observation model and a conjugate prior pdf in a form

$$f(\Theta) \propto A(\Theta)^{\nu_0} \exp(\langle V_0, C(\Theta)\rangle),$$

for some ν_0 and a vector V_0 such that $\int A(\Theta)^{\nu_0} \exp(\langle V_0, C(\Theta)\rangle)\mathrm{d}\Theta < +\infty$, the corresponding posterior pdf has a form

$$f(\Theta|d^{1:t}) \propto A(\Theta)^{\nu_t} \exp(\langle V_t, C(\Theta)\rangle), \tag{2.39}$$

where

$$\begin{aligned} \nu_t &= \nu_0 + t, \\ V_t &= V_0 + \sum_{\tau=1}^{t} B(\Psi_\tau). \end{aligned} \tag{2.40}$$

It is seen that for given functions $A(\cdot)$ and $C(\cdot)$ the posterior pdf (2.39) is entirely represented by the parameters ν_t and V_t.

Approximate learning

Observation models with conjugate prior pdfs are often too simple for practical applications. On that account, more complex models must be employed, which makes the learning technically difficult. A compromise between complexity of models and feasibility of learning can be achieved by approximate learning. We briefly mention two algorithms designed for Bayesian estimation of dynamic models – the quasi-Bayes algorithm [29] and the projection-based algorithm [3].

The quasi-Bayes algorithm is designed for estimation of finite probabilistic mixtures, i.e., observation models which can be expressed in a form

$$f(\Delta_t|a_t, \phi_{t-1}, \Theta) = \sum_{c=1}^{\mathring{c}} \alpha_c m(\Delta_t|a_t, \phi_{t-1}, \Theta_c), \tag{2.41}$$

where $\mathring{c} \in \mathbb{N}$, $\Theta \equiv (\alpha_1, \ldots, \alpha_{\mathring{c}}, \Theta_1, \ldots, \Theta_{\mathring{c}})$, α_c are non-negative weights such that $\sum_{c=1}^{\mathring{c}} \alpha_c = 1$, and pdfs $m\left(\Delta_t\,|a_t, \phi_{t-1}, \Theta_c\right)$, referred to as components of the mixture, are supposed to be from an exponential

family. The components are supposed to be of the same type, typically Gaussian pdfs, and differ only in values of their parameters Θ_c.

The quasi-Bayes algorithm [29] constructs an approximate posterior pdf in a product form

$$f(\Theta|d^{1:t}) = Di_\alpha(\kappa_t)\prod_{c=1}^{\mathring{c}} f(\Theta_c|d^{1:t}), \tag{2.42}$$

where $\kappa_t \equiv (\kappa_{1;t},\ldots,\kappa_{\mathring{c};t})$ is a vector with positive entries, $Di_\alpha(\kappa)$ is a Dirichlet pdf of $\alpha \equiv (\alpha_1,\ldots,\alpha_{\mathring{c}})$, i.e.,

$$Di_\alpha(\kappa_t) = \frac{\prod_{c=1}^{\mathring{c}} \alpha_c^{\kappa_{c;t}-1}}{B(\kappa_t)}, \quad B(\kappa_t) = \frac{\prod_{c=1}^{\mathring{c}}\Gamma(\kappa_{c;t})}{\Gamma(\sum_{c=1}^{\mathring{c}}\kappa_{c;t})},$$

and $\Gamma(\cdot)$ is the Gamma function [1]. Approximate posterior pdfs of component parameters $f(\Theta_c|d^{1:t})$ in (2.42) are supposed to be in a conjugate form to the observation models of components $m(\Delta_t|a_t,\phi_{t-1},\Theta_c)$. Data updating of the approximate posterior pdf (2.42) is then provided in the following way. For all $c \in c^*$,

$$\begin{aligned} \kappa_{c;t+1} &= \kappa_{c;t} + w_{c;t+1} \\ f(\Theta_c|d^{1:t+1}) &\propto \left(m\left(\Delta_{t+1}\left|a_{t+1},d^{1:t},\Theta_c\right.\right)\right)^{w_{c;t+1}} f(\Theta_c|d^{1:t}), \end{aligned} \tag{2.43}$$

where

$$w_{c;t+1} = \frac{\kappa_{c;t}\int m\left(\Delta_{t+1}\left|a_{t+1},\phi_t,\Theta_c\right.\right) f(\Theta_c|d^{1:t})\mathrm{d}\Theta_c}{\sum_{\tilde{c}=1}^{\mathring{c}}\kappa_{\tilde{c};t}\int m\left(\Delta_{t+1}\left|a_{t+1},\phi_t,\Theta_{\tilde{c}}\right.\right) f\left(\Theta_{\tilde{c}}\left|d^{1:t}\right.\right)\mathrm{d}\Theta_{\tilde{c}}}.$$

From (2.43) it is clear that the approximate posterior pdfs $f(\Theta_c|d^{1:t+1})$ remain in the conjugate form and their evaluation reduces to updating of finite dimensional statistics analogous to that in (2.39).

The Projection-based algorithm [3] can be seen as a generalization of the quasi-Bayes algorithm. Approximate posterior pdfs are searched within a pre-specified class of pdfs in the following way: the approximate posterior pdf at time $t+1$ is a pdf, which minimizes the Kullback-Leibler divergence from the correctly updated approximate pdf in time t. Experimental results indicate that the projection-based algorithm provides better results then the well established quasi-Bayes algorithm [5]. It is also worth to be mentioned that the projection-based algorithm can be used for estimation of more complex models then the finite mixtures (2.41); see [4].

Data-driven design

In general, the loss function (2.21) may depend on actions a_t, observations Δ_t, and unobservable internals Θ_t. However, often the objectives of the decision making do not involve internals Θ_t, typically if they represent unknown parameters of the observation model. A design with a loss function independent of the unobservable internals is called data-driven. The merit of the data-driven design is that the otherwise technically difficult evaluation of $\mathcal{V}_{\mathring{t}+1}(d^{1:\mathring{t}})$, see (2.32), becomes trivial as $\mathcal{V}_{\mathring{t}+1}(d^{1:\mathring{t}}) = L(d^{1:\mathring{t}})$.

Additive loss function

A loss function $L(d^{1:\mathring{t}},\Theta^{1:\mathring{t}})$ is called additive if it can be expressed in a form

$$L\left(d^{1:\mathring{t}},\Theta^{1:\mathring{t}}\right) = \sum_{t=1}^{\mathring{t}} l_t\left(d_t,\Theta_t\right), \tag{2.44}$$

where $l_t : d_t^* \times \Theta_t^* \to [C,+\infty]$ and $C \in \mathbb{R}$. Compared to the general case employing (2.31) and (2.32), the design with an additive loss function can be performed in an easier way. For functions $\mathcal{V}_t(d^{1:t-1})$ defined so that

$$\mathcal{V}_t\left(d^{1:t-1}\right) = \min_{\{f(a_{\tilde{t}}|d^{1:\tilde{t}-1})\}_{\tilde{t}=t}^{\mathring{t}}} \mathsf{E}\left[\sum_{\tau=t}^{\mathring{t}} l_t\left(d_\tau,\Theta_\tau\right)\middle| d^{1:t-1}\right]$$

we get a recursive relation

$$\mathcal{V}_t\left(d^{1:t-1}\right) = \min_{f(a_t|d^{1:t-1})} \mathsf{E}\left[l_t\left(d_t, \Theta_t\right) + \mathcal{V}_{t+1}\left(d^{1:t}\right) \middle| d^{1:t-1}\right] \tag{2.45}$$

starting with

$$\mathcal{V}_{\mathring{t}+1}\left(d^{1:\mathring{t}}\right) = 0.$$

Receding horizon

In general, the computational complexity related to evaluation of functions $\mathcal{V}_{\mathring{t}+1}(d^{1:\mathring{t}})$ grows exponentially with the length of the decision horizon $\mathring{t}$, which makes the computations practically infeasible even for moderate $\mathring{t}$. Receding horizon represents a widely used approximation method, which overcomes this difficulty. It is based on the following idea. In each time t, the approximate decision strategy $R_t \equiv \{f(a_\tau|d^{1:\tau-1})\}_{\tau=t}^{t+\mathcal{T}-1}$, where $\mathcal{T} \in \mathbb{N}$, is designed so that it minimizes the conditional expected loss

$$\mathsf{E}\left[L_t\left(d^{t:t+\mathcal{T}-1}, \Theta^{t:t+\mathcal{T}-1}\right) \middle| d^{1:t-1}\right]. \tag{2.46}$$

In (2.46), $L_t\left(d^{t:t+\mathcal{T}-1}, \Theta^{t:t+\mathcal{T}-1}\right)$ is a loss function to the receding horizon (time $t-1+\mathcal{T}$), which is selected by the decision maker. From the decision strategy R_t only the initial decision rule $f(a_t|d^{1:t-1})$ is used. In the next time $t+1$, an approximate decision strategy R_{t+1} is newly designed. For a given length of the receding horizon $\mathcal{T}$, the complexity of the design grows only linearly with the decision horizon $\mathring{t}$.

For the receding horizon to be applicable in practice, it is essential that suitable loss functions $L_t\left(d^{t:t+\mathcal{T}-1}, \Theta^{t:t+\mathcal{T}-1}\right)$ can be easily found. For example, in case of an additive loss function (2.44) the loss functions $L_t\left(d^{t:t+\mathcal{T}-1}, \Theta^{t:t+\mathcal{T}-1}\right)$, are typically selected in a form

$$L_t\left(d^{t:t+\mathcal{T}-1}, \Theta^{t:t+\mathcal{T}-1}\right) = \sum_{\tau=t}^{t+\mathcal{T}-1} l_t(d_\tau, \Theta_\tau). \tag{2.47}$$

Simplified outer model of the system

In case of the receding horizon strategy applied to models with time-invariant internals Θ, see (2.34), the computational complexity can be further reduced if the posterior pdf is "learned enough", i.e., if it is based on a large set of data records or on a strong prior information. In this case, it can be supposed that the posterior pdfs would not change much to the receding horizon and the posterior pdfs $f(\Theta|d^{1:t}), \ldots, f(\Theta|d^{1:t+T-1})$, can be approximated by the actual posterior pdf $f(\Theta|d^{1:t-1})$.

It is often technically easier to work with the observation model $f(\Delta_t|a_t, \phi_{t-1}, \Theta)$ for some fixed Θ instead of the predictive pdf $f\left(\Delta_t\middle| a_t, d^{1:t-1}\right) = \int f(\Delta_t|a_t, \phi_{t-1}, \Theta) f(\Theta|d^{1:t-1}) \mathrm{d}\Theta$. On that account, further simplification can be achieved by substituting a point estimate of the parameter to the observation model. In this case, at time t a point estimation $\hat{\Theta}$ is established from the posterior pdf $f(\Theta|d^{1:t-1})$ using a suitable loss function. The outer models $f\left(\Delta_\tau\middle| a_\tau, d^{1:\tau-1}\right)$, for $\tau \in \{t, \ldots, t+T-1\}$, are then approximated by $f\left(\Delta_\tau\middle| a_\tau, \phi_{\tau-1}, \Theta\right)|_{\Theta=\hat{\Theta}}$. This method is referred to as a certainty-equivalence strategy.

2.4 Fully Probabilistic Design

This paragraph is focused on a special design method called the fully probabilistic design (FPD) [27], [29], which can be taken as an alternative to the decision making based on the minimization of the expected loss (2.33). Although the FPD can be performed with both data and internals [30], we confine to the data-driven design in tis work.

Within the FPD, the objectives of the decision making are described by a so-called ideal pdf

$${}^I f(d^{1:\mathring{t}}) = \prod_{t=1}^{\mathring{t}} {}^I f(\Delta_t|a_t, d^{1:t-1})\, {}^I f(a_t|d^{1:t-1}),$$

which is simply interpreted as a desired distribution on the set of all trajectories. The FPD defines the optimal decision strategy ${}^{O}R$ as a randomized decision strategy which minimizes the Kullback-Leibler divergence from the ideal pdf, i.e.,

$$ {}^{O}R \in \underset{(f(a_t|d^{1:t-1}))_{t=1}^{\mathring{t}}}{\operatorname{argmin}} \mathsf{D}\left(f\left(d^{1:\mathring{t}}\right) \middle|\middle| {}^{I}f\left(d^{1:\mathring{t}}\right)\right), \tag{2.48} $$

where the joint pdf $f(d^{1:\mathring{t}}) = \prod_{t=1}^{\mathring{t}} f(\Delta_t|a_t, d^{1:t-1}) f(a_t|d^{1:t-1})$ is determined by the outer model of the system (2.30), or its approximation (typically using the certainty-equivalence approach, see Section 2.3.3), and the searched randomized decision strategy.

The main asset of the FPD is that it has an explicit solution [29]. It can be found recursively using the dynamic programing. The optimal solution of (2.48) is given by the following formulas:

$$ \begin{aligned} {}^{O}f(a_t|d^{1:t-1}) &= \frac{{}^{I}f(a_t|d^{1:t-1}) \exp(-\omega_t(a_t, d^{1:t-1}))}{\gamma_{t-1}(d^{1:t-1})} \\ \omega_t(a_t, d^{1:t-1}) &= \int f(\Delta_t|a_t, d^{1:t-1}) \ln \frac{f(\Delta_t|a_t, d^{1:t-1})}{{}^{I}f(\Delta_t|a_t, d^{1:t-1})\gamma_t(d^{1:t})} \, \mathrm{d}\Delta_t \\ \gamma_{t-1}(d^{1:t-1}) &= \int {}^{I}f(a_t|d^{1:t-1}) \exp(-\omega_t(a_t, d^{1:t-1})) \mathrm{d}a_t \end{aligned} \tag{2.49} $$

The recursion is performed backwards and starts in $\mathring{t}$ with $\gamma_{\mathring{t}}(d^{1:\mathring{t}}) = 1$.

In connection with the multiple participant decision making, an advantage of the fully probabilistic design is that it allows to establish common objectives in an easy way; see Section 4. However, it should be mentioned that the FPD suffers also from some conceptual shortcomings. They are briefly discussed in Appendix B.

Chapter 3

Towards Multiple Participant Decision Making

In the multiple participant decision making several decision makers, shortly referred to as participants, deal with, at least partially, overlapping parts of a system. Our approach to the decision making with multiple participants is determined by two basic assumptions:

- Neither participants' states of knowledge about the system nor their objectives are required to be a priori consistent in any way.
- Computational resources of individual participants are limited.

The assumption on limited computational resources forces us to search for a decision making procedure in a distributed form, i.e., all computations are to be performed by individual participants and should involve only quantities related to the parts of the system treated by particular participants.

The inconsistency of objectives and states of knowledge of individual participants causes that if the participants act independently, as if they were the only decision makers in their parts of the system, then the decision making can easily became inefficient. Typical features of such inefficiency are:

- The participants compete in the sense that each participant makes effort to draw the trajectory of the system closer to its objectives. It can cause a temporal shift to more expensive actions which, due to difference of the objectives, do not have the desired effect. As a result, the competition can lead to simultaneous increasing of losses of individual participants.
- Each participant designs its strategy using only its own incomplete knowledge about the part of the system it deals with. This knowledge could be possibly strengthen using the information provided by the participants which deal with the same part of the system. In case of individually acting participants, such information is completely ignored.

A common reason of the discussed inefficiency is that the participants reach information about objectives and knowledge of other participants only indirectly through observed data. Such an information exchange is slow and thus ineffective. It is expected that the participants employing decision making procedures which take the diversity of objectives into account and are able to exploit the knowledge provided by other participants have a potential to perform better than the individually acting participants. Our aim is to design methods which allow the participants to share the information on their objectives and knowledge about the system with other participants dealing with the same part of the system and use them to enhance the designed strategies. As we seek a distributed solution, these methods are to be applied by individual participants. However, a design of the cooperation methods is a matter of Chapters 4 and 5. In this Chapter, we focus rather on a formulation of the multiple participant decision making.

It turns out that the formulation of the problem itself, especially specification of a criterion according to which the quality of sets of participants' decision strategies could be compared, is a hard task. Although a wide class of such criterions can be considered, none of them follows directly from the original decision

tasks of the individual participants. In order to set down a unique criterion, it is necessary to state additional conditions, which are, however, inevitably tightly related to a concrete application. For a further discussion a particular set of such conditions is proposed thereinafter. It allows us to illustrate possible ways of solution of some partial problems related to multiple participant decision making and point out the bottlenecks faced.

This chapter starts with a formal description of a single participant, Section 3.1. The formalism is extended to the multiple participants in Section 3.2. In Section 3.3, we formulate basic assumptions on which our approach to the multiple participant decision making is based. Its general aspects are shortly discussed in Section 3.4 and in Section 3.5 they are further inspected for a particular form of the multiple participant decision making, namely, for the fully cooperating participants. The chapter is closed with a short summary, Section 3.6.

3.1 Single Participant

In this work the term participant means a particular implementation of a Bayesian decision maker performing a regular Bayesian decision making, generally described in Section 2.3, or a decision maker performing fully probabilistic design; see Section 2.4. As the participant represents a basic element of the multiple participant decision making, we summarize its main features here. We start with the participant performing a regular Bayesian decision making.

1. The participant deals with a sequence of random quantities (data) $d^{1:\mathring{t}} \equiv (d_1, \ldots, d_{\mathring{t}})$, where $d_t \equiv (a_t, \Delta_t)$ consist of actions a_t generated by the participant according to a randomized decision strategy $\left(f(a_t|d^{1:t-1})\right)_{t=1}^{\mathring{t}}$, and observations Δ_t. Unobservable quantities are not considered except an unknown constant parameter Θ.

2. The participant models the system by a parametric, finite-memory observation model
$$f(\Delta_t|a_t, d^{1:t-1}, \Theta) = f(\Delta_t|a_t, d^{t-T:t-1}, \Theta) = f(\Delta_t|a_t, \phi_{t-1}, \Theta),$$
where $T \in \mathbb{N}$, and ϕ_{t-1} is a subvector of $d^{t-T:t-1}$ with a given fixed structure, e.g., $\phi_{t-1} \equiv (\Delta_{t-1}, a_{t-1}, \Delta_{t-2}, \Delta_{t-3})$, for all $t \in t^*$. The observation model is considered to be time-invariant, i.e., for all $t \in \{2, \ldots, \mathring{t}\}$,
$$f(\Delta_t|a_t, \phi_{t-1}, \Theta) = f(\Delta_{t-1}|a_{t-1}, \phi_{t-2}, \Theta)|_{\Delta_{t-1}=\Delta_t, a_{t-1}=a_t, \phi_{t-2}=\phi_{t-1}}.$$

3. Initial knowledge about the unknown parameter Θ is described by a prior pdf $f(\Theta) \equiv f(\Theta|d^{1-T:0})$. Initial data $d^{1-T:0}$ needed for a construction of vectors $\phi_0, \ldots \phi_{T-1}$ are supposed to be given. For simplicity, $d^{1-T:0}$ are typically omitted in conditional pdfs.

4. Participant's objectives are described by a loss function
$$L : \left(d^{1:\mathring{t}}\right)^* \to [C, +\infty],$$
for some $C \in \mathbb{R}$.

5. Learning is performed by sequential evaluation of the posterior pdf $f(\Theta|d^{1:t})$ according to (2.35) or, more often, using an approximate technique, such as the quasi-Bayes algorithm (2.43) or a projection-based algorithm.

6. An optimal decision strategy
$$R \equiv \left(f\left(a_t\middle|\, d^{1:t-1}\right)\right)_{t=1}^{\mathring{t}}$$
is designed according to (2.33) so that it minimizes expected loss (2.25). At this point, some approximation techniques, e.g., a receding horizon, or a certainty-equivalence strategy, see Section 2.3.3, are frequently involved.

A participant performing the FPD differs from the regular Bayesian one in points (4) and (6). Namely:

- Participant's objectives are expressed by a joint ideal pdf ${}^{I}f(d^{1:\mathring{t}})$.
- An optimal decision strategy is designed by the FPD; see Section 2.4.

3.2 Multiple Participants

The multiple participant decision making arises whenever some part of the world is supposed to be influenced by more than one decision maker (participant). In other words, in the multiple participant decision making we assume a group of participants such that the sets of random quantities modelled or generated by individual participants, at least partially, overlap.

In the case of multiple participants, a part of the world which is of an interest of at least one participant is referred to as a system. A part of the world which is of an interest of the participant p is labeled as an environment of the participant p.

Before we approach to the description of the participants themselves, we introduce a notation regarding the system.

- The system is described by a sequence of data $d^{1:\mathring{t}} \equiv (d_1, \dots d_{\mathring{t}})$.
- Quantities d_t are n-dimensional random vectors, for $n \in \mathbb{N}$, with entries $d_{t;i}, i \in \{1, 2, \dots, n\}$, i.e., $d_t \equiv (d_{t;1}, d_{t;2}, \dots, d_{t;n})$.

In the rest of this section, notation related to the multiple participant decision making is introduced. Remind, that it is focused merely on a formal description. The ideas behind the multiple participant decision making are discussed thereinafter.

For $\mathring{p} \in \mathbb{N}$, let us consider $\mathring{p}$ participants, each of which is an instance of a participant generally described in Section 3.1, except that the participants need not necessarily treat the complete system. For all $p \in p^* \equiv \{1, \dots, \mathring{p}\}$, let ${}^{p}i^* \subset \{1, 2, \dots, n\}$, ${}^{p}i^* \neq \emptyset$.

Each participant p

- deals with a sequence of data ${}^{p}d^{1:\mathring{t}} \equiv ({}^{p}d_1, \dots, {}^{p}d_{\mathring{t}})$ describing its environment, where for all $t \in t^*$, ${}^{p}d_t \equiv (d_{t;i})_{i \in {}^{p}i^*}$,
- splits data ${}^{p}d_t$ into its actions ${}^{p}a_t$ and observations ${}^{p}\Delta_t$ at time t, i.e., ${}^{p}d_t \equiv ({}^{p}a_t, {}^{p}\Delta_t)$,
- models its observations by a parametric model

  $$ {}^{p}f({}^{p}\Delta_t | {}^{p}a_t, {}^{p}\phi_{t-1}, {}^{p}\Theta), $$

  where ${}^{p}\phi_{t-1}$ is, for some ${}^{p}T \in \mathbb{N}$, a subvector of ${}^{p}d^{t-{}^{p}T:t-1}$ with a fixed structure,
- has a prior pdf ${}^{p}f({}^{p}\Theta)$ of the unknown parameter ${}^{p}\Theta$,
- describes its objectives by a loss function ${}^{p}L : ({}^{p}d^{1:\mathring{t}})^* \to [{}^{p}C, +\infty]$, ${}^{p}C \in \mathbb{R}$, or by an ideal pdf ${}^{pI}f({}^{p}d^{1:\mathring{t}})$ defined on its data ${}^{p}d^{1:\mathring{t}}$, in dependence on which approach to the design of a decision strategy it employs; see Section 3.1,
- evaluates (approximately) its posterior pdf ${}^{p}f({}^{p}\Theta | {}^{p}d^{1:t})$,
- applies its decision strategy

  $$ {}^{p}R \equiv \left({}^{p}f({}^{p}a_t | {}^{p}d^{1:t-1}) \right)_{t \in t^*} $$

  designed by the regular Bayesian decision making or FPD using its loss function ${}^{p}L\left({}^{p}d^{1:\mathring{t}}\right)$ or its ideal pdf ${}^{pI}f\left({}^{p}d^{1:\mathring{t}}\right)$, respectively.

According to the relation between the environments of the individual participants and the system, it holds that $\bigcup_{p \in p^*} {}^{p}i^* = \{1, 2, \dots, n\}$. For $p_1, p_2 \in p^*$, such that $p_1 \neq p_2$ and ${}^{p_1}i^* \cap {}^{p_2}i^* \neq \emptyset$, a participant p_2 is called a neighbour of the participant p_1. The pair of participants p_1 and p_2 is shortly referred to as neighbours. Data, at time t, in common of the neighbours p_1, p_2 are denoted ${}^{p_1,p_2}d_t$, i.e.,

$$ {}^{p_1,p_2}d_t \equiv (d_{t;i})_{i \in {}^{p_1}i^* \cap {}^{p_2}i^*}. \tag{3.1} $$

In what follows, we utilize a distinction of decision making tasks according to their domain: decision making tasks solved by individual participants, i.e., those defined by models ${}^{p}f({}^{p}\Delta_t|\,{}^{p}a_t,\,{}^{p}\phi_{t-1},\,{}^{p}\Theta)$, priors ${}^{p}f({}^{p}\Theta)$, and loss functions ${}^{p}L\left({}^{p}d^{1:\mathring{t}}\right)$ or ideal pdfs ${}^{p}If\left({}^{p}d^{1:\mathring{t}}\right)$ on data ${}^{p}d^{1:\mathring{t}}$, are referred to as local (decision making) tasks. Analogously, a decision making task defined on the system is referred to as global (decision making) task. For convenience, the terms local and global task could be used also in a more general sense, i.e., not only for Bayesian or FPD decision making tasks.

3.3 Basic Assumptions

In this section, elementary assumptions on which our approach to the multiple participant decision making is based are formulated. The assumptions are stated in a rather vague form as their purpose is just to outline what kind of decision making problems is aimed at.

1. First of all, we suppose that there is at least one pair of neighbours, i.e., there exist distinct $p_1, p_2 \in p^*$, such that ${}^{p_1}i^* \cap {}^{p_2}i^* \neq \emptyset$. This assumption can be replaced by a more strict one, which ensures that the multiple participant decision task cannot be trivially split into two independent multiple participant decision tasks: for all $i_1^*, i_2^* \subset \{1,\ldots,n\}$, such that $i_1^* \neq \emptyset, i_2^* \neq \emptyset$, and $i_1^* \cup i_2^* = \{1,\ldots,n\}$, there exists $p \in p^*$ so that ${}^{p}i^* \cap i_1^* \neq \emptyset$ and ${}^{p}i^* \cap i_2^* \neq \emptyset$.

   Both the above stated assumptions are reasonable for practical applications, nevertheless, for the problem formulation they are not necessary. Thus, regarding sets ${}^{p}i^*$, it is sufficient to suppose only that $\forall p \in p^*$, ${}^{p}i^* \neq \emptyset$. Note that it is impossible for any two participants to have in their common a random quantity which represents actions of both of them. However, an action of a participant may stand as an observation of another participant.

2. No restrictions are put on the parametric models of individual participants as well as on their prior pdfs. It means that different participants may model a part of the system in their common by different models, parameterized by different parameters, and with different structures of state vectors. Predictive pdfs $\int {}^{p}f({}^{p}\Delta_t|\,{}^{p}a_t,\,{}^{p}\phi_{t-1},\,{}^{p}\Theta)\,{}^{p}f\left({}^{p}\Theta\middle|\,{}^{p}d^{1:t-1}\right)\mathrm{d}\,{}^{p}\Theta$ may also differ in their marginal pdfs on common parts of the system.

3. No consistency of objectives (loss functions or ideal pdfs) of individual participants is required: objectives of different participants regarding a part of the system in their common may arbitrarily differ.

4. Computational resources of the participants are limited. This crucial assumption, fully respecting reality, leads to the same constraints as in case of a single participant: parametric models, decision strategies prior/posterior pdfs, loss functions, and ideal pdfs must be from given fixed classes and parameterized by finite-dimensional parameters. Furthermore, the limited computational resources restrict an extent of a decision making tasks that can be solved by a single decision maker. This forces us to seek the solution in a distributed form. Namely, we assume that the decision strategies of the individual participants are completely designed by the participants themselves, whereas each participant p deals only with its environment, i.e., with data ${}^{p}d^{1:\mathring{t}}$. Nevertheless, for a formulation of the distributed multiple participant decision making it turns out to be necessary to consider a virtual global decision maker dealing with the complete system.

5. To reach a desired effect of the multiple participant decision making, the participants are to be endowed with mechanisms for cooperation with their neighbours.

6. A normative theory of the multiple participant decision making is aimed at. In other words, we want to avoid descriptive approaches modelling decision making within groups of real beings. Of course, there is no clear line between normative and descriptive approach, see [47], however, we want at least to attempt to formulate some kind of rational basis for the multiple participant decision making.

7. For simplicity, it is assumed that

- all participants observe the system and apply their actions simultaneously,
- all participants have the same decision horizon $\mathring{t}$,
- data structures of individual participants, i.e., sets ${}^{p}i^{*}$, are fixed,
- all participants perform regular Bayesian decision making or all participants perform the FPD.

Although these assumptions may be restrictive in practical applications, they are not related to the essence of the multiple participant decision making.

3.4 Problem Statement

As stated in the introduction of this chapter, the objective of our approach to the multiple participant decision making is to provide methods for cooperation of the participants which enable them to design decision strategies that are "better" than strategies designed by individually acting participants. The enhanced strategies are to be designed on conditions stated in Section 3.3, especially respecting limited computational abilities of the participants, i.e., the desired methods are to be applied by the participants within their environments only. To complete a rough formulation of the problem it is necessary to specify what is meant by "better" decision strategies. In other words, we need to establish a partial order on $\mathring{p}$-tuples of decision strategies designed by the individual participants. The desired partial order represents some kind of a global task. It should be stressed that if we are going to enhance the overall behaviour of the system in any sense, such a global task must always exist.

In case of a single participant, a unique preference order on data and a unique assessment of uncertainty are considered. To extend the preference order to all decision strategies it just remains to select a suitable method for treating the uncertainty. In case of multiple participants we are facing a more complex problem - we have $\mathring{p}$ preference orders on data and $\mathring{p}$ descriptions of uncertainty. Both of them are specified only partially, i.e., on the environments of the individual participants, and they need not be consistent in any way. Moreover, although the individual participants are Bayesian decision makers, it does not automatically imply that the global task must be based on the same principles as the local ones.

Obviously, there is a wide range of possible formulations of the multiple participant decision making depending especially on

- the approach to multiple objectives,
- treating the uncertainty (in general),
- the interpretation of ambiguity in assessments of uncertainty.

Any specific formulation cannot be derived solely from the attributes of individual participants – their loss functions, parametric models, and prior pdfs. In any case the formulation depends on a particular application, i.e., on desired properties of the solution and additional assumptions, which can be stated under given conditions.

A systematic classification of possible formulations of the multiple participant decision making is out of the scope of this work. Instead of it, we attempt to outline a formulation of a single specific case – fully cooperating participants with independent sources of prior information. This particular case serves as a basis for a further discussion on methods for cooperation of participants.

Notes:

- The multiplicity of objectives makes the formulation of the multiple participant decision making close to fields dealing with multi-objective optimization. However, known results, e.g., [34] cannot be used directly as the discussed problem is somewhat more complex due to multiple, mutually inconsistent, uncertainty assessments.
- Remind, that at this stage we are dealing only with a formulation of the problem and not with its solution. To design a distributed solution, or at least its approximation, directly by some segmentation of the global task is a hard problem indeed. However, the global task represents a unique criterion by which the $\mathring{p}$-tuples of decision strategies can be compared. Having such a criterion, we

may design various ad-hoc distributed "solutions" directly and evaluate their performance ex-post, at least in an experimental way.

- Due to the insufficiencies of the FPD discussed in Appendix B, we focus in the rest of this chapter on participants performing regular Bayesian decision making.

3.5 Fully Cooperating Participants

In this section, we attempt to outline a way in which a global task for fully cooperating participants with independent sources of prior information could be formulated. As it is seen below, even in this relatively simple case, the formulation is a hard task. There arise many problems which either require further inspection or must be eliminated by additional assumptions at the cost of restrictions on the class of possible applications.

At this point, it is convenient to make in advance an assumption on the general form of the global task: The global task is formulated as a Bayesian one in the sense of Section 2.3. It should be stressed, that this is a technical assumption only and as a necessity it does not follow from anything stated up to know. However, the assumption can be, at least partially, justified in the following way: All the participants are Bayesian decision makers. It means, that their decision strategies are designed according to principles which ensure certain kind of rationality of the decision making [25], [10]. It seems to be natural to require these principles to be satisfied also in the global task.

The Bayesian nature of the global task implies that the global objectives and global uncertainty can be treated separately. This fact significantly simplifies the formulation.

The desired features of fully cooperating participants are:

- The global objective is, in some sense, a compromise among objectives of individual participants. Objectives of any participant are not to be preferred to objectives of other participants.
- For a global uncertainty assessment all prior knowledge provided by individual participants is exploited. At the same time, the added parasitic prior information should be minimal. By the parasitic information we mean a prior information for which there is no evidence.
- The participants are supposed to act in a fair way – they are willing to provide all their knowledge and information on their objectives.

3.5.1 Form of the Global Task

The purpose of the global task is to provide a partial order on $\mathring{p}$-tuples of decision strategies ${}^{p}f\left({}^{p}a_t\middle|\,{}^{p}d^{1:t-1}\right)$ of the individual participants. Assuming a fixed vector of past data $d^{1:t-1}$, each participant p generates its action ${}^{p}a_t$ according to its decision strategy independently of actions generated by other participants. Furthermore, action ${}^{p}a_t$ is generated irrespectively of data ${}^{\tilde{p}}d^{1:t-1}$ unknown to the participant p. As a result, global decision strategies $\left(f\left(a_t\middle|\,d^{1:t-1}\right)\right)_{t\in t^*}$ which can be represented by $\mathring{p}$-tuples of decision strategies of the individual participants $\left({}^{p}f\left({}^{p}a_t\middle|\,{}^{p}d^{1:t-1}\right)\right)_{t\in t^*}$ restrict to those which exhibit the above indicated conditional independences, i.e., to those with decision rules in a form

$$f\left(a_t\middle|\,d^{1:t-1}\right)=\prod_{p=1}^{\mathring{p}}f\left({}^{p}a_t\middle|\,{}^{p}d^{1:t-1}\right). \tag{3.2}$$

$a_t\equiv\left({}^{1}a_t,\ldots,{}^{\mathring{p}}a_t\right)$ in (3.2) denote global actions, i.e., random vectors which consist of actions of individual participants.

Let us denote the set of all global decision strategies which can be represented by $\mathring{p}$-tuples of local decision strategies $\bar{R}^*$. To acquire an order on $\mathring{p}$-tuples of decision strategies of individual participants it is then sufficient to establish an order on the set $\bar{R}^*$. However, it turns out to be convenient to formulate the global task as a more general one – without the restriction of decision strategies to the set $\bar{R}^*$. Then, the formulation of the global task consists in construction of a suitable global loss function $L(d^{1:\mathring{t}})$ characterizing the global preference order on data and probabilities on the system expressing uncertainty

quantification in dependence on the global decision strategy. The probabilities are represented by a system of pdfs $\left(f_R(d^{1:\mathring{t}})\right)_{R\in R^*}$, where R^* is a set of all global decision strategies. This approach not only makes the formulation easier, but, among others, it allows to evaluate limitations of the distributed solution. For example, the difference

$$\min_{\bar{R}\in\bar{R}^*} \mathsf{E}_{\bar{R}}\left[L(d^{1:\mathring{t}})\right] - \min_{R\in R^*} \mathsf{E}_R\left[L(d^{1:\mathring{t}})\right],$$

where $\mathsf{E}_{\bar{R}}[\cdot]$ and $\mathsf{E}_R[\cdot]$ denote expectations with respect to $f_{\bar{R}}(d^{1:t})$ and $f_R(d^{1:t})$ respectively, expresses an unavoidable drop of quality of the decision making caused by employing local decision strategies instead of the global one.

3.5.2 Uncertainty Description

Since the global task is assumed to be a Bayesian one, the uncertainty is to be quantified by a joint pdf on the system $f_R\left(d^{1:\mathring{t}}\right)$ depending on a randomized strategy

$$R \equiv \left(f\left(a_t\middle|\, d^{1:t-1}\right)\right)_{t=1}^{\mathring{t}}.$$

We assume that $f_R\left(d^{1:\mathring{t}}\right)$ can be expressed similarly as in case of individual participants. Namely,

$$\begin{aligned} f_R\left(d^{1:\mathring{t}}\right) &= \int \prod_{t=1}^{\mathring{t}} f\left(\Delta_t\middle|\, a_t, d^{1:t-1}, \Theta\right) f\left(a_t\middle|\, d^{1:t-1}, \Theta\right) f(\Theta)\,\mathrm{d}\Theta = \\ &= \int \prod_{t=1}^{\mathring{t}} f\left(\Delta_t\middle|\, a_t, \phi_{t-1}, \Theta\right) f\left(a_t\middle|\, d^{1:t-1}\right) f(\Theta)\,\mathrm{d}\Theta. \end{aligned} \tag{3.3}$$

In (3.3):

- Δ_t denote global observations, i.e., random vectors consisting of all entries of d_t, which are not actions of any participant. Remind, that the global observations Δ_t need not consist of all entries of all observations ${}^p\Delta_t$ of individual participants, because observations of a participant may contain actions of another one,
- $f\left(\Delta_t\middle|\, a_t, \phi_{t-1}, \Theta\right)$ is a time-invariant global parametric model with a state vector ϕ_{t-1} of a fixed structure,
- pdfs $f\left(a_t\middle|\, d^{1:t-1}\right)$ represent a global decision strategy,
- pdf $f(\Theta)$ is a global prior pdf.

Again, the existence of a suitable time-invariant parametric model $f\left(\Delta_t\middle|\, a_t, \phi_{t-1}, \Theta\right)$ and a prior pdf $f(\Theta)$ is rather a technical assumption at this stage. Conditional independence of a_t and Θ given $d^{1:t-1}$ is implied by the natural conditions of decision making (2.23), which are assumed to be valid for all participants.

In what follows, we focus on constructing of a suitable global parametric model and a global prior pdf from local parametric models and local prior pdfs so that they reflect general requirements for the fully cooperating participants stated in Section 3.5. Namely, all available prior knowledge provided by individual participants is to be exploited, while supposing that these pieces of information are not intentionally modified. Note, that the following parts represent just a sketch of possible a approach. On that account it is formulated in a rather vague way and the technical details are not to be taken too exactly. For convenience, we also assume that the parameters Θ and ${}^p\Theta$ are individual pdfs themselves. The parametric models are then represented directly by the sets Θ^* and ${}^p\Theta^*$.

Global Parametric Model

According to the general assumptions stated in Section 3.3, the parametric models of individual participants need not be consistent in the sense that any two neighbours may model a part of the system in their common by arbitrarily different parametric models. For further progression, it is worth to emphasize the following, often contradicting, aspects of a choice of a parametric model.

- Parametric model ${}^{p}\Theta^{*}$ represents a strong prior information assigning zero probability to models out of ${}^{p}\Theta^{*}$.
- Parametric model must respect limited computational abilities of a participant.

Ideally, the choice of a parametric model should be supported by a prior knowledge based, e.g., on a theoretical analysis of a particular system, or eventually should reflect a lack of a prior knowledge. However, it is often practically impossible because of the related computational complexity. Then, the parametric model must be selected from a class of parametric models for which feasible algorithms are available. As a result, the choice of a parametric model introduces into the decision making task a parasitic prior information.

The below proposed approach to the construction of a global parametric model comes out of distinction of the sources of inconsistency of local parametric models. The basic idea is as follows: if the inconsistency of the local parametric models is completely caused by computational restrictions of individual participants, i.e., not by a prior knowledge, there is not any real contradiction among them. The global parametric model is then to be selected so that it, in some sense, contains all local parametric models. In other words, the global parametric model should be able to model relations among quantities which could be modelled by any one of the local parametric models. The amount of parasitic prior information added in this stage should be as small as possible.

The construction of a global parametric model becomes more complex if the inconsistency is caused by different prior information. The reason is that, in dependence on a particular parametric model, such information can be stronger than any evidence represented by an arbitrarily large, but finite, number of observations. If there appears this kind of inconsistency in prior information pieces, the above outlined approach can not be applied. Instead, more detailed models of the inconsistency must be employed.

In what follows we assume that the inconsistency is solely caused by the technical limitations and thus we are able, at least theoretically, to select a suitable global parametric model.

Notes:

- For a particular participant p, let us split global observations, actions, and state vectors to quantities considered by the p-th participant and the remaining ones: $\Delta \equiv ({}^{p}\Delta, \bar{}^{p}\Delta), a \equiv ({}^{p}a, \bar{}^{p}a)$, and $\phi \equiv ({}^{p}\phi, \bar{}^{p}\phi)$. Time indices are omitted for simplicity. Then, the global parametric model induces a parametric model of observations ${}^{p}\Delta$ depending on the global actions a and state vector ϕ as a set of marginal pdfs

$$f\left({}^{p}\Delta \middle| a, \phi, \Theta\right), \Theta \in \Theta^{*}. \tag{3.4}$$

  Furthermore, because the participant p do not care about quantities $\bar{}^{p}a$ and $\bar{}^{p}\phi$, its observations ${}^{p}\Delta$ can be considered to be conditionally independent of $({}^{\bar{p}}a, {}^{\bar{p}}\phi)$ given $({}^{p}a, {}^{p}\phi)$, for all ${}^{p}\Theta \in {}^{p}\Theta^{*}$, within the model used by this participant. Thus, its parametric model ${}^{p}f\left({}^{p}\Delta \middle| {}^{p}a, {}^{p}\phi, {}^{p}\Theta\right), {}^{p}\Theta \in {}^{p}\Theta^{*}$ can be equivalently substituted by a parametric model

$$ {}^{p}f\left({}^{p}\Delta \middle| a, \phi, {}^{p}\Theta\right), {}^{p}\Theta \in {}^{p}\Theta^{*}, \tag{3.5}$$

  where

$${}^{p}f\left({}^{p}\Delta \middle| a, \phi, {}^{p}\Theta\right) = {}^{p}f\left({}^{p}\Delta \middle| {}^{p}a, \bar{}^{p}a, {}^{p}\phi, \bar{}^{p}\phi, {}^{p}\Theta\right) \equiv {}^{p}f\left({}^{p}\Delta \middle| {}^{p}a, {}^{p}\phi, {}^{p}\Theta\right)$$

  for all ${}^{p}\Delta \in {}^{p}\Delta^{*}$, ${}^{p}a \in {}^{p}a^{*}$, $\bar{}^{p}a \in \bar{}^{p}a^{*}$, ${}^{p}\phi \in {}^{p}\phi^{*}$ and $\bar{}^{p}\phi \in \bar{}^{p}\phi^{*}$. This direction seems to a be suitable way to define what it means that a global parametric model includes a local one. Namely, the global parametric model Θ^{*} can be considered to include the local model ${}^{p}\Theta^{*}$ if for any ${}^{p}\Theta \in {}^{p}\Theta^{*}$ there is some $\Theta \in \Theta^{*}$ such that

$${}^{p}f\left({}^{p}\Delta \middle| a, \phi, \ {}^{p}\Theta\right) = f\left({}^{p}\Delta \middle| a, \phi, \Theta\right)$$

for all ${}^{p}\Delta \in {}^{p}\Delta^{*}, a \in {}^{p}a^{*}$, and $\phi \in \phi^{*}$, where ${}^{p}f\left({}^{p}\Delta \middle| a, \phi, {}^{p}\Theta\right)$ and $f\left({}^{p}\Delta \middle| a, \phi, \Theta\right)$ are "extended" local model (3.5) and "restricted" global model (3.4), respectively.

- The amount of a parasitic prior information introduced to the decision making should be as small as possible. In case of a parametric model, such information is represented especially by causeless exclusion of some Θ from Θ^{*}. To illustrate it, consider a parametric model Θ^{*} of two random quantities Δ_1, Δ_2. Assume that there is an evidence which indicates that pdfs $f\left(\Delta_1 \middle| \Theta_1\right)$ for $\Theta_1 \in \Theta_1^{*}$ and pdfs $f\left(\Delta_2 \middle| \Delta_1, \Theta_{2|1}\right)$ for $\Theta_{2|1} \in \Theta_{2|1}^{*}$ are to be considered for modelling of Δ_1 and the dependence of Δ_2 on Δ_1, respectively. Then the parametric model $f\left(\Delta_1, \Delta_2 \middle| \Theta\right), \Theta \in \Theta^{*}$, should contain all marginal pdfs in Θ_1^{*} and all conditional pdfs in $\Theta_{2|1}^{*}$, in the sense that Θ^{*} must satisfy

$$\forall \Theta_1 \in \Theta_1^{*}, \exists \Theta \in \Theta^{*}, f\left(\Delta_1 \middle| \Theta_1\right) = f\left(\Delta_1 \middle| \Theta\right) \tag{3.6}$$

and

$$\forall \Theta_{2|1} \in \Theta_{2|1}^{*}, \exists \Theta \in \Theta^{*}, f\left(\Delta_2 \middle| \Delta_1, \Theta_{2|1}\right) = f\left(\Delta_2 \middle| \Delta_1, \Theta\right). \tag{3.7}$$

Moreover, if no other evidence is available, then the parametric model $f\left(\Delta_1, \Delta_2 \middle| \Theta\right), \Theta \in \Theta^{*}$, should contain all combinations of marginal pdfs $f\left(\Delta_1 \middle| \Theta_1\right)$ and conditional pdfs $f\left(\Delta_2 \middle| \Delta_1, \Theta_{2|1}\right)$, i.e., it should hold

$$\forall \Theta_1 \in \Theta_1^{*}, \forall \Theta_{2|1} \in \Theta_{2|1}^{*}, \exists \Theta \in \Theta^{*}, f\left(\Delta_1, \Delta_2 \middle| \Theta\right) = f\left(\Delta_1 \middle| \Theta_1\right) f\left(\Delta_2 \middle| \Delta_1, \Theta_{2|1}\right). \tag{3.8}$$

If Θ^{*} satisfy (3.6) and (3.7) but does not satisfy (3.8) then it is impossible to define prior pdf $f\left(\Theta\right)$ so that marginal pdf of Δ_1 and conditional pdf of Δ_2 given Δ_1 are independent. Recall, that pdfs are supposed to coincide with the parameters. In this case it may easily happen that, e.g., an observation of Δ_1 has an impact on the posterior pdf of conditional pdf of Δ_2 given Δ_1. Such a property is undesirable if there is no prior evidence on relation of marginal pdfs of Δ_1 and conditional pdfs of Δ_2 given Δ_1

In case of the multiple participant decision making, the local parametric models provide information on marginal and conditional pdfs which are to be included in the global parametric model. The above discussion indicates that in order to minimize an impact of a parasitic information, the global parametric model should be required to include not only the local parametric models but also some kind of their combinations. Nevertheless, it is out of the scope of this work to bring any particular procedure for constructing the global parametric model, or, at least, well formulated conditions which are to be satisfied by the global model.

Global Prior Pdf

In the introduction of Section 3.5 we have made an assumption that the prior information of individual participants come from independent sources. Then, the pieces of prior information can be taken as complementing ones. This leads to the idea of construction of the global prior pdf: the pieces of participants' prior information expressed in a common suitable form are to be taken similarly as independent blocks of observations and used for evaluation of the global prior pdf via procedure analogous to evaluation of a posterior pdf. Of course, this is a hard task in general. However, in a special, yet practically important, case in which the individual prior pdfs correspond to some posterior ones a further progression is somewhat easier. Nevertheless, as it seen below, even in this relatively simple case a general solution is not straightforward at all. In order to avoid technical difficulties, which are not substantial for the problem addressed, we assume that all data quantities are discrete. Pdfs of data quantities are then to be taken with respect to an underlying counting measure.

We discuss the addressed problem in two steps. In both of them, the prior pdfs ${}^{p}f\left({}^{p}\Theta\right)$ of individual participants are supposed to correspond to posterior ones for some sequences of, possibly virtual, observations with respect to some prior pdfs ${}^{p}f_0\left({}^{p}\Theta\right)$, which are agreed to represent complete lack of knowledge. These ancillary prior pdfs we refer to as pre-prior pdfs in order to distinguish them from common prior pdfs.

First, we assume that regression vectors of the parametric models of the individual participants are empty. Then, the local prior pdfs can be expressed in a form

$$ {}^{p}f\left({}^{p}\Theta\right) \propto {}^{p}f_0\left({}^{p}\Theta\right) \prod_{\tau \in {}^{p}\mathcal{T}^*} {}^{p}f\left({}^{p}\Delta_\tau \middle| {}^{p}\Theta\right), \tag{3.9} $$

for some sequences of observations $\left({}^{p}\Delta_\tau\right)_{\tau \in {}^{p}\mathcal{T}^*}$, where the sets ${}^{p}\mathcal{T}^*$ are supposed to be pairwise disjoint.

For fixed pre-prior pdfs ${}^{p}f_0({}^{p}\Theta)$, the prior pdfs ${}^{p}f\left({}^{p}\Theta\right)$, taken as functions of $\left({}^{p}\Delta_\tau\right)_{\tau \in {}^{p}\mathcal{T}^*}$, can be seen as statistics in a Bahadur sense, i.e., mappings to arbitrary target spaces, see [46]. It is obvious that, except for very special cases, the sequences $\left({}^{p}\Delta_\tau\right)_{\tau \in {}^{p}\mathcal{T}^*}$ for which (3.9) holds are not determined uniquely. Instead, ${}^{p}f\left({}^{p}\Theta\right)$ determine sets ${}^{p}A \subset \left({}^{p}\Delta^*\right)^{\left|{}^{p}\mathcal{T}^*\right|}$, where $|\cdot|$ denotes cardinality of a set, of all data sequences $\left({}^{p}\Delta_\tau\right)_{\tau \in {}^{p}\mathcal{T}^*}$ for which (3.9) is satisfied:

$$ {}^{p}A \equiv \left\{ \left({}^{p}\Delta_\tau\right)_{\tau \in {}^{p}\mathcal{T}^*} \in \left({}^{p}\Delta^*\right)^{\left|{}^{p}\mathcal{T}^*\right|} \middle| \; {}^{p}f\left({}^{p}\Theta\right) \propto {}^{p}f_0\left({}^{p}\Theta\right) \prod_{\tau \in {}^{p}\mathcal{T}^*} {}^{p}f\left({}^{p}\Delta_\tau \middle| {}^{p}\Theta\right) \right\}. $$

For the sets ${}^{p}A$, it holds

$$ {}^{p}f\left({}^{p}\Theta\right) \propto {}^{p}f_0\left({}^{p}\Theta\right) {}^{p}P\left(\left({}^{p}\Delta_\tau\right)_{\tau \in {}^{p}\mathcal{T}^*} \in {}^{p}A \middle| {}^{p}\Theta\right), \tag{3.10} $$

where ${}^{p}P(\cdot|\,{}^{p}\Theta)$ denotes the probability with the pdf ${}^{p}f\left(\cdot\middle|\,{}^{p}\Theta\right)$, i.e.,

$$ {}^{p}P\left(\left({}^{p}\Delta_\tau\right)_{\tau \in {}^{p}\mathcal{T}^*} \in {}^{p}A \middle| {}^{p}\Theta\right) = \sum_{\left({}^{p}\Delta_\tau\right)_{\tau \in {}^{p}\mathcal{T}^*} \in {}^{p}A} {}^{p}f\left(\left({}^{p}\Delta_\tau\right)_{\tau \in {}^{p}\mathcal{T}^*} \middle|\, {}^{p}\Theta\right). $$

In order to express the knowledge represented by the sets ${}^{p}A$ in a unified way, we establish sets ${}^{p}B \subset \left(\Delta^*\right)^{\left|{}^{p}\mathcal{T}^*\right|}$ carrying the same information. For that purpose, we utilize projections ${}^{p}S : d^* \to {}^{p}d^*$. Using the notation introduced in Section 3.2, the projection ${}^{p}S$ is defined for all $d_t = (d_{t;i})_{i=1}^{n} \in d^*$ by

$$ {}^{p}S(d_t) = \left(d_{t;i}\right)_{i \in {}^{p}i^*}. \tag{3.11} $$

Roughly speaking, the sets ${}^{p}B$ consist of all data sequences $\left(\Delta_\tau\right)_{\tau \in {}^{p}\mathcal{T}^*} \in \left(\Delta^*\right)^{\left|{}^{p}\mathcal{T}^*\right|}$ such that subvectors of their elements corresponding to data ${}^{p}d_\tau$ fulfill (3.9). More precisely,

$$ {}^{p}B \equiv \left\{ \left(\Delta_\tau\right)_{\tau \in {}^{p}\mathcal{T}^*} \in \left(\Delta^*\right)^{\left|{}^{p}\mathcal{T}^*\right|} \middle| \left({}^{p}S\left(\Delta_\tau\right)\right)_{\tau \in {}^{p}\mathcal{T}^*} \in {}^{p}A \right\}. $$

It is easy to verify that for any global parametric model it holds

$$ P\left(\left(\Delta_\tau\right)_{\tau \in {}^{p}\mathcal{T}^*} \in {}^{p}B \middle|\, \Theta\right) = P\left(\left({}^{p}S\left(\Delta_\tau\right)\right)_{\tau \in {}^{p}\mathcal{T}^*} \in {}^{p}A \middle|\, \Theta\right), $$

for all $\Theta \in \Theta^*$. The sets ${}^{p}B$ represent the pieces of prior knowledge corresponding to prior pdfs ${}^{p}f\left({}^{p}\Theta\right)$. For a suitable global pre-prior pdf $f_0(\Theta)$, the global prior pdf based on the prior information pieces provided by all participants can be then expressed as

$$ f\left(\Theta\right) \propto f_0(\Theta) P\left(\left(\Delta_\tau\right)_{\tau \in {}^{1}\mathcal{T}^*} \in {}^{1}B, \ldots, \left(\Delta_\tau\right)_{\tau \in {}^{\mathring{p}}\mathcal{T}^*} \in {}^{\mathring{p}}B | \Theta\right). \tag{3.12} $$

In case that all participants deal with the complete system and all local parametric models as well as the global one have the same sufficient statistics, the evaluation of the global prior pdf (3.12) reduces to a simple application of the Bayes rule

$$ f\left(\Theta\right) \propto f_0(\Theta) f\left(\left(\Delta_\tau\right)_{\tau \in \cup_{p \in {}^{p}\mathcal{P}^*} {}^{p}\mathcal{T}^*} \middle| \Theta\right), $$

where $\left(\Delta_\tau\right)_{\tau \in {}^{\mathring{p}}\mathcal{T}^*}$ are arbitrary sequences of observations such that (3.9) holds. In a general case, the evaluation of (3.12) is a hard task due to the involved "set observations" ${}^{p}B$. Its practical applicability is thus conditioned by availability of suitable procedures for approximation of the resulting pdfs.

Now, assume that the local parametric models have nonempty regression vectors. Particularly, we consider parametric models ${}^{p}f\left({}^{p}\Delta_t\middle|\,{}^{p}a_t,\,{}^{p}\Theta\right)$. Similarly as in the preceding case, we assume that the local prior pdfs can be expressed as posterior ones with suitable pre-prior pdfs

$$ {}^{p}f\left({}^{p}\Theta\right) \propto f_0({}^{p}\Theta) \prod_{\tau\in {}^{p}\mathcal{T}^*} {}^{p}f\left({}^{p}\Delta_\tau\middle|\,{}^{p}a_\tau,\,{}^{p}\Theta\right), \tag{3.13} $$

for some data sequences $\left({}^{p}d_\tau\right)_{\tau\in {}^{p}\mathcal{T}^*} \equiv \left({}^{p}\Delta_\tau,\,{}^{p}a_\tau\right)_{\tau\in {}^{p}\mathcal{T}^*}$ and pairwise disjoint sets ${}^{p}\mathcal{T}^*$. Again, we can construct sets

$$ {}^{p}A \equiv \left\{ \left({}^{p}d_\tau\right)_{\tau\in {}^{p}\mathcal{T}^*} \in \left({}^{p}d^*\right)^{\left|{}^{p}\mathcal{T}^*\right|} \middle|\; {}^{p}f\left({}^{p}\Theta\right) \propto {}^{p}f_0\left({}^{p}\Theta\right) \prod_{\tau\in {}^{p}\mathcal{T}^*} {}^{p}f\left({}^{p}\Delta_\tau\middle|\,{}^{p}a_\tau,\,{}^{p}\Theta\right) \right\} $$

of all data sequences from $\left({}^{p}d^*\right)^{\left|{}^{p}\mathcal{T}^*\right|}$ for which (3.13) holds and corresponding sets

$$ {}^{p}B \equiv \left\{ \left(d_\tau\right)_{\tau\in {}^{p}\mathcal{T}^*} \in \left(d^*\right)^{\left|{}^{p}\mathcal{T}^*\right|} \middle|\; \left({}^{p}S\left(d_\tau\right)\right)_{\tau\in {}^{p}\mathcal{T}^*} \in {}^{p}A \right\} $$

of data sequences from $\left(d^*\right)^{\left|{}^{p}\mathcal{T}^*\right|}$. However, in this case the information $\left(d_\tau\right)_{\tau\in {}^{p}\mathcal{T}^*} \in {}^{p}B$ cannot be used for evaluation of the global prior pdf analogous to (3.12), except for a very special case in which for all sequences $\left(\Delta_\tau, a_\tau\right)_{\tau\in {}^{p}\mathcal{T}^*}, \left(\tilde{\Delta}_\tau, \tilde{a}_\tau\right)_{\tau\in {}^{p}\mathcal{T}^*} \in {}^{p}B$ it holds $a_\tau = \tilde{a}_\tau$ for all $\tau \in {}^{p}\mathcal{T}^*$.

The information $\left(d_\tau\right)_{\tau\in {}^{p}\mathcal{T}^*} \in {}^{p}B$ could be exploited for construction of the global prior pdf in case that the global parametric model $f\left(\Delta_t\middle|\,a_t,\Theta\right)$ can be suitably extended to a joint pdf $f\left(\Delta_t, a_t\middle|\,\Theta\right)$, i.e., if an appropriate pdf $f\left(a_t\middle|\,\Theta\right)$ is available. It is, e.g., if the participants are known to generate their actions ${}^{p}a_t$ according to time-invariant strategies ${}^{p}f\left({}^{p}a_t\right)$ independently of past data. In such case, the suitable $f\left(a_t\middle|\,\Theta\right)$ can be selected as $\prod_{p\in p^*} {}^{p}f\left({}^{p}a_t\right)$, for all $\Theta \in \Theta^*$. The extended global parametric model is then

$$ f\left(\Delta_t, a_t\middle|\,\Theta\right) = f\left(\Delta_t\middle|\,a_t,\Theta\right) \prod_{p\in p^*} {}^{p}f\left({}^{p}a_t\right), $$

and the global prior pdf has a form analogous to (3.12)

$$ f\left(\Theta\right) \propto f_0(\Theta) P\left(\left(d_\tau\right)_{\tau\in {}^{1}\mathcal{T}^*} \in {}^{1}B, \ldots, \left(d_\tau\right)_{\tau\in {}^{\mathring{p}}\mathcal{T}^*} \in {}^{\mathring{p}}B \middle|\Theta\right). $$

In practice the participants apply strategies varying in time and depending on all past data in general. Knowing these local decision strategies ${}^{p}f\left({}^{p}a_t\middle|\,{}^{p}d^{1:t-1}\right)$ we can extend the global parametric model $f\left(\Delta_t\middle|\,a_t,\Theta\right) = f\left(\Delta_t\middle|\,a_t,d^{1:t-1},\Theta\right)$ by

$$ f\left(a_t\middle|\,d^{1:t-1},\Theta\right) = \prod_{p\in p^*} {}^{p}f\left({}^{p}a_t\middle|\,{}^{p}d^{1:t-1}\right). $$

However, due to the dependence of the resulting extended global parametric model

$$ f\left(\Delta_t, a_t\middle|\,d^{1:t-1},\Theta\right) = f\left(\Delta_t\middle|\,a_t,\Theta\right) \prod_{p\in p^*} {}^{p}f\left({}^{p}a_t\middle|\,{}^{p}d^{1:t-1}\right) $$

on past data $d^{1:t-1}$ it cannot be used for evaluation of the global prior pdf analogously to (3.12). Another difficulty is that the local decision strategies ${}^{p}f\left({}^{p}a_t\middle|\,{}^{p}d^{1:t-1}\right)$ vary in time, whereas the virtual data $\left({}^{p}d_\tau\right)_{\tau\in {}^{p}\mathcal{T}^*}$ are not related to particular moments. $\left({}^{p}d_\tau\right)_{\tau\in {}^{p}\mathcal{T}^*}$ are rather sequences of action-observation pairs. For this reason it is generally impossible for the participants to provide strategies ${}^{p}f\left(a_\tau\middle|\,d^{1:\tau-1}\right)$ corresponding to data $\left(d_\tau\right)_{\tau\in\cup_{p\in p^*} {}^{p}\mathcal{T}^*}$ in terms of time.

More rigorous approach to the construction of the global prior pdf using the information $\left(d_\tau\right)_{\tau\in {}^{p}\mathcal{T}^*} \in {}^{p}B$ is to create a set of prior pdfs

$$ F \equiv \left\{ f\left(\Theta\right) \in \mathcal{F}(\Theta) \middle|\; f\left(\Theta\right) \propto f_0(\Theta) \prod_{p\in p^*} \prod_{\tau\in {}^{p}\mathcal{T}^*} f\left(\Delta_\tau\middle|\,a_\tau,\Theta\right), \left(d_\tau\right)_{\tau\in {}^{1}\mathcal{T}^*} \in {}^{1}B, \ldots, \left(d_\tau\right)_{\tau\in {}^{\mathring{p}}\mathcal{T}^*} \in {}^{\mathring{p}}B \right\}. \tag{3.14} $$

The pdfs in F represent possible global prior pdfs whereas there is not any order in a sense "more likely then" given on them. Apart from technical difficulties related to treating a set of prior pdfs, this approach has also a more fundamental drawback: it leads to a loss of a weak order, see Appendix A, on a set of decision strategies, which means, in fact, a deflection from the Bayesian framework. To illustrate it, we recall a basic property of the Bayesian decision making: for a single prior pdf, preference order $\prec$ on a set of decision strategies R^* induced by the expected loss, i.e., for $R_1, R_2 \in R^*$, $R_1 \prec R_2$ iff

$$\int L(d) f_{R_1}(d|\Theta) f(\Theta) \mathrm{d}d\mathrm{d}\Theta > \int L(d) f_{R_2}(d|\Theta) f(\Theta) \mathrm{d}d\mathrm{d}\Theta, \tag{3.15}$$

is a weak order. This property implies, that a mutual incomparability of the decision strategies is an equivalence relation, and the equivalence classes form a strictly ordered set. It ensures that the optimal decision strategies, being maximal elements of the ordered set $(R^*, \prec)$, possess appealing properties, e.g., that for an optimal strategy R and an arbitrary strategy $\tilde{R}$ it holds

$$\tilde{R} \prec R \quad \text{or} \quad \tilde{R} \text{ is optimal.} \tag{3.16}$$

However, if there is a set F of possible prior pdfs, it is generally impossible to define a preference order on R^* via (3.15) due to the ambiguity of a prior pdf. Instead, we can define a relation $\prec'$ on R^* so that $R_1 \prec' R_2$ iff

$$\int L(d) f_{R_1}(d|\Theta) f(\Theta) \mathrm{d}d\mathrm{d}\Theta \geq \int L(d) f_{R_2}(d|\Theta) f(\Theta) \mathrm{d}d\mathrm{d}\Theta, \tag{3.17}$$

for all $f(\Theta) \in F$ and the inequality in (3.17) is strict for some $f(\Theta) \in F$. The relation $\prec'$ is an analogy of the relation induced by dominance of decision strategies [8], in which case the inequality (3.17) must hold for all $f(\Theta) \in \mathcal{F}(\Theta)$. Compared to $\prec$ defined by (3.15), the relation $\prec'$ is only a strict partial order on R^*. Restricting the choice of a resulting strategy to the maximal elements of $(R^*, \prec')$, certain rationality of the decision making is guaranteed. It is similar as a restriction to non-dominated strategies in case that a prior pdf is not considered. However, the maximal elements of a partially ordered set do not possess any property similar to (3.16). While the maximal elements of $(R^*, \prec)$ are, due to (3.16), taken as "equally good", which corresponds to the fact that all of them lead to the same value of the expected loss, the maximal elements of $(R, \prec')$ are just incomparable.

Notes:

- The importance of the above outlined construction of the global prior pdf, at least the case we are able to complete, consists in the fact that under suitable selection of the pre-prior pdfs it represents an exact solution of the problem addressed. Any other more generally applicable method should ideally reduce to the outlined solution if it is applied to local prior pdfs corresponding to some posterior ones.

- Some kind of non-informative prior pdfs [10] could serve well as the pre-prior pdfs of the individual participants as well as of the global task.

- The assumption that the sets ${}^{p}\tau^*$ in (3.9) are pairwise disjoint gives an interpretation to what we have labeled as an independence of sources of prior information.

- The assumption on the independence of the sources of prior information, is substantial for the construction of the global prior pdf by (3.12) or classes of prior pdfs (3.14). However, in practice the independence can be hardly guaranteed. The problem of combining overlapping information pieces is widely addressed in the literature, see, e.g., [21] or [13]. Nevertheless, the existing methods commonly assume that the kind of overlapping as well as its measure are known, which makes them unsuitable for our approach to the multiple participant decision making. As it is practically impossible to detect the overlapping of information from the local prior pdfs themselves, we expect that if the independence of the sources of prior information is not given explicitly, then any attempt to treat multiple sources of prior information in an exact manner will lead to a loss of the weak order on decision strategies.

- Evaluation of (3.12) leads to a global prior pdf in an intractable form even if the global parametric model is from an exponential family. If this approach is to be applied in practice, some kind of approximation of the prior pdfs must be employed.

3.5.3 Preference Description

Because the global decision task is constructed as a Bayesian one, the preferences on the global level are to be described by a loss function on the system. Due to the separated assessments of uncertainty and preferences in the Bayesian decision making [10], we suppose that the global loss function $L\left(d^{1:\mathring{t}}\right)$ should depend on loss functions ${}^{p}L\left({}^{p}d^{1:\mathring{t}}\right)$ of individual participants and not on their parametric models or prior pdfs. Contrary to local parametric models and prior pdfs, which could be taken as supplemental blocks of information, the local loss functions describe more or less antagonistic objectives and the global loss function must be searched as some kind of compromise among them. In accordance with the requirements for fully cooperating participants stated in Section 3.5, the global loss function is to be constructed as an even compromise among preferences given by loss function of individual participants so that objectives of any participant are in no way preferred to objectives of others.

Problems of this type are widely addressed in the literature. The most famous result in this direction comes from Arrow [6]. Arrow in his impossibility theorem proved that any procedure which assigns a group preference ordering to a set of individual preference orderings violates at least one of five conditions, which are generally accepted as necessary ones for the procedure to be fair. However, Arrow's theorem deals only with qualitative preference orderings, i.e., preferences expressed in terms of a weak order. In our case, the point of departure is somewhat different, as due to presence of uncertainty a quantitative component of preferences must be taken into account.

Construction of a global loss function from a group of local ones (or equivalently in terms of utility functions) has been addressed by Keeney [32], [33]. In [32] Keeny proved that under presence of uncertainty a group utility function satisfies conditions analogous to that proposed by Arrow if and only if it is a linear combination of individuals' utility functions with nonnegative coefficients and at least two of them are positive. Although the Keeney's proof has some week points, e.g., he implicitly supposes that the group utility function depends on the individual ones only pointwise, we find the linear combination of local loss functions to be a suitable candidate for a global loss function for fully cooperating participants.

It is well known and easily verified that any positive linear transformation of a loss function preserves the order on randomized strategies induced via expected loss. It means that the loss functions of individual participants are not determined uniquely, which disallows mutual comparability of the preferences. In order to avoid the impact of a positive linear transformation of a local loss function on the global one, we select for each participant a unique loss function representing its preferences on a common scale. Particularly, we select a loss function which assigns the loss 0 to the most preferred data and 1 to the least preferred ones in case of a non-constant bounded loss function, and 0 in case of a constant one, i.e.,

$$ {}^{p}\mathcal{L}({}^{p}d^{1:\mathring{t}}) = \begin{cases} 0 & \text{if } {}^{p}L({}^{p}d^{1:\mathring{t}}) \text{ is constant} \\ \dfrac{{}^{p}L({}^{p}d^{1:\mathring{t}}) - \min_{{}^{p}d^{1:\mathring{t}}} {}^{p}L({}^{p}d^{1:\mathring{t}})}{\max_{{}^{p}d^{1:\mathring{t}}} {}^{p}L({}^{p}d^{1:\mathring{t}}) - \min_{{}^{p}d^{1:\mathring{t}}} {}^{p}L({}^{p}d^{1:\mathring{t}})} & \text{otherwise} \end{cases} . \tag{3.18} $$

In case of unbounded loss functions, the inter-participant comparison of preferences cannot be performed merely on the basis of the loss functions themselves and must be given explicitly.

The normalized local loss functions ${}^{p}\mathcal{L}({}^{p}d^{1:\mathring{t}})$ are defined on the environments of individual participants. In order to combine them into a global loss function it is convenient to extend them to a common domain, that is, to the complete system. The extension can be done naturally as the preferences of a participant p about data $d^{1:\mathring{t}} \in \left(d^{1:\mathring{t}}\right)^{*}$ are completely determined by a part of data $d^{1:\mathring{t}}$ which is related to its environment, i.e., by the projections $({}^{p}S(d_1), \ldots, {}^{p}S(d_{\mathring{t}}))$, where ${}^{p}S(\cdot)$ is defined by (3.11). The extended loss functions, distinguished from the original ones by the arguments, are then defined by

$$ {}^{p}\mathcal{L}\left(d^{1:\mathring{t}}\right) = {}^{p}\mathcal{L}\left({}^{p}S(d_1), \ldots, {}^{p}S(d_{\mathring{t}})\right). \tag{3.19} $$

Now, consider a global loss function $L(d^{1:\mathring{t}})$ in a form of a linear combination of the extended local loss functions (3.19)

$$L(d^{1:\mathring{t}}) = \sum_{p=1}^{\mathring{p}} {}^{p}\alpha \, {}^{p}\mathcal{L}(d^{1:\mathring{t}}) \tag{3.20}$$

with nonnegative weights ${}^{p}\alpha$. Due to the invariance of preferences with respect to positive linear transformations of a loss function, we can assume that the weights ${}^{p}\alpha$ satisfy $\sum_{p\in p^*} {}^{p}\alpha = 1$. It is immediately seen that for any positive weights ${}^{p}\alpha$ the global loss functions (3.20) have the following properties which make them suitable candidates for a fair global loss function in terms of conditions stated in Section 3.5.

- If for any $d, \tilde{d} \in \left(d^{1:\mathring{t}}\right)^*$ all participants do not prefer d to $\tilde{d}$, then neither d is preferred to $\tilde{d}$ in the global task. If, in addition, there is at least one participant which prefer $\tilde{d}$ to d then $\tilde{d}$ is preferred to d in the global task.
- For the quantitative description of preferences it holds

$$\min_{p\in p^*} {}^{p}\mathcal{L}(d^{1:\mathring{t}}) \leq L(d^{1:\mathring{t}}) \leq \max_{p\in p^*} {}^{p}\mathcal{L}(d^{1:\mathring{t}}).$$

The weights ${}^{p}\alpha$ can be interpreted as the importance of the objectives of the p-th participant – the higher ${}^{p}\alpha$ is, the more the global preferences are influenced by the preferences of p-th participant. For a global loss function $L(d^{1:\mathring{t}})$, which does not prefer objectives of any participant to objectives of the others, the weights ${}^{p}\alpha$ are to be selected evenly, i.e., ${}^{p}\alpha = 1/\mathring{p}$, for all p. Note that (3.20) with weights ${}^{p}\alpha = 1/\mathring{p}$ is a compromise among ${}^{p}L({}^{p}d^{1:\mathring{t}})$ in the sense that, for any participant p, the "partial loss" ${}^{p}\alpha \, {}^{p}\mathcal{L}(d^{1:\mathring{t}})$ depends only on a relative position of $d^{1:\mathring{t}}$ with respect the less and to the most preferred data of the participant p. This specificity becomes more obvious especially in case that the local loss functions ${}^{p}L({}^{p}d^{1:\mathring{t}})$ are expressed, e.g., as a negative profit in a common currency with maximal values being significantly different for individual participants.

3.6 Summary

To make the design of a distributed decision strategy meaningful, it is necessary to formulate some kind of a global task, which enables to compare quality of the $\mathring{p}$-tuples of participants' strategies. However, no unique global task follows from the original participants' decision tasks directly. On that account, additional requirements on a "nature" of the global task must be stated. We have made an attempt to formulate the global task as a Bayesian one, Section 3.4, so that it fits the requirements for fully cooperating participants stated in Section 3.5. In order to simplify the construction of the global prior pdf, we have focused on a special case in which the prior information can be transformed to sets of virtual observations.

While a suitable global preference assessment can be simply acquired as a weighted sum of normalized local loss functions, the treatment of the uncertainty in the global task is a much harder problem. The outlined solution reveals a number of, more or less, technical problems. The most significant ones are related to the choice of a suitable global parametric model, form of a global prior pdf resulting from (3.12), and exploitation of information in a form of non-conjugated local prior pdfs.

However, apart from the technical issues, the presented approach suffers also from a conceptual shortcoming. Namely, the requirements for the global task

- to be formulated as a Bayesian one,
- to exploit complete information provided by individual participants,
- not to introduce an information for which there is no evidence

are, in general, contradicting. This problem arises, e.g., if the participants are not able to provide information about data in the state vector of the global task, or if the prior information pieces of the individual participants cannot be guaranteed to come from independent sources.

If we omit a possibility to leave out problematic pieces of information, a solution of the discussed problem could be based on employing a more general form of the global task and adding an information supported by an evidence. Especially the later case seems to be worth of attention. In practice, it means that a more detailed modelling should be employed. For example, on a participant level all quantities in state vectors of parametric models are to be modelled. On a global level, modelling of dependencies among prior information pieces provided by individual participants should be used. However, such models are inevitably highly dependent on a particular application.

Employing a more general form, i.e., a non-Bayesian one, of the global task would be a rigorous way to treat the incomplete information. Imprecise probabilities [50] could be appropriate means for this purpose. Although, from the practical point of view, this approach would be more complex than the strictly Bayesian one, there are indications that it could be feasible; see, e.g., [45].

In this chapter we have focused on a construction of the global task and left the design of the cooperation methods open. However, the form of the global task suggests what objectives are to be followed by the cooperation. Namely, the knowledge about the system behaviour possessed by any of the participants should be provided to all others which are able to exploit it, i.e., to its neighbours. Furthermore, the objectives of the individual participants should be harmonized so that the new objectives are not contradictory for any pair of neighbours whereas they are close to the original ones. While the harmonization of the objectives is clearly related to the fully cooperating participants discussed in this chapter, the knowledge sharing is supposed to be employed in a much wider class of multiple participant decision making tasks. These observations motivates the need for practically feasible, though possibly approximate, cooperation methods proposed in the following chapters.

Chapter 4

Global Objective Setting

In this chapter, a method is proposed which enables the participants employing the fully probabilistic design, see Section 2.4, to create the global objectives so that they are close to a set of local ones. Within the FPD, the objectives are described by so-called ideal pdfs. The problem of constructing the global objectives can be then roughly formulated in the following way: for a set of given (local) ideal pdfs of the individual participants defined on their environments find a common ideal pdf on the system so that it is, in some sense, close to the given local ideal pdfs. The acquired global ideal pdf can be taken as an analogy of the global loss function 3.20, which characterizes compromise objectives for a group of participants performing the regular Bayesian decision making.

In Section 4.1, we propose a function through which the proximity of a joint pdf and a set of marginal pdfs is measured. The global ideal pdf is then searched as a pdf which minimizes the established function. Its properties are investigated in Section 4.2. In Section 4.3, basic elements of an iterative algorithm for an approximate solution of the optimization task are presented. An application of the proposed method in the distributed multiple participant decision making is briefly discussed in Section 4.4.

4.1 Notation and Problem Formulation

The problem addressed in this chapter is not strictly related to the rest of the decision making process. For this reason, some of the items in the notation, which are not essential in this chapter, can be omitted. Changes in the notation used throughout the chapter are as follows:

- Actions and observations need not be distinguished. On that account, the considered random quantities are simply denoted by d, possibly with additional subscripts or superscripts.
- Time evolution of the system need not be considered. Thus, the time subscripts are omitted.
- The only pdfs related to individual participants are their ideal pdfs. For this reason, the left superscript I, used for the ideal pdfs of individual participants, is omitted. It is used only for the searched common ideal pdf.

With the modified notation, the considered system is described by a vector random quantity $d \equiv (d_1, d_2, \ldots, d_n), n \in \mathbb{N}$. For each participant $p \in p^* \equiv (1, 2, \ldots, \mathring{p})$:

- ${}^{p}d \equiv (d_i)_{i \in {}^{p}i^*}$, where $\emptyset \neq {}^{p}i^* \subset \{1, 2, \ldots, n\}$, is a random quantity treated by the p-th participant,
- ideal pdf ${}^{p}f({}^{p}d)$ is specified by the p-th participant on its data ${}^{p}d$,
- ${}^{p}\alpha$ is a given nonnegative weight (the meaning of which is clarified later).

Without loss of generality, we assume that $\bigcup_{p \in p^*} {}^{p}i^* = \{1, 2, \ldots, n\}$, and $\sum_{p \in p^*} {}^{p}\alpha = 1$.

As a measure of proximity of a joint pdf $f\,(d)$ to a set of pdfs ${}^{p}f\,({}^{p}d)$ we use a function $\mathcal{D}$ acting on the set $\mathcal{F}$ of all pdfs of d, which, for given pdfs ${}^{p}f({}^{p}d)$ and weights ${}^{p}\alpha$, is defined by

$$\mathcal{D}(f) = \sum_{p \in p^*} {}^{p}\alpha \mathsf{D}\left({}^{p}f({}^{p}d) || f({}^{p}d)\right). \tag{4.1}$$

The problem of selecting a common ideal pdf ${}^{I}f\,(d)$ is then formulated as follows:

For given pdfs ${}^{p}f({}^{p}d)$ and weights ${}^{p}\alpha$, find ${}^{I}f(d)$ so that

$${}^{I}f(d) \in \operatorname*{argmin}_{f \in \mathcal{F}} \mathcal{D}(f), \tag{4.2}$$

where $\mathcal{D}(f)$ is defined by (4.1).

The particular choice of the function $\mathcal{D}$ is motivated by the following reasons, which stem from the ideas behind the fully probabilistic design; see Section 2.4.

- Within the FPD framework the proximity of a pair of pdfs is measured by the Kullback-Leibler divergence.

- Whenever the Kullback-Leibler divergence is used as a measure of proximity of a pdf $f(d)$ and its approximation $g(d)$, the arguments should be ordered as $\mathsf{D}\,(f||g)$. It can be justified in the following way: $\mathsf{D}\,(f||g) = \int f(d)(\ln f(d) - \ln g(d))\mathrm{d}d$ can be interpreted as a negative value of the expected log-likelihood of pdf $g(d)$ for data distributed according to $f(d)$, shifted by an additive term $\int f(d) \ln f(d)\mathrm{d}d$, so that its minimum, reached for $g = f$, is 0. In the addressed problem, the ideal pdfs ${}^{p}f({}^{p}d)$ play the role of the original pdfs and ${}^{I}f({}^{p}d)$ is the searched approximation; thus, the Kullback-Leibler divergence $\mathsf{D}\left({}^{p}f({}^{p}d) \middle|\middle| {}^{I}f({}^{p}d)\right)$ is used. Another reasoning can be found, e.g., in [9].

- For any $p \in p^*$, the value of $\mathcal{D}(f)$ depends on ${}^{p}f({}^{p}d)$ only through the corresponding marginal pdf $f\,({}^{p}d)$.

- Requirements on $\mathsf{D}\left({}^{p}f({}^{p}d) \middle|\middle| {}^{I}f({}^{p}d)\right)$ to be small for all $p \in p^*$ are contradicting in general. For positive weights ${}^{p}\alpha$, the minimization of the weighted sum (4.1) ensures that the resulting pdf ${}^{I}f\,(d)$ is a non-dominated solution, in the sense that there is no other ${}^{I}\tilde{f}(d)$ such that

  $$\mathsf{D}\left({}^{p}f\,({}^{p}d) \middle|\middle| {}^{I}\tilde{f}({}^{p}d)\right) \leq \mathsf{D}\left({}^{p}f\,({}^{p}d) \,||\, {}^{I}f({}^{p}d)\right),$$

  for all $p \in p^*$, and the inequality is strict for some p. At the same time, the weights ${}^{p}\alpha$ represent "tuning knobs", which allow to reflect the "importance" of the objectives of the individual participants.

- The objectives described by local ideal pdfs are considered to be consistent if for any two participants $p, \tilde{p} \in p^*$ it holds

  $${}^{p}f\left({}^{p,\tilde{p}}d\right) = {}^{\tilde{p}}f\left({}^{p,\tilde{p}}d\right),$$

  where ${}^{p,\tilde{p}}d$ are data common to participants p and $\tilde{p}$ defined by (3.1). If the objectives given by pdfs ${}^{p}f({}^{p}d)$ are consistent, then a reasonable common ideal pdf ${}^{I}f(d)$ must satisfy ${}^{I}f({}^{p}d) = {}^{p}f({}^{p}d)$ for all $p \in p^*$. This requirement is reflected by (4.1), because the property of the Kullback-Leibler divergence (2.3) ensures that $\mathcal{D}({}^{I}f) = 0$ if, $\forall p \in p^*$, ${}^{I}f({}^{p}d) = {}^{p}f({}^{p}d)$, and $\mathcal{D}({}^{I}f) > 0$ otherwise.

In general, the set of pdfs minimizing $\mathcal{D}$ can have more than one element. As there is no reason to prefer any particular element of this set, the searched pdf ${}^{I}f(d)$ can be selected as any one of them.

4.2 General Solution

To find a pdf ${}^{I}f(d)$ satisfying (4.2) is, in general, a hard problem. We start this section with two examples the solution of which is trivial. The first one, in which all participants deal with the complete system, is for its importance formulated as a lemma.

Lemma 4.2.1 *Let $\mathring{p} \geq 1$, and $\forall p \in p^*$, ${}^{p}i^* \equiv \{1\}$. Then, for ideal pdfs ${}^{p}f(d)$ and weights ${}^{p}\alpha$, the function $\mathcal{D}$ defined by* (4.1) *has a unique minimizer ${}^{I}f(d) = \sum_{p\in p^*} {}^{p}\alpha\, {}^{p}f(d)$.*

Proof: For an arbitrary $f(d) \in \mathcal{F}$, $\mathcal{D}(f)$ can be written, using (2.17) and the linearity of the Kerridge inaccuracy in the first argument (2.12), as

$$\begin{aligned}\mathcal{D}(f) &= \sum_{p\in p^*} {}^{p}\alpha \mathsf{D}\left({}^{p}f(d)\,||f(d)\right) = -\sum_{p\in p^*} {}^{p}\alpha \mathsf{H}\left({}^{p}f(d)\right) + \sum_{p\in p^*} {}^{p}\alpha \mathsf{K}\left({}^{p}f(d), f(d)\right) = \\ &= -\sum_{p\in p^*} {}^{p}\alpha \mathsf{H}\left({}^{p}f(d)\right) + \mathsf{K}\left(\sum_{p\in p^*} {}^{p}\alpha\, {}^{p}f(d), f(d)\right). \end{aligned} \tag{4.3}$$

In (4.3), the sum of entropies is independent of $f(d)$ and the Kerridge inaccuracy has - due to (2.13) - a unique minimizer $\sum_{p\in p^*} {}^{p}\alpha\, {}^{p}f(d)$. □

The following example is also one of those few the solution of which is trivial. It is, however, very important, because it represents the most general case with two participants.

Example 4.2.1 *Let $d \equiv (d_1, d_2, d_3)$, $\mathring{p} = 2$, ${}^{1}i^* \equiv \{1,2\}$, and ${}^{2}i^* \equiv \{2,3\}$. The ideal pdfs ${}^{1}f({}^{1}d)$, ${}^{2}f({}^{2}d)$, and weights ${}^{1}\alpha$, ${}^{2}\alpha$ are supposed to be given. Then, $\mathcal{D}(f)$ can be written as*

$$\begin{aligned}\mathcal{D}(f) &= {}^{1}\alpha \int {}^{1}f(d_2)\, {}^{1}f(d_1|d_2) \ln \frac{{}^{1}f(d_1|d_2)}{f(d_1|d_2)} \mathrm{d}d_1 \mathrm{d}d_2 + {}^{2}\alpha \int {}^{2}f(d_2)\, {}^{2}f(d_3|d_2) \ln \frac{{}^{2}f(d_3|d_2)}{f(d_3|d_2)} \mathrm{d}d_3 \mathrm{d}d_2 + \\ &+ \left(\sum_{p=1}^{2} \int {}^{p}\alpha\, {}^{p}f(d_2) \ln \frac{{}^{p}f(d_2)}{f(d_2)} \mathrm{d}d_2\right). \end{aligned} \tag{4.4}$$

From the property of the Kullback-Leibler divergence (2.3)*, it is clear that the first term in* (4.4) *is minimized (equal to 0) if $f(d_1|d_2) = {}^{1}f(d_1|d_2)$. Similarly, the second term is minimized if $f(d_3|d_2) = {}^{2}f(d_3|d_2)$. The last term in* (4.4) *is, according to Lemma 4.2.1, minimized by $f(d_2) = \sum_{p=1}^{2} {}^{p}\alpha\, {}^{p}f(d_2)$. As all three terms in* (4.4) *can be minimized independently, $f(d_1, d_2, d_3)$ minimizes $\mathcal{D}$ if it holds*

$$f(d_1|d_2) = {}^{1}f(d_1|d_2),\ f(d_3|d_2) = {}^{2}f(d_3|d_2),\ \text{and}\ f(d_2) = \sum_{p=1}^{2} {}^{p}\alpha\, {}^{p}f(d_2). \tag{4.5}$$

Thus, ${}^{I}f(d)$ minimizing (4.4) *is, e.g.,*

$${}^{I}f(d) = {}^{1}f(d_1|d_2)\, {}^{2}f(d_3|d_2) \left(\sum_{p=1}^{2} {}^{p}\alpha\, {}^{p}f(d_2)\right).$$

Note that if ${}^{1}f(d_2)$ and ${}^{2}f(d_2)$ are positive on d_2^*, then the conditions (4.5) *are also necessary for $f(d)$ to minimize $\mathcal{D}$.*

The example demonstrates that for $\mathring{p} = 2$ the common ideal pdf has the following appealing properties:

- In spite of the ambiguity of pdfs minimizing (4.4), the marginal pdfs ${}^{I}f\left({}^{1}d\right)$ and ${}^{I}f\left({}^{2}d\right)$ of the common ideal pdf are determined unambiguously:

$$\begin{aligned}{}^{I}f\left({}^{1}d\right) &= {}^{I}f(d_1, d_2) = {}^{1}f(d_1|d_2) \left(\sum_{p=1}^{2} {}^{p}\alpha\, {}^{p}f(d_2)\right), \\ {}^{I}f\left({}^{2}d\right) &= {}^{I}f(d_2, d_3) = {}^{2}f(d_3|d_2) \left(\sum_{p=1}^{2} {}^{p}\alpha\, {}^{p}f(d_2)\right). \end{aligned}$$

- The marginal pdf ${}^{I}f\,(d_2)$, describing objectives on a part of the system which is in common of both participants, is a mixture (with weights ${}^{p}\alpha$) of the corresponding marginal pdfs of the original ideal pdfs, i.e.,

$$ {}^{I}f(d_2) = \sum_{p=1}^{2} {}^{p}\alpha\, {}^{p}f(d_2). $$

- The conditional pdfs ${}^{I}f(d_1|d_2)$, ${}^{I}f(d_3|d_2)$, describing common objectives on parts of the system treated by only one of the participants given the quantity in common of both participants, reflect the original objectives of individual participants:

$$ {}^{I}f(d_1|d_2) = {}^{1}f(d_1|d_2),\ {}^{I}f(d_3|d_2) = {}^{2}f(d_3|d_2). $$

The next example represents the simplest case the solution of which is nontrivial.

Example 4.2.2 *Let* $d \equiv (d_1, d_2)$, $\mathring{p} = 3$, ${}^{1}i^* \equiv \{1\}$, ${}^{2}i^* \equiv \{2\}$, *and* ${}^{3}i^* \equiv \{1,2\}$.
In this case

$$ \begin{aligned} \mathcal{D}(f) &= {}^{1}\alpha \int {}^{1}f(d_1) \ln \frac{{}^{1}f(d_1)}{f(d_1)} \mathrm{d}d_1 + {}^{2}\alpha \int {}^{2}f(d_2) \ln \frac{{}^{2}f(d_2)}{f(d_2)} \mathrm{d}d_2 + \\ &\quad + {}^{3}\alpha \int {}^{3}f(d_1, d_2) \ln \frac{{}^{3}f(d_1, d_2)}{f(d_1, d_2)} \mathrm{d}d_1, \mathrm{d}d_2. \end{aligned} \tag{4.6} $$

Contrary to the preceding example, the pdf minimizing (4.6) *cannot be found directly, because the integrals in* (4.6) *cannot be minimized independently.*

The rest of this chapter is focused on the general case described by (4.2). If not said otherwise, it is assumed that $\mathring{p}$, sets ${}^{p}i^*$, pdfs ${}^{p}f({}^{p}d)$, and weights ${}^{p}\alpha$ are given and thus objects defined by them, e.g., the function $\mathcal{D}$, are well specified. In what follows, an additional notation will prove useful: ${}^{\bar{p}}d$ denotes a data vector describing the part of the system not treated by the p-th participant, i.e., for all $p \in p^*$,

$$ {}^{\bar{p}}d \equiv (d_i)_{i \in \{1,\ldots,n\} \setminus {}^{p}i^*}. $$

Furthermore, we define two particular subsets of $\mathcal{F}$:

$$ \mathcal{G} \equiv \{f\,(d) \in \mathcal{F} | \mathcal{D}(f) < +\infty\} \tag{4.7} $$

$$ \mathcal{H} \equiv \{f\,(d) \in \mathcal{F}\,| \forall p \in p^*, \forall\, {}^{p}d \in {}^{p}d^*, f\,({}^{p}d) \geq {}^{p}\alpha\, {}^{p}f\,({}^{p}d)\,\} \tag{4.8} $$

The sets $\mathcal{G}$ and $\mathcal{H}$ defined by (4.7) and (4.8) have the following elementary properties:

1. $\mathcal{G}$ is nonempty. To prove it, consider an arbitrary $h(d) \in \mathcal{F}$ such that $h(d) > 0$ on d^*. For a pdf

$$ f\,(d) \equiv \sum_{p \in p^*} {}^{p}\alpha\, {}^{p}f\,({}^{p}d)\, h({}^{\bar{p}}d|\, {}^{p}d) $$

it holds

$$ \begin{aligned} \mathcal{D}(f) &= \sum_{p \in p^*} {}^{p}\alpha \int {}^{p}f\,({}^{p}d) \ln \frac{{}^{p}f\,({}^{p}d)}{\int \sum_{r \in p^*} {}^{r}\alpha\, {}^{r}f\,({}^{r}d)\, h({}^{\bar{r}}d|\, {}^{r}d) \mathrm{d}\, {}^{\bar{p}}d} \mathrm{d}\, {}^{p}d \\ &\leq \sum_{p \in p^*} {}^{p}\alpha \int {}^{p}f\,({}^{p}d) \ln \frac{{}^{p}f\,({}^{p}d)}{\int {}^{p}\alpha\, {}^{p}f\,({}^{p}d)\, h({}^{\bar{p}}d|\, {}^{p}d) \mathrm{d}\, {}^{\bar{p}}d} \mathrm{d}\, {}^{p}d = -\sum_{p \in p^*} {}^{p}\alpha \ln {}^{p}\alpha, \end{aligned} $$

and thus $f\,(d) \in \mathcal{G}$.

2. ${}^{I}f(d) \in \mathcal{G}$. This property follows directly from the definition (4.7) of the set $\mathcal{G}$ and the fact that $\mathcal{G} \neq \emptyset$. For this reason, it is sufficient to search for

$$ {}^{I}f(d) \in \underset{f \in \mathcal{G}}{\operatorname{argmin}}\, \mathcal{D}(f) $$

instead of (4.2).

3. $\mathcal{H} \subset \mathcal{G}$. The inclusion follows from definition (4.1) of the operator $\mathcal{D}$ and definition (2.1) of the Kullback-Leibler divergence.

4. Sets $\mathcal{G}$ and $\mathcal{H}$ are convex. The convexity of the set $\mathcal{G}$ follows from the convexity of the Kullback-Leibler divergence (2.4). The convexity of the set $\mathcal{H}$ follows directly from its definition.

A crucial role for derivation of properties of ${}^{I}f(d)$ minimizing the function $\mathcal{D}$ has the operator $A : \mathcal{G} \to \mathcal{G}$ defined by

$$ Af = \sum_{p=1}^{\mathring{p}} {}^{p}\alpha f\left({}^{\bar{p}}d\middle|\,{}^{p}d\right) {}^{p}f\left({}^{p}d\right). \tag{4.9} $$

Note that the operator A is well defined in the sense that if for some p it holds $f\left({}^{p}d\right)|_{{}^{p}d={}^{p}\tilde{d}} = 0$ for some ${}^{p}\tilde{d} \in {}^{p}d^{*}$, then it holds ${}^{p}f\left({}^{p}d\right)|_{{}^{p}d={}^{p}\tilde{d}} = 0$, because $f(d) \in \mathcal{G}$. The ambiguity in $f\left({}^{\bar{p}}d\middle|\,{}^{p}d\right)|_{{}^{p}d={}^{p}\tilde{d}}$ is then irrelevant.

A key property of the operator A is given by the following proposition.

Proposition 4.2.1 $\forall f(d) \in \mathcal{G}$,

$$ \mathcal{D}(f) - \mathcal{D}(Af) \geq \mathsf{D}\left(Af||f\right). $$

Proof: The proof is straightforward and is based on definitions of the Kerridge inaccuracy (2.10) and the Kullback-Leibler divergence (2.1), properties of the Kerridge inaccuracy (2.15), (2.13), and on definition (4.9) of the operator A.

$$ \mathcal{D}(f) - \mathcal{D}(Af) \overset{(4.1)}{=} \sum_{p=1}^{\mathring{p}} {}^{p}\alpha \int_{{}^{p}d^{*}} {}^{p}f\left({}^{p}d\right) \ln \frac{(Af)({}^{p}d)}{f\left({}^{p}d\right)} \mathrm{d}\,{}^{p}d $$

$$ \begin{aligned} &= \sum_{p=1}^{\mathring{p}} {}^{p}\alpha \int_{d^{*}} {}^{p}f\left({}^{p}d\right) f\left({}^{\bar{p}}d\middle|\,{}^{p}d\right) \ln \frac{(Af)({}^{p}d)}{f\left({}^{p}d\right)} \mathrm{d}d \\ &= \sum_{p=1}^{\mathring{p}} {}^{p}\alpha \int_{d^{*}} {}^{p}f\left({}^{p}d\right) f\left({}^{\bar{p}}d\middle|\,{}^{p}d\right) \ln \frac{(Af)({}^{p}d) f\left({}^{\bar{p}}d\middle|\,{}^{p}d\right)}{f\left({}^{p}d\right) f\left({}^{\bar{p}}d\middle|\,{}^{p}d\right)} \mathrm{d}d \qquad (4.10) \\ &\overset{(2.10)}{=} \sum_{p=1}^{\mathring{p}} {}^{p}\alpha \left(\mathsf{K}\left({}^{p}f\left({}^{p}d\right) f\left({}^{\bar{p}}d\middle|\,{}^{p}d\right), f(d)\right) - \mathsf{K}\left({}^{p}f\left({}^{p}d\right) f\left({}^{\bar{p}}d\middle|\,{}^{p}d\right), (Af)({}^{p}d) f\left({}^{\bar{p}}d\middle|\,{}^{p}d\right)\right)\right) \\ &\overset{(2.15),(2.13)}{\geq} \sum_{p=1}^{\mathring{p}} {}^{p}\alpha \left(\mathsf{K}\left({}^{p}f\left({}^{p}d\right) f\left({}^{\bar{p}}d\middle|\,{}^{p}d\right), f(d)\right) - \mathsf{K}\left({}^{p}f\left({}^{p}d\right) f\left({}^{\bar{p}}d\middle|\,{}^{p}d\right), (Af)(d)\right)\right) \\ &\overset{(2.10)}{=} \sum_{p=1}^{\mathring{p}} {}^{p}\alpha \int_{d^{*}} {}^{p}f\left({}^{p}d\right) f\left({}^{\bar{p}}d\middle|\,{}^{p}d\right) \ln \frac{(Af)(d)}{f(d)} \mathrm{d}d \overset{(4.9),(2.1)}{=} \mathsf{D}\left(Af||f\right) \end{aligned} $$

Remind, that the function $\mathcal{D}$ is defined using the Kullback-Leibler divergence. On that account, the convention $0 \ln 0 = 0$, adopted in its definition (2.1), is employed in the above expressions. Then, e.g., the equality (4.10) holds also in case that $f\left({}^{\bar{p}}d\middle|\,{}^{p}d\right) = 0$ for some $d \in d^{*}$. □

A direct consequence of Proposition 4.2.1 gives a necessary condition for ${}^{I}f(d)$ to be a minimizer of $\mathcal{D}(f)$.

Proposition 4.2.2 *If* ${}^{I}f(d) \in \operatorname{argmin}_{f\in\mathcal{G}} \mathcal{D}(f)$, *then it holds*

$$A\,{}^{I}f = {}^{I}f. \tag{4.11}$$

Proof: Suppose that ${}^{I}f(d) \in \operatorname{argmin}_{f\in\mathcal{G}} \mathcal{D}(f)$ and $A\,{}^{I}f \neq {}^{I}f$. Then $\mathsf{D}\left(A\,{}^{I}f \middle|\middle| {}^{I}f\right) > 0$ and, due to Proposition 4.2.1, it holds

$$\mathcal{D}(A\,{}^{I}f) \leq \mathcal{D}(f) - \mathsf{D}\left(A\,{}^{I}f \middle|\middle| {}^{I}f\right) < \mathcal{D}(f),$$

which is in a contradiction with ${}^{I}f(d) \in \operatorname{argmin}_{f\in\mathcal{G}} \mathcal{D}(f)$. □

Lemma 4.2.2 $\forall f(d) \in \mathcal{G},\ Af \in \mathcal{H}$.

Proof: For all $p \in p^*$, it holds

$$(Af)({}^{p}d) = \int \sum_{r\in p^*} {}^{r}\alpha\, {}^{r}f({}^{r}d)\, f\left({}^{\bar{r}}d \middle|\, {}^{r}d\right) \mathrm{d}\,{}^{\bar{p}}d \geq {}^{p}\alpha\, {}^{p}f({}^{p}d).$$

□

Lemma 4.2.3

$$\operatorname*{argmin}_{f\in\mathcal{G}} \mathcal{D}(f) \subset \mathcal{H}$$

Proof: The lemma follows directly from Proposition 4.2.2 and Lemma 4.2.2. □

The opposite implication to Proposition 4.2.2 does not hold in general. However, under an additional assumption, the equality $A\,{}^{I}f = {}^{I}f$ provides also a sufficient condition for ${}^{I}f(d)$ to be a minimizer of the function $\mathcal{D}$.

Proposition 4.2.3 *Let* ${}^{I}f(d) \in \mathcal{H}$, ${}^{I}f(d) > 0$ *on* d^*, *and* $A\,{}^{I}f = {}^{I}f$. *Then it holds*

$${}^{I}f(d) \in \operatorname*{argmin}_{f\in\mathcal{G}} \mathcal{D}(f).$$

Proof: In this proof we follow the basic idea of the calculus of variations. For fixed pdfs $f(d) \in \mathcal{H}$, such that $f(d) > 0$ on d^*, and $h(d) \in \mathcal{F}$, let us define a function $q_{f,h} : [0,1] \to \mathbb{R}$,

$$q_{f,h}(\omega) \equiv \mathcal{D}((1-\omega)f + \omega h) = \sum_{p\in p^*} \int {}^{p}f({}^{p}d) \ln \frac{{}^{p}f({}^{p}d)}{(1-\omega)f({}^{p}d) + \omega h({}^{p}d)} \mathrm{d}\,{}^{p}d.$$

First, we prove that $q_{f,h}$ has a derivative on a (right) neighbourhood of 0, and we evaluate it. For all $p \in p^*$ and ${}^{p}d \in {}^{p}d^*$, it holds

$$\begin{aligned} &\left| \frac{\partial}{\partial\omega} \left({}^{p}f({}^{p}d) \ln \frac{{}^{p}f({}^{p}d)}{(1-\omega)f({}^{p}d) + \omega h({}^{p}d)} \right) \right| \\ &\qquad = \left| {}^{p}f({}^{p}d)\, \frac{h({}^{p}d) - f({}^{p}d)}{(1-\omega)f({}^{p}d) + \omega h({}^{p}d)} \right| \leq {}^{p}f({}^{p}d)\, \frac{h({}^{p}d) + f({}^{p}d)}{(1-\omega)f({}^{p}d)} \leq \frac{h({}^{p}d) + f({}^{p}d)}{{}^{p}\alpha(1-\omega)}, \end{aligned} \tag{4.12}$$

where the last inequality follows from $f(d) \in \mathcal{H}$, see (4.8). Thus, for all $p \in p^*$, the expression 4.12 has an integrable upper bound independent of ω on $[0, \omega_0]$, for some $\omega_0 > 0$, which ensures that the derivative of $q_{f,h}(\omega)$ exists on some right neighbourhood of 0. Its value in $\omega = 0$ is equal to

$$\begin{aligned} \left.\frac{\partial q_{f,h}(\omega)}{\partial\omega}\right|_{\omega=0} &= \sum_{p\in p^*} {}^{p}\alpha \int_{{}^{p}d^*} {}^{p}f({}^{p}d)\, \frac{f({}^{p}d) - h({}^{p}d)}{f({}^{p}d)} \mathrm{d}\,{}^{p}d \\ &= 1 - \int_{d^*} \left(\sum_{p\in p^*} {}^{p}\alpha \frac{{}^{p}f({}^{p}d)}{f({}^{p}d)} \right) h(d)\mathrm{d}d \overset{(4.9)}{=} 1 - \int_{d^*} \frac{Af(d)}{f(d)} h(d)\mathrm{d}d. \end{aligned} \tag{4.13}$$

Now, assume that ${}^{I}f(d) \in \mathcal{H}$, ${}^{I}f(d) > 0$ on d^*, $A\,{}^{I}f = {}^{I}f$, and $\mathcal{D}(f) < \mathcal{D}({}^{I}f)$ for some $f(d) \in \mathcal{G}$. Then, because $\mathcal{D}$ is a convex function on $\mathcal{G}$, it holds

$$\left.\frac{\partial q_{{}^{I}f,f}(\omega)}{\partial \omega}\right|_{\omega=0} = \lim_{\varepsilon \to 0+} \frac{\mathcal{D}((1-\varepsilon)\,{}^{I}f + \varepsilon f) - \mathcal{D}({}^{I}f)}{\varepsilon} \leq \mathcal{D}(f) - \mathcal{D}({}^{I}f) < 0. \tag{4.14}$$

Simultaneously, according to (4.13) it holds

$$\left.\frac{\partial q_{{}^{I}f,f}(\omega)}{\partial \omega}\right|_{\omega=0} = 1 - \int \frac{A\,{}^{I}f(d)}{{}^{I}f(d)} f(d)\,\mathrm{d}d = 0,$$

which is in a contradiction with (4.14). □

Note that without the assumption that ${}^{I}f(d) > 0$ on d^* the implication in Proposition 4.2.3 need not hold, as it is illustrated by the following example. On the other hand, most likely this assumption is not necessary and could be replaced by a weaker one.

Example 4.2.3 *Let* $d \equiv (d_1, d_2), d_1^* \equiv d_2^* \equiv \{0,1\}, \mathring{p} = 2$, ${}^{1}d \equiv d_1$, ${}^{2}d \equiv d_2$, ${}^{1}\alpha, {}^{2}\alpha \in (0,1)$, ${}^{1}\alpha + {}^{2}\alpha = 1$, *and*

$${}^{1}f(d_1) \equiv \begin{cases} p & \text{for } d_1 = 0 \\ 1-p & \text{for } d_1 = 1 \end{cases}, \quad {}^{2}f(d_2) \equiv \begin{cases} q & \text{for } d_2 = 0 \\ 1-q & \text{for } d_2 = 1 \end{cases},$$

for some $p, q \in (0,1)$.

For any ${}^{I}f(d_1, d_2)$ *such that the marginal pdfs* ${}^{I}f(d_1)$ *and* ${}^{I}f(d_2)$ *are equal to pdfs* ${}^{1}f(d_1)$ *and* ${}^{2}f(d_2)$, *respectively, it holds* ${}^{I}f(d) \in \operatorname{argmin}_{f \in \mathcal{G}} \mathcal{D}(f)$ *and* $\mathcal{D}({}^{I}f) = 0$. *Obviously, such* ${}^{I}f(d)$ *exists. It can be selected, e.g., as*

$${}^{I}f(d_1, d_2) \equiv \begin{cases} \min(p,q) & \text{for } d_1 = 0, d_2 = 0 \\ p - \min(p,q) & \text{for } d_1 = 0, d_2 = 1 \\ q - \min(p,q) & \text{for } d_1 = 1, d_2 = 0 \\ 1 - \max(p,q) & \text{for } d_1 = 1, d_2 = 1 \end{cases}.$$

Now, consider a pdf $\tilde{f}(d_1, d_2)$ *defined by*

$$\tilde{f}(d_1, d_2) \equiv \begin{cases} {}^{1}\alpha p + {}^{2}\alpha q & \text{for } d_1 = 0, d_2 = 0 \\ {}^{1}\alpha(1-p) + {}^{2}\alpha(1-q) & \text{for } d_1 = 1, d_2 = 1 \\ 0 & \text{otherwise} \end{cases}.$$

It is easy to verify that $A\tilde{f} = \tilde{f}$ *for any* $p, q \in (0,1)$, *but* $\tilde{f} \in \operatorname{argmin}_{f \in \mathcal{G}} \mathcal{D}(f)$ *only if* $p = q$.

Proposition 4.2.2 offers an appealing interpretation of ${}^{I}f(d)$ defined by (4.2): assume, for a while, that all participants specify their ideal pdfs on the complete system. Without considering any criterion, like (4.2), the convex combination $\sum_{p \in p^*} {}^{p}\alpha\, {}^{p}f(d)$ seems to be a suitable candidate for a common ideal pdf. Now, assume that the participants specify their ideal pdfs ${}^{p}f({}^{p}d)$ only on their environments. It can be interpreted so that every participant p accepts any objectives regarding ${}^{\bar{p}}d$. In other words, any ideal pdf, say ${}^{p}\tilde{f}(d)$, such that ${}^{p}\tilde{f}({}^{p}d) = {}^{p}f({}^{p}d)$ is accepted by the p-th participant as a description of its objectives. What conditional pdfs ${}^{p}\tilde{f}({}^{\bar{p}}d|\,{}^{p}d)$ should be used to extend ${}^{p}f({}^{p}d)$ to ${}^{p}\tilde{f}(d)$? A natural answer is "the ones on which the participants agree". Then, the desired extensions ${}^{p}\tilde{f}({}^{\bar{p}}d|\,{}^{p}d)$ are the corresponding conditional pdfs ${}^{I}f({}^{\bar{p}}d|\,{}^{p}d)$. Altogether, such a common ideal pdf must satisfy the condition

$${}^{I}f(d) = \sum_{p \in p^*} {}^{p}\alpha\, {}^{p}f({}^{p}d)\; {}^{I}f\left({}^{\bar{p}}d\middle|\,{}^{p}d\right),$$

i.e., for the common ideal pdf ${}^{I}f(d)$ it must hold ${}^{I}f = A\,{}^{I}f$. Of course, from this consideration it does not follow that such ${}^{I}f(d)$ exists. It just says, that if it exists, then it could be taken as a suitable candidate for a common ideal pdf.

4.3 Iterative Algorithm

As an analytical solution of the equation (4.11) is not known (up to few trivial cases, e.g., Example 4.4), Proposition 4.2.2 itself cannot be used to find candidates for the common ideal pdf. However, under some additional assumptions, an approximation of ${}^{I}f\,(d) \in \arg\min_{f\in\mathcal{G}} \mathcal{D}(f)$ can be found using an iterative algorithm based on the propositions stated in Section 4.2. A core of the algorithm consist in repetitive application of the operator A defined by (4.9). Namely, for an arbitrary pdf $\varphi_0(d) \in \mathcal{G}$, we consider a sequence of pdfs $(\varphi_k(d))_{k=0}^{+\infty}$ defined recursively by

$$\varphi_{k+1} = A\varphi_k. \tag{4.15}$$

Proposition 4.2.1 ensures that $(\mathcal{D}(\varphi_k))_{k=0}^{+\infty}$ is a non-increasing sequence. Particularly, if it is guaranteed that $\varphi_k(d) > 0$ on d^*, then it holds, according to Lemma 4.2.2 and Proposition 4.2.3, that

$$\begin{aligned} &\mathcal{D}(\varphi_{k+1}) < \mathcal{D}(\varphi_k) \quad if \quad \varphi_k \neq \varphi_{k+1}, \\ &\varphi_k \in \operatorname*{argmin}_{f\in\mathcal{G}} \mathcal{D}(f) \quad if \quad \varphi_k = \varphi_{k+1}. \end{aligned}$$

However, to this point, nothing guarantees that $\mathcal{D}(\varphi_k) - \mathcal{D}(\varphi_{k+1})$ being arbitrarily small, yet positive, for some positive $\varphi_k(d)$, implies that $\mathcal{D}(\varphi_k)$ is close to $\mathcal{D}({}^{I}f)$. In other words, still it is not assured that $\lim_{k\to\infty} \mathcal{D}(\varphi_k) = \mathcal{D}({}^{I}f)$, even if it is provided that $\varphi_k(d) > 0$ on d^*. The convergence and some other issues are discussed in the following paragraphs. The discrete and continuous case are treated separately.

4.3.1 Iterative Algorithm for Discrete Quantities

Suppose that $d_1, \ldots, d_n$ are discrete random quantities with values in finite sets $d_1^*, \ldots, d_n^*$. In this case, the convergence $\lim_{k\to\infty} \mathcal{D}(\varphi_k) = \mathcal{D}({}^{I}f)$ can be easily proved, e.g., if for some $\varepsilon > 0$ it holds $\varphi_k(d) > \varepsilon$ on d^*, for all $k \in \mathbb{N}$. This property of $\varphi_k(d)$ is guaranteed, for example, if for some $p \in p^*$ it holds ${}^{p}d = d$ and ${}^{p}f\,(d) > 0$ on d^*. The convergence is proven by Proposition 4.3.1. In its proof a lower estimate on the Kullback-Leibler divergence of binary random variables proposed by the following lemma is employed.

Lemma 4.3.1 *Let $s, t \in (0,1)$ satisfy $\frac{s}{t} \geq C$ and $t \geq \varepsilon$, for some $C > 1$ and $\varepsilon > 0$. Then it holds*

$$s \ln \frac{s}{t} + (1-s) \ln \frac{1-s}{1-t} \geq C\varepsilon \ln C + (1 - C\varepsilon) \ln \frac{1-C\varepsilon}{1-\varepsilon}.$$

Proof: Let us consider a function $u(a,b) \equiv a \ln \frac{a}{b} + (1-a) \ln \frac{1-a}{1-b}$ for $a, b \in (0,1)$. As for $a > b > 0$, $\frac{\partial}{\partial a} u(a,b) = \ln \frac{a(1-b)}{b(1-a)} > 0$, it holds

$$u(s,t) \geq u(Ct,t). \tag{4.16}$$

Now, define a function

$$v(b) \equiv u(Cb,b) = Cb \ln C + (1 - Cb) \ln \frac{1-Cb}{1-b},$$

for $b \in [0,1)$. We prove that its derivative

$$\frac{\mathrm{d}}{\mathrm{d}b} v(b) = C \ln \frac{C - Cb}{1 - Cb} + \frac{1-C}{1-b}$$

is positive for $b > 0$: For $b = 0$, it holds

$$\left.\frac{\mathrm{d}}{\mathrm{d}b} v(b)\right|_{b=0} = C \ln C + 1 - C > 0, \tag{4.17}$$

because $(C \ln C + 1 - C)|_{C=1} = 0$ and $\frac{\mathrm{d}}{\mathrm{d}C}\,(C \ln C + 1 - C) = \ln C > 0$ for $C > 1$. For the second derivative of $v(b)$ it holds

$$\frac{\mathrm{d}^2}{\mathrm{d}b^2} v(b) = \frac{(C-1)^2}{(1-b)^2(1-Cb)} > 0 \tag{4.18}$$

for $b < \frac{1}{C}$. From 4.17 and 4.18 it follows that $\frac{\mathrm{d}}{\mathrm{d}b} v(b) > 0$ for $b \in [0, \frac{1}{C})$ and thus $v(t) \geq v(\varepsilon)$, which, together with 4.16, proves the lemma.

□

Proposition 4.3.1 *Suppose that the sequence* $(\varphi_k(d))_{k=1}^{+\infty}$ *of pdfs defined by* (4.15)*, for some* $\varphi_0(d) \in \mathcal{G}$*, has the property*

$$\exists \varepsilon > 0, \forall k \in \mathbb{N}, \forall d \in d^*, \varphi_k(d) > \varepsilon.$$

Then it holds

$$\lim_{k\to\infty} \mathcal{D}(\varphi_k) = \mathcal{D}({}^I f), \tag{4.19}$$

where ${}^I f\,(d)$ *satisfies* (4.2)*.*

Proof: Suppose that (4.19) does not hold. Then, because from Proposition 4.2.1 it follows that $(\mathcal{D}(\varphi_k))_{k=1}^{+\infty}$ is decreasing, it holds

$$\exists c > 0, \forall\, {}^I f \in \operatorname*{argmin}_{f\in\mathcal{G}} \mathcal{D}(f), \forall k \in \mathbb{N},\ \mathcal{D}(\varphi_k) - \mathcal{D}({}^I f) \geq c. \tag{4.20}$$

As stated in the proof of Proposition 4.2.3, it holds

$$\frac{\mathrm{d}}{\mathrm{d}\omega}\, \mathcal{D}((1-\omega)\varphi_k + \omega\, {}^I f)\big|_{\omega=0} = 1 - \int \left(\sum_{p\in p^*} {}^p\alpha \frac{{}^p f\,({}^p d)}{\varphi_k({}^p d)} \right) {}^I f\,(d)\,\mathrm{d}d. \tag{4.21}$$

Due to convexity of $\mathcal{D}(\cdot)$, it also holds

$$\frac{\mathrm{d}}{\mathrm{d}\omega}\, \mathcal{D}((1-\omega)\varphi_k + \omega\, {}^I f)\big|_{\omega=0} \leq \mathcal{D}({}^I f) - \mathcal{D}(\varphi_k). \tag{4.22}$$

Then, from definition (4.15) of φ_{k+1}, definition (4.9) of the operator A, and from relations (4.21), (4.22), and (4.20) it follows that

$$\begin{aligned} \int \frac{\varphi_{k+1}(d)}{\varphi_k(d)}\, {}^I f\,(d)\,\mathrm{d}d \overbrace{=}^{(4.15)} \int \frac{A\varphi_k(d)}{\varphi_k(d)}\, {}^I f\,(d)\,\mathrm{d}d \overbrace{=}^{(4.9),(4.21)} 1 - \frac{\mathrm{d}}{\mathrm{d}\omega}\mathcal{D}((1-\omega)\varphi_k + \omega\, {}^I f) \\ \overbrace{\geq}^{(4.22)} 1 + \mathcal{D}(\varphi_k) - \mathcal{D}({}^I f) \overbrace{\geq}^{(4.20)} 1 + c, \end{aligned} \tag{4.23}$$

and thus for some $\tilde{d}_k \in d^*$ it must hold

$$\frac{\varphi_{k+1}(\tilde{d}_k)}{\varphi_k(\tilde{d}_k)} \geq 1 + c.$$

According to Proposition 2.2.1, it holds that

$$\mathsf{D}\left(\varphi_{k+1}(d)||\varphi_k(d)\right) \geq \varphi_{k+1}(\tilde{d}_k) \ln \frac{\varphi_{k+1}(\tilde{d}_k)}{\varphi_k(\tilde{d}_k)} + \left(1 - \varphi_{k+1}(\tilde{d}_k)\right) \ln \frac{1-\varphi_{k+1}(\tilde{d}_k)}{1-\varphi_k(\tilde{d}_k)}. \tag{4.24}$$

From Lemma 4.3.1 applied to (4.24) it follows that for all $k \in \mathbb{N}$ it holds

$$\mathsf{D}\left(\varphi_{k+1}(d)||\varphi_k(d)\right) \geq \varepsilon(1+c) \ln \frac{\varepsilon(1+c)}{\varepsilon} + (1 - \varepsilon(1+c)) \ln \frac{1-\varepsilon(1+c)}{1-\varepsilon}, \tag{4.25}$$

which is positive, as it represents a Kullback Leibler divergence of two non-equal pdfs of a binary random quantity. Because, according to Proposition 4.2.1,

$$\mathcal{D}(\varphi_k) - \mathcal{D}(\varphi_{k+1}) \geq \mathsf{D}\left(\varphi_{k+1}(d)||\varphi_k(d)\right),$$

it follows from (4.25) that $\lim_{k\to+\infty} \mathcal{D}(\varphi_k) = -\infty$, which is in a contradiction with the non-negativity of the function $\mathcal{D}$. □

For discrete random quantities, an implementation of the iterative algorithm based on (4.15) is straightforward. Pdfs of discrete quantities are typically represented by multi-dimensional arrays, the

elements of which are the individual probabilities. The evaluation of $\varphi_{k+1}(d)$ then consist in a recalculation of the array entries according to (4.15). The number of algebraic operations performed in each iteration is proportional to $\mathring{p}\mathring{d}$, where $\mathring{d}$ denotes a cardinality of d^*.

An important point, which has not to be mentioned yet, is a stopping rule. Proposition 4.3.1 says that, under the given assumptions, for an arbitrary initial approximation $\varphi_0(d) \in \mathcal{G}$, an arbitrarily good approximation (in the sense of a value of $\mathcal{D}(\cdot)$) can be acquired by repetitive application of the operator A; however, to this point we are not able to evaluate the quality of the approximation. According to Proposition 4.2.3, it holds that if, for some k, it is fulfilled $A\varphi_k = \varphi_k$, then $\varphi_k(d)$ minimizes $\mathcal{D}(\cdot)$. Nevertheless, in practice one can hardly select an initial approximation $\varphi_0(d)$ so that an optimal solution is found within a finite number of iterations.

To judge a quality of the approximation $\varphi_k(d)$, a lower estimate of $\mathcal{D}({}^I f)$ based on (4.22) can be used. Namely, from (4.22), (4.21), and the definition (4.9) of the operator A it follows that for positive $\varphi_k(d)$ it holds

$$\mathcal{D}({}^I f) \geq \mathcal{D}(\varphi_k) + 1 - \int \frac{A\varphi_k(d)}{\varphi_k(d)} \, {}^I f\,(d)\,\mathrm{d}d. \tag{4.26}$$

The lower estimate of $\mathcal{D}({}^I f)$ is then acquired by replacing $\int \frac{A\varphi_k(d)}{\varphi_k(d)} \, {}^I f\,(d)\,\mathrm{d}d$ in (4.26) by its upper estimate independent of the unknown ${}^I f\,(d)$. For d^* being finite, the simplest estimate is

$$\int \frac{A\varphi_k(d)}{\varphi_k(d)} \, {}^I f\,(d)\,\mathrm{d}d \leq \max_{d \in d^*} \frac{A\varphi_k(d)}{\varphi_k(d)}, \tag{4.27}$$

which gives a lower bound for $\mathcal{D}({}^I f)$:

$$\mathcal{D}({}^I f) \geq \mathcal{D}(\varphi_k) + 1 - \max_{d \in d^*} \frac{A\varphi_k(d)}{\varphi_k(d)}. \tag{4.28}$$

For (4.27) to be a suitable estimate for a stopping rule, it is necessary to show that the right-hand side of (4.28) converges to $\mathcal{D}({}^I f)$. Under the assumptions of Proposition 4.3.1, the convergence is guaranteed by the following proposition.

Proposition 4.3.2 *Suppose that the sequence $(\varphi_k(d))_{k=1}^{+\infty}$ of pdfs defined by (4.15), for some $\varphi_0(d) \in \mathcal{G}$, has the property*

$$\exists \varepsilon > 0, \forall k \in \mathbb{N}, \forall d \in d^*, \varphi_k(d) > \varepsilon.$$

Then, it holds

$$\max_{d \in d^*} \frac{A\varphi_k(d)}{\varphi_k(d)} \to 1. \tag{4.29}$$

Proof: Suppose that (4.29) does not hold. Then, because $\int \varphi_k(d)\mathrm{d}d = \int A\varphi_k(d)\mathrm{d}d = 1$, there exist a strictly increasing sequence $(k_j)_{j=1}^{\infty}, k_j \in \mathbb{N}$, and $c > 0$ so that, for all $j \in \mathbb{N}$,

$$\frac{A\varphi_{k_j}(\tilde{d}_j)}{\varphi_{k_j}(\tilde{d}_j)} \geq 1 + c,$$

for some $\tilde{d}_j \in d^*$. The rest of the proof is an analogy of the proof of Proposition 4.3.1. □

A stopping rule for the recursive evaluation of approximations $\varphi_k(d)$ based on the estimate (4.27) has a form

$$\text{stop if } \max_{d \in d^*} \frac{A\varphi_k(d)}{\varphi_k(d)} - 1 \leq \zeta, \tag{4.30}$$

where $\zeta > 0$ is a predefined threshold specifying a precision of the resulting approximation. If the condition (4.30) is fulfilled for some $\varphi_k(d)$, then, according to (4.28), it hodls that $\mathcal{D}(\varphi_k) - \mathcal{D}({}^I f) \leq \zeta$. Furthermore, Proposition 4.3.2 guarantees that, under the given assumptions, the stopping condition (4.30) is fulfilled within a finite number of iterations.

The estimate (4.27) is too rough for (4.30) to be an efficient stopping rule. A more efficient, but computationally more expensive, stopping rule can be obtained from (4.26) by employing a more accurate estimate of $\int \frac{A\varphi_k(d)}{\varphi_k(d)} {}^{I}f(d)\,\mathrm{d}d$. For example, using (4.11) and the definition (4.9) of the oprator A, we get

$$\begin{aligned}\int \frac{A\varphi_k(d)}{\varphi_k(d)} {}^{I}f(d)\,\mathrm{d}d &= \sum_{p=1}^{\mathring{p}} {}^{p}\alpha \int_{{}^{p}d^*} {}^{p}f({}^{p}d) \int_{\bar{p}d^*} \frac{A\varphi_k({}^{p}d,\bar{p}d)}{\varphi_k({}^{p}d,\bar{p}d)} {}^{I}f\left(\bar{p}d\middle|{}^{p}d\right)\mathrm{d}\bar{p}d\,\mathrm{d}{}^{p}d \\ &\leq \sum_{p=1}^{\mathring{p}} {}^{p}\alpha \int_{{}^{p}d^*} {}^{p}f({}^{p}d) \left(\max_{\bar{p}d\in\bar{p}d^*} \frac{A\varphi_k({}^{p}d,\bar{p}d)}{\varphi_k({}^{p}d,\bar{p}d)}\right)\mathrm{d}{}^{p}d \qquad (4.31)\end{aligned}$$

As it holds that

$$\sum_{p=1}^{\mathring{p}} {}^{p}\alpha \int_{{}^{p}d^*} {}^{p}f({}^{p}d) \left(\max_{\bar{p}d\in\bar{p}d^*} \frac{A\varphi_k({}^{p}d,\bar{p}d)}{\varphi_k({}^{p}d,\bar{p}d)}\right)\mathrm{d}{}^{p}d \leq \max_{d\in d^*} \frac{A\varphi_k(d)}{\varphi_k(d)},$$

the inequality (4.31) can provide a more accurate upper estimate of $\int \frac{A\varphi_k(d)}{\varphi_k(d)} {}^{I}f(d)\,\mathrm{d}d$ then (4.27).

4.3.2 Iterative Algorithm for Continuous Quantities

In applications, continuous random quantities are of a great importance. However, for continuous random quantities, an implementation of the iterative algorithm based on (4.15) is much more difficult.

A proof of convergence of $(\mathcal{D}(\varphi_k))_{k=1}^{+\infty}$ to $\mathcal{D}({}^{I}f)$ cannot be done so easily as in the case of discrete quantities. Namely, if $(\mathcal{D}(\varphi_k))_{k=1}^{+\infty} \to \mathcal{D}({}^{I}f)$ does not hold, then, similarly as in the proof of Proposition 4.3.1, it must hold for some $c>0$, all $k\in\mathbb{N}$, and all ${}^{I}f(d)$ satisfying (4.2) that

$$\int \frac{\varphi_{k+1}(d)}{\varphi_k(d)} {}^{I}f(d)\,\mathrm{d}d \geq 1+c.$$

Then, it must hold, e.g., that sets

$$M_k \equiv \left\{ d\in d^* \,\middle|\, \frac{\varphi_{k+1}(d)}{\varphi_k(d)} \geq 1+\frac{c}{2} \right\}$$

are non-empty. Although it can proved, for example, that M_k satisfy

$$\int_{M_k} {}^{I}f(d)\,\mathrm{d}d \geq \frac{c}{2\left(\mathring{p}-1-\frac{c}{2}\right)},$$

which is positive for $\mathring{p}\geq 2$, it is difficult to find reasonable conditions on the initial approximation $\varphi_0(d)$ and the pdfs ${}^{p}f(d)$ which guarantee that for some $\varepsilon>0$ it holds

$$\int_{M_k} \varphi_k(d)\mathrm{d}d \geq \varepsilon,$$

for all $k\in\mathbb{N}$.

Another problem related to continuous quantities is a form of the pdfs $\varphi_k(d)$. Because the operator A employs both conditioning and mixing operations, it is probably impossible to find a sufficiently rich class of pdfs which can be parameterized by a finite dimensional parameter and is closed with respect to the application of the operator A. A way to overcome this difficulty is to search the approximations of ${}^{I}f(d)$ within a given class of pdfs.

In the rest of this paragraph, we focus on a modification of the iterative algorithm so that the approximations $\varphi_k(d)$ remain in a form of a finite Gaussian mixture. This class of pdfs has been selected because Gaussian mixtures represent quite general approximators and a wide set of learning and design algorithms is available for them; see [27].

First step towards the modified algorithm is a "relaxation" of the relation (4.15) so that $\varphi_{k+1}(d)$ is newly allowed to be searched within a specified set of pdfs: Using the property (2.7) of the Kullback-Leibler divergence, the definition (4.1) of the function $\mathcal{D}$ can be equivalently written in a form

$$\mathcal{D}(\varphi_k(d)) = \sum_{p=1}^{\mathring{p}} {}^p\alpha \mathsf{D}\left({}^pf\left({}^pd\right)\varphi_k({}^{\bar{p}}d|\,{}^pd)\middle|\middle|\varphi_k(d)\right). \tag{4.32}$$

Each term in (4.32) is a Kullback-Leibler divergence of a pdf of the quantity d, namely ${}^pf\left({}^pd\right)\varphi_k({}^{\bar{p}}d|\,{}^pd)$, from $\varphi_k(d)$. Due to Lemma 4.2.1, the new approximation $\varphi_{k+1}(d)$ defined by (4.15) can be interpreted as a pdf closest to the pdfs ${}^pf\left({}^pd\right)\varphi_k({}^{\bar{p}}d|\,{}^pd)$, $p \in p^*$, in the sense that

$$\varphi_{k+1}(d) \in \underset{\varphi\in\mathcal{F}}{\operatorname{argmin}} \sum\nolimits_{p=1}^{\mathring{p}} {}^p\alpha \mathsf{D}\left({}^pf\left({}^pd\right)\varphi_k({}^{\bar{p}}d|\,{}^pd)\middle|\middle|\varphi(d)\right). \tag{4.33}$$

If, instead of (4.33), $\varphi_{k+1}(d)$ is required to satisfy a weaker condition

$$\sum_{p=1}^{\mathring{p}} {}^p\alpha \mathsf{D}\left({}^pf\left({}^pd\right)\varphi_k({}^{\bar{p}}d|\,{}^pd)\middle|\middle|\varphi_{k+1}(d)\right) \leq \sum_{p=1}^{\mathring{p}} {}^p\alpha \mathsf{D}\left({}^pf\left({}^pd\right)\varphi_k({}^{\bar{p}}d|\,{}^pd)\middle|\middle|\varphi_k(d)\right), \tag{4.34}$$

the property $\mathcal{D}(\varphi_{k+1}) \leq \mathcal{D}(\varphi_k)$ remains preserved, because, according to (2.8), it holds

$$\mathcal{D}(\varphi_{k+1}) = \sum_{p=1}^{\mathring{p}} {}^p\alpha \mathsf{D}\left({}^pf\left({}^pd\right)||\varphi_{k+1}({}^pd)\right) \leq \sum_{p=1}^{\mathring{p}} {}^p\alpha \mathsf{D}\left({}^pf\left({}^pd\right)\varphi_k({}^{\bar{p}}d|\,{}^pd)\middle|\middle|\varphi_{k+1}(d)\right). \tag{4.35}$$

The relation (4.35) together with (4.34) and (4.32) then gives the desired inequality $\mathcal{D}(\varphi_{k+1}) \leq \mathcal{D}(\varphi_k)$. The condition (4.34) for $\varphi_{k+1}(d)$ allows to search it within a specified class of pdfs.

The second step is to design a method which, for some class of pdfs $\mathcal{M} \subset \mathcal{F}$ and a pdf $\varphi_k(d) \in \mathcal{M}$, finds $\varphi_{k+1}(d) \in \mathcal{M}$ so that the condition (4.34) is fulfilled and the equality in (4.34) holds, ideally, only if

$$\varphi_k(d) \in \underset{\varphi\in\mathcal{M}}{\operatorname{argmin}} \sum\nolimits_{p=1}^{\mathring{p}} {}^p\alpha \mathsf{D}\left({}^pf\left({}^pd\right)\varphi_k({}^{\bar{p}}d|\,{}^pd)\middle|\middle|\varphi(d)\right). \tag{4.36}$$

For $\mathcal{M}$ being a class of finite Gaussian mixtures, such method can be acquired as a generalization of the well known EM algorithm [17] proposed below. We start with a brief description of the EM algorithm itself.

EM algorithm

Generally speaking, the EM algorithm is a method for finding maximum likelihood (ML) estimates with incomplete observations. Assume that we are to find a ML estimate

$$\hat{\Theta} \in \arg\max_{\Theta\in\Theta^*} f\left(d^{1:\mathring{t}}\middle|\,\Theta\right), \tag{4.37}$$

where $f\left(d^{1:\mathring{t}}\middle|\,\Theta\right) = \prod_{t=1}^{\mathring{t}} f\left(d_t|\,\Theta\right)$. Note that here $d_1, \ldots, d_{\mathring{t}}$ stand for general random quantities, i.e., they are not necessarily related to the system considered in the rest of this chapter. If $f\left(d_t|\,\Theta\right)$ are given as marginal pdfs of joint pdfs

$$f\left(d_t, c_t|\,\Theta\right), \tag{4.38}$$

where c_t are non-observed quantities, the optimization in (4.37) is typically a hard task. Note that c_t can be real missing observations as well as "fictitious" quantities used for definition of $f\left(d_t|\,\Theta\right)$. A typical example of the later case are probabilistic mixtures, i.e., pdfs in a form

$$f\left(d_t|\,\Theta\right) = \sum_{c=1}^{\mathring{c}} \omega_c m(d_t|\theta_c), \tag{4.39}$$

where pdfs $m(d_t|\theta_c)$, called components of the mixture, are from a given parametric class of pdfs and differ by values of their parameters θ_c. ω_c are nonnegative weights, such that $\sum_{c=1}^{c^*} \omega_c = 1$, and $\Theta \equiv (\omega_1, \ldots, \omega_{c^*}, \theta_1, \ldots, \theta_{c^*})$. Probabilistic mixture (4.39) can be taken as a marginal pdf of

$$f\left(d_t, c_t | \Theta\right) = \omega_{c_t} m(d_t|\theta_{c_t}), \tag{4.40}$$

where the quantities c_t, with values in c^*, are interpreted as identifiers of components.

The EM algorithm constructs a sequence of point estimates $\hat{\Theta}^{(i)}$ of the unknown parameter Θ in an iterative way. In i-th iteration, estimate $\hat{\Theta}^{(i)}$ is evaluated from $\hat{\Theta}^{(i-1)}$ in two steps:

E-step: The expected value of the log-likelihood of the parameter Θ, given the observations $d^{1:\mathring{t}}$ and the preceding estimate $\hat{\Theta}^{(i-1)}$, is evaluated as a function of Θ:

$$q^{(i)}(\Theta) \equiv \mathsf{E}\left[\ln f\left(d^{1:\mathring{t}}, c^{1:\mathring{t}} \middle| \Theta\right) \middle| d^{1:\mathring{t}}, \hat{\Theta}^{(i-1)}\right]. \tag{4.41}$$

The expectation in (4.41) is performed with respect to

$$f\left(c^{1:\mathring{t}} \middle| d^{1:\mathring{t}}, \hat{\Theta}^{(i-1)}\right) = \prod_{t=1}^{\mathring{t}} f\left(c_t \middle| d_t, \hat{\Theta}^{(i-1)}\right),$$

where $f\left(c_t \middle| d_t, \hat{\Theta}^{(i-1)}\right) \equiv f\left(c_t | d_t, \Theta\right)|_{\Theta=\hat{\Theta}^{(i-1)}}$ are conditional pdfs acquired from (4.38) for $\Theta = \hat{\Theta}^{(i-1)}$.

M-step: New estimate $\hat{\Theta}^{(i)}$ is selected as a maximizer of $q^{(i)}(\Theta)$ from the E-step

$$\hat{\Theta}^{(i)} \in \arg\max_{\Theta \in \Theta^*} q^{(i)}(\Theta). \tag{4.42}$$

Note that a necessary assumption for the EM algorithm to have a practical asset is that the optimization in (4.42) is easier than the direct evaluation of the ML estimate (4.37). Estimation of parameters of probabilistic mixtures with components from an exponential family meets this assumption.

Two important properties of the EM algorithm should be stressed:

1. EM algorithm converges monotonically in the sense that

$$f\left(d^{1:\mathring{t}} \middle| \hat{\Theta}^{(i)}\right) \geq f\left(d^{1:\mathring{t}} \middle| \hat{\Theta}^{(i-1)}\right).$$

2. EM algorithm does not guarantee a convergence to the global maximum. This drawback is typically treated by repetitious runs of the EM algorithm with different initial estimates $\hat{\Theta}^{(0)}$. For details see, e.g., [23].

Generalized EM algorithm

The above described EM algorithm can be easily generalized so that it can be used for approximating a given pdf. The generalization is based on the following fact: maximum-likelihood estimation is equivalent to the minimization of the Kerridge inaccuracy (2.11) of an empirical pdf and the searched parametric pdf. Indeed, let

$$r(d) \equiv \frac{1}{\mathring{t}} \sum_{t=1}^{\mathring{t}} \delta(d - d_t)$$

be an empirical pdf from independent observations $d^{1:\mathring{t}}$. The log-likelihood $\ln f\left(d^{1:\mathring{t}} \middle| \Theta\right)$ can be expressed in a form

$$\ln f\left(d^{1:\mathring{t}} \middle| \Theta\right) = \mathring{t} \int r(d) \ln(f\left(d | \Theta\right)) \mathrm{d}d = -\mathring{t} \mathsf{K}\left(r(d), f\left(d | \Theta\right)\right),$$

and thus

$$\arg\max_{\Theta\in\Theta^*} f\left(d^{1:\mathring{t}}\middle|\,\Theta\right) = \arg\min_{\Theta\in\Theta^*} \mathsf{K}\left(r(d), f\left(d|\,\Theta\right)\right). \tag{4.43}$$

In this way, the E-step (4.41) of the EM algorithm can be equivalently rewritten as

$$q^{(i)}(\Theta) = \mathring{t}\int r(d) f\left(c\middle|\,d,\hat{\Theta}^{(i-1)}\right)\ln f\left(d,c|\,\Theta\right)\mathrm{d}c\mathrm{d}d = -\mathring{t}\mathsf{K}\left(r(d) f\left(c\middle|\,d,\hat{\Theta}^{(i-1)}\right), f\left(d,c|\,\Theta\right)\right).$$

An iteration of the EM algorithm, i.e., the E-step and M-step together, can be then expressed as

$$\hat{\Theta}^{(i)} \in \arg\min_{\Theta\in\Theta^*} \mathsf{K}\left(r(d) f\left(c\middle|\,d,\hat{\Theta}^{(i-1)}\right), f\left(d,c|\,\Theta\right)\right). \tag{4.44}$$

Notice, that the fact that $r(d)$ is an empirical pdf plays no role in (4.43) and (4.44). On that account, we can expect that the problem of finding

$$\hat{\Theta} \in \arg\min_{\Theta\in\Theta^*} \mathsf{K}\left(h(d), f\left(d|\,\Theta\right)\right), \tag{4.45}$$

where $h(d)$ is an arbitrary pdf of d, can be solved by a generalized EM algorithm, in which the i-th iteration consist in evaluating $\hat{\Theta}^{(i)}$ so that

$$\hat{\Theta}^{(i)} \in \arg\min_{\Theta\in\Theta^*} \mathsf{K}\left(h(d) f\left(c\middle|\,d,\hat{\Theta}^{(i-1)}\right), f\left(d,c|\,\Theta\right)\right). \tag{4.46}$$

The following proposition claims that the monotone convergence of the generalized EM algorithm is preserved.

Proposition 4.3.3 *Consider an arbitrary pdf $h(d)$ and a parametric model $f\left(d|\,\Theta\right)$ given in a form $f\left(d|\,\Theta\right) = \int f\left(d,c|\,\Theta\right)\mathrm{d}c$ for some $f\left(d,c|\,\Theta\right)$, $\Theta\in\Theta^*$. For an arbitrary $\hat{\Theta}^{(i-1)}\in\Theta^*$, such that*

$$\mathsf{K}\left(h(d) f\left(c\middle|\,d,\hat{\Theta}^{(i-1)}\right), f\left(d,c|\,\Theta\right)\right)$$

is finite for some $\Theta\in\Theta^*$, let

$$\hat{\Theta}^{(i)} \in \arg\min_{\Theta\in\Theta^*} \mathsf{K}\left(h(d) f\left(c\middle|\,d,\hat{\Theta}^{(i-1)}\right), f\left(d,c|\,\Theta\right)\right). \tag{4.47}$$

Then, it holds

$$\mathsf{K}\left(h(d), f\left(d\middle|\,\hat{\Theta}^{(i)}\right)\right) \le \mathsf{K}\left(h(d), f\left(d\middle|\,\hat{\Theta}^{(i-1)}\right)\right).$$

Proof: According to the elementary property (2.15) of the Kerridge inaccuracy, it holds

$$\mathsf{K}\left(h(d), f\left(d\middle|\,\hat{\Theta}^{(i-1)}\right)\right) =$$

$$= \mathsf{K}\left(h(d) f\left(c\middle|\,d,\hat{\Theta}^{(i-1)}\right), f\left(d,c\middle|\,\hat{\Theta}^{(i-1)}\right)\right) - \mathsf{K}\left(h(d) f\left(c\middle|\,d,\hat{\Theta}^{(i-1)}\right), f\left(c\middle|\,d,\hat{\Theta}^{(i-1)}\right)\right). \tag{4.48}$$

For the first term in (4.48) it holds, according to (4.47), that

$$\mathsf{K}\left(h(d) f\left(c\middle|\,d,\hat{\Theta}^{(i-1)}\right), f\left(d,c\middle|\,\hat{\Theta}^{(i-1)}\right)\right) \ge \mathsf{K}\left(h(d) f\left(c\middle|\,d,\hat{\Theta}^{(i-1)}\right), f\left(d,c\middle|\,\hat{\Theta}^{(i)}\right)\right).$$

From (2.14) and (2.13) it follows for the second term in (4.48) that

$$\mathsf{K}\left(h(d) f\left(c\middle|\,d,\hat{\Theta}^{(i-1)}\right), f\left(c\middle|\,d,\hat{\Theta}^{(i-1)}\right)\right) \le \mathsf{K}\left(h(d) f\left(c\middle|\,d,\hat{\Theta}^{(i-1)}\right), f\left(c\middle|\,d,\hat{\Theta}^{(i)}\right)\right).$$

Substituting these two inequalities into (4.48) and using (2.15) we get

$$\begin{aligned}
&\mathsf{K}\left(h(d), f\left(d\middle|\hat{\Theta}^{(i-1)}\right)\right) \geq \\
&\quad \geq \mathsf{K}\left(h(d) f\left(c\middle|d,\hat{\Theta}^{(i-1)}\right), f\left(d,c\middle|\hat{\Theta}^{(i)}\right)\right) - \mathsf{K}\left(h(d) f\left(c\middle|d,\hat{\Theta}^{(i-1)}\right), f\left(c\middle|d,\hat{\Theta}^{(i)}\right)\right) = \\
&\quad = \mathsf{K}\left(h(d), f\left(d\middle|\hat{\Theta}^{(i)}\right)\right).
\end{aligned}$$

□

Notes:

- For $h(d)$ being a pdf corresponding to a probability distribution which is absolutely continuous with respect to the probability distributions determined by pdfs $f(d|\Theta)$ for all $\Theta \in \Theta^*$, we get even easier proof of the monotone convergence of the generalized EM algorithm using the relation (2.17) between the Kullback-Leibler divergence and the Kerridge inaccuracy. From the elementary properties of the Kullback Leibler divergence, namely (2.7) and (2.8), and (4.47) we get:

$$\begin{aligned}
\mathsf{D}\left(h(d)\middle\|f\left(d\middle|\hat{\Theta}^{(i-1)}\right)\right) &= \mathsf{D}\left(h(d) f\left(c\middle|d,\hat{\Theta}^{(i-1)}\right)\middle\|f\left(d,c\middle|\hat{\Theta}^{(i-1)}\right)\right) \geq \\
&\geq \mathsf{D}\left(h(d) f\left(c\middle|d,\hat{\Theta}^{(i-1)}\right)\middle\|f\left(d,c\middle|\hat{\Theta}^{(i)}\right)\right) \geq \mathsf{D}\left(h(d)\middle\|f\left(d\middle|\hat{\Theta}^{(i)}\right)\right)
\end{aligned}$$

- In the generalized version of the EM algorithm, the convergence is also guaranteed only to a local extreme.
- Similarly as in the case of the EM algorithm, the generalized EM algorithm has a practical asset only if the minimization (4.46) is technically easier then the direct minimization of (4.45).

Generalized EM algorithm for Gaussian mixtures

A resulting form of the generalized EM algorithm depends on the particular choice of the parametric model $f(d|\Theta)$. As stated in the beginning of this subsection, we focus on parametric models in the form of finite Gaussian mixtures, i.e.,

$$f(d|\Theta) = \sum_{c=1}^{\mathring{c}} \omega_c m(d|\theta_c), \tag{4.49}$$

where

- $\mathring{c} \in \mathbb{N}$, $\omega_1, \ldots, \omega_{\mathring{c}}$ are non-negative weights, such that $\sum_{c=1}^{\mathring{c}} \omega_c = 1$,
- conditional pdfs $m(d|\theta_c)$ – referred to as components – are Gaussian pdfs, i.e.,

$$m(d|\theta_c) \equiv \mathcal{N}_d(\mu_c, \Sigma_c) \equiv \frac{1}{(2\pi)^{\frac{n}{2}} (\det \Sigma_c)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(d-\mu_c)'\Sigma_c^{-1}(d-\mu_c)\right), \tag{4.50}$$

  where n is the dimension of the data vector d, $\mu_c \in \mathbb{R}^n$, $\Sigma_c \in \mathbb{R}^{n,n}$ is positive definite, det denotes determinant, and $\theta_c \equiv (\mu_c, \Sigma_c)$ is a vector of parameters of the c-th component,
- $\Theta \equiv (\omega_1, \ldots, \omega_{\mathring{c}}, \theta_1, \ldots, \theta_{\mathring{c}})$ is a vector of parameters of the parametric model.

Considering the identifier of components c to be a random quantity, the parametric model (4.49) can be taken as a marginal pdf of

$$f(d, c|\Theta) = \omega_c m(d|\theta_c). \tag{4.51}$$

Suppose that a pdf $h(d)$ of a multi-dimensional random quantity $d \equiv (d_1, \ldots, d_n)$, which is to be approximated in the sense of (4.45), fulfills

$$\left| \int d_k d_l h(d_k, d_l) \mathrm{d}d_k \mathrm{d}d_l \right| < +\infty, \tag{4.52}$$

for all $k, l \in \{1, \ldots, n\}$, and that for the $(i-1)$-th point estimate $\hat{\Theta}^{(i-1)} \in \hat{\Theta}$ it holds $\hat{\omega}_c^{(i-1)} > 0$, for all $c \in c^*$. In the i-th iteration of the generalized EM algorithm, given by (4.46), the conditional pdf $f\left(c \middle| d, \hat{\Theta}^{(i-1)}\right)$ is required: from the Bayes rule and (4.51) we get

$$f\left(c \middle| d, \hat{\Theta}^{(i-1)}\right) = \frac{f\left(d, c \middle| \hat{\Theta}^{(i-1)}\right)}{f\left(c \middle| \hat{\Theta}^{(i-1)}\right)} = \frac{\hat{\omega}_c^{(i-1)} m(d|\hat{\theta}_c^{(i-1)})}{\sum_{\tilde{c} \in c^*} \hat{\omega}_{\tilde{c}}^{(i-1)} m(d|\hat{\theta}_{\tilde{c}}^{(i-1)})}. \tag{4.53}$$

Substituting(4.51) and (4.53) into the Kerridge inaccuracy in (4.46), we get

$$\mathsf{K}\left(h(d) f\left(c \middle| d, \hat{\Theta}^{(i-1)}\right), f\left(d, c \middle| \Theta\right)\right) = -\sum_{c \in c^*} \int_{d^*} h(d) \frac{\hat{\omega}_c^{(i-1)} m(d|\hat{\theta}_c^{(i-1)})}{\sum_{\tilde{c} \in c^*} \hat{\omega}_{\tilde{c}}^{(i-1)} m(d|\hat{\theta}_{\tilde{c}}^{(i-1)})} \left(\ln \omega_c + \ln m(d|\theta_c)\right) \mathrm{d}d. \tag{4.54}$$

From (4.54) it is clear that minimization of $\mathsf{K}\left(h(d) f\left(c \middle| d, \hat{\Theta}^{(i-1)}\right), f\left(d, c \middle| \Theta\right)\right)$ can be done separately for ω_c and θ_c. Denoting

$$\xi_c^{(i-1)}(d) \equiv h(d) \frac{\hat{\omega}_c^{(i-1)} m(d|\hat{\theta}_c^{(i-1)})}{\sum_{\tilde{c} \in c^*} \hat{\omega}_{\tilde{c}}^{(i-1)} m(d|\hat{\theta}_{\tilde{c}}^{(i-1)})}$$

and

$$\eta_c^{(i-1)} \equiv \int \xi_c^{(i-1)}(d) \mathrm{d}d,$$

we get from (4.54)

$$\hat{\omega}^{(i)} \in \arg\min_{\omega \in \omega^*} \left(-\sum_{c \in c^*} \eta_c^{(i-1)} \ln \omega_c \right), \tag{4.55}$$

which is fulfilled if, for all $c \in c^*$,

$$\hat{\omega}_c^{(i)} = \eta_c^{(i-1)}, \tag{4.56}$$

because $\eta_c^{(i-1)}$ are nonnegative and $\sum_{c \in c^*} \eta_c^{(i-1)} = 1$. Due to the assumption that $\hat{\omega}_c^{(i-1)} > 0$ for all $c \in c^*$, it holds also $\hat{\omega}_c^{(i)} > 0$.

For $\hat{\theta}_c^{(i)}$ we get

$$\hat{\theta}_c^{(i)} \in \arg\min_{\theta_c \in \theta_c^*} \left(-\int \xi_c^{(i-1)}(d) \ln m(d|\theta_c) \right) \mathrm{d}d,$$

which, for the components $m(d|\theta_c)$ being Gaussian pdfs (4.50), has a form

$$\hat{\theta}_c^{(i)} \in \arg\min_{\theta_c \in \theta_c^*} \left(-\int \xi_c^{(i-1)}(d) \left[-\frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln\left(\det \Sigma_c\right) - \frac{1}{2} (d - \mu_c)' \Sigma_c^{-1} (d - \mu_c) \right] \mathrm{d}d \right). \tag{4.57}$$

Finding $\hat{\theta}_c^{(i)}$ is analogous to a standard procedure of finding a maximum likelihood estimate of parameters of a Gaussian pdf. Here, we partially follow the procedure used in [46].

Let us denote the expression minimized in (4.57), taken as a function of μ_c and Σ_c, as $\mathcal{I}_c^{(i-1)}(\mu_c, \Sigma_c)$. Using a notation

$$\begin{aligned} m_c^{(i-1)} &\equiv \frac{1}{\eta_c^{(i-1)}} \int \xi_c^{(i-1)}(d)\, d \,\mathrm{d}d, \\ Q_c^{(i-1)} &\equiv \frac{1}{\eta_c^{(i-1)}} \int \xi_c^{(i-1)}(d)(d - m_c^{(i-1)})(d - m_c^{(i-1)})' \mathrm{d}d, \end{aligned}$$

$\mathcal{I}_c^{(i-1)}(\mu_c, \Sigma_c)$ can be written as

$$\begin{aligned}
&\mathcal{I}_c^{(i-1)}(\mu_c, \Sigma_c) \\
&\quad = \frac{n}{2}\eta_c^{(i-1)} \ln(2\pi) + \frac{1}{2}\eta_c^{(i-1)} \ln(\det \Sigma_c) + \frac{1}{2}\int \xi_c^{(i-1)}(d)\,(d-\mu_c)'\,\Sigma_c^{-1}\,(d-\mu_c)\,\mathrm{d}d \\
&\quad = \frac{n}{2}\eta_c^{(i-1)} \ln(2\pi) + \frac{1}{2}\eta_c^{(i-1)} \ln(\det \Sigma_c) + \frac{1}{2}\int \xi_c^{(i-1)}(d)\mathrm{tr}\left((d-\mu_c)\,(d-\mu_c)'\,\Sigma_c^{-1}\right)\mathrm{d}d \\
&\quad = \frac{n}{2}\eta_c^{(i-1)} \ln(2\pi) + \frac{1}{2}\eta_c^{(i-1)} \ln(\det \Sigma_c) + \frac{1}{2}\mathrm{tr}\left(\left(\int \xi_c^{(i-1)}(d)\left(d-m_c^{(i-1)}\right)\left(d-m_c^{(i-1)}\right)'\mathrm{d}d\right.\right. \\
&\qquad \left.\left. + \eta_c^{(i-1)}\left(m_c^{(i-1)}-\mu_c\right)\left(m_c^{(i-1)}-\mu_c\right)'\right)\Sigma_c^{-1}\right) \\
&\quad = \frac{n}{2}\eta_c^{(i-1)} \ln(2\pi) + \frac{1}{2}\eta_c^{(i-1)} \ln(\det \Sigma_c) + \frac{1}{2}\eta_c^{(i-1)}\mathrm{tr}\left(Q_c^{(i-1)}\Sigma_c^{-1}\right) \\
&\qquad + \frac{1}{2}\eta_c^{(i-1)}\left(m_c^{(i-1)}-\mu_c\right)'\Sigma_c^{-1}\left(m_c^{(i-1)}-\mu_c\right). \qquad (4.58)
\end{aligned}$$

From (4.58) it follows that for any fixed positive definite matrix $\Sigma_c \in \mathbb{R}^{n,n}$ the function $\mathcal{I}_c^{(i-1)}(\mu_c, \Sigma_c)$ has the global minimum at $\mu_c = m_c^{(i-1)}$, which is independent of Σ_c. Thus, it holds

$$\hat{\mu}_c^{(i)} = \frac{1}{\eta_c^{(i-1)}}\int \xi_c^{(i-1)}(d)\,d\,\mathrm{d}d. \qquad (4.59)$$

To obtain $\hat{\Sigma}_c^{(i)}$, we need to find a Σ_c which minimizes

$$\mathcal{I}_c^{(i-1)}\left(\hat{\mu}_c^{(i)}, \Sigma_c\right) = \frac{n}{2}\eta_c^{(i-1)} \ln(2\pi) + \frac{1}{2}\eta_c^{(i-1)} \ln\left(\det \Sigma_c\right) + \frac{1}{2}\eta_c^{(i-1)}\mathrm{tr}\left(Q_c^{(i-1)}\Sigma_c^{-1}\right). \qquad (4.60)$$

For its partial derivative with respect to Σ_c^{-1}, which is technically more convenient then Σ_c, it holds

$$\frac{\partial}{\partial \Sigma_c^{-1}}\mathcal{I}_c^{(i-1)}\left(\hat{\mu}_c^{(i)}, \Sigma_c\right) = -\frac{1}{2}\eta_c^{(i-1)}\Sigma_c' + \frac{1}{2}\eta_c^{(i-1)}\left(Q_c^{(i-1)}\right)', \qquad (4.61)$$

see, e.g., [42]. The derivative (4.61) is equal to 0 for $\Sigma_c = Q_c^{(i-1)}$. To show that (4.60) has its minimum at $Q_c^{(i-1)}$, we prove that

$$\mathcal{I}_c^{(i-1)}\left(\hat{\mu}_c^{(i)}, \Sigma_c\right) - \mathcal{I}_c^{(i-1)}\left(\hat{\mu}_c^{(i)}, Q_c^{(i-1)}\right)$$

is nonnegative for all positive definite $\Sigma_c \in \mathbb{R}^{n,n}$. Substituting $\Sigma_c = Q_c^{(i-1)}$ into (4.60), we get

$$\mathcal{I}_c^{(i-1)}\left(\hat{\mu}_c^{(i)}, Q_c^{(i-1)}\right) = \frac{n}{2}\eta_c^{(i-1)} \ln(2\pi) + \frac{1}{2}\eta_c^{(i-1)} \ln\left(\det Q_c^{(i-1)}\right) + \frac{1}{2}\eta_c^{(i-1)} n. \qquad (4.62)$$

The difference of (4.60) and (4.62) is then

$$\begin{aligned}
&\mathcal{I}_c^{(i-1)}\left(\hat{\mu}_c^{(i)}, \Sigma_c\right) - \mathcal{I}_c^{(i-1)}\left(\hat{\mu}_c^{(i)}, Q_c^{(i-1)}\right) \\
&\quad = \frac{1}{2}\eta_c^{(i-1)}\left(\ln\left(\det \Sigma_c\right) - \ln\left(\det Q_c^{(i-1)}\right)\right) - \frac{1}{2}\eta_c^{(i-1)} n + \frac{1}{2}\eta_c^{(i-1)}\mathrm{tr}\left(Q_c^{(i-1)}\Sigma_c^{-1}\right) \\
&\quad = \frac{1}{2}\eta_c^{(i-1)}\left(-\ln\left(\det\left(Q_c^{(i-1)}\Sigma_c^{-1}\right)\right) - n + \mathrm{tr}\left(Q_c^{(i-1)}\Sigma_c^{-1}\right)\right) \\
&\quad = \frac{1}{2}\eta_c^{(i-1)}\left(-\ln\left(\lambda_1 \cdots \lambda_n\right) - n + \left(\lambda_1 + \cdots + \lambda_n\right)\right), \qquad (4.63)
\end{aligned}$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of the matrix $Q_c^{(i-1)}\Sigma_c^{-1}$. As $x - 1 - \ln x \geq 0$ for all $x \geq 0$, and $\lambda_1, \ldots, \lambda_n$ are positive, because $Q_c^{(i-1)}\Sigma_c^{-1}$ is positive definite, it holds that

$$\sum_{i=1}^{n}\left(\lambda_i - 1 - \ln \lambda_i\right) \geq 0. \qquad (4.64)$$

From (4.64) it follows that (4.63) is nonnegative, and thus it holds

$$\hat{\Sigma}_c^{(i)} = \frac{1}{\eta_c^{(i-1)}} \int \xi_c^{(i-1)}(d)(d - \hat{\mu}_c^{(i)})(d - \hat{\mu}_c^{(i)})' \mathrm{d}d. \tag{4.65}$$

Due to the assumption (4.52), it is guaranteed that $\hat{\mu}_c^{(i)}$ and $\hat{\Sigma}_c^{(i)}$ are finite for all $c \in c^*$.

Iterative algorithm for Gaussian mixtures

Using Proposition 4.3.3, and relations (4.55), (4.59), and (4.65), a basic version of an algorithm approximating ${}^{I}f\,(d)$, defined by 4.2, by a finite Gaussian mixture can be designed. In the algorithm we employ projective operators ${}^{p}S$ previously introduced in Section 3.5. Namely, ${}^{p}S : \mathbb{R}^n \to \mathbb{R}^{{}^{p}\mathring{i}}$ is defined for $x \equiv (x_1, \ldots, x_n)$ by

$${}^{p}S(x) = (x_i)_{i \in {}^{p}i^*}\,. \tag{4.66}$$

Analogously we define operators ${}^{pp}S : \mathbb{R}^{n,n} \to \mathbb{R}^{{}^{p}\mathring{i},\,{}^{p}\mathring{i}}$ defined for $X \equiv (x_{ij})_{i,j=1}^{n}$ by

$${}^{pp}S(X) = (x_{ij})_{i \in {}^{p}i^*,\, j \in {}^{p}i^*}\,. \tag{4.67}$$

With the operators ${}^{p}S$ and ${}^{pp}S$, the parameters of marginal pdfs of a Gaussian pdf can be easily expressed; see, e.g., [46]. Namely, for $\mathcal{N}_d(\mu, \Sigma)$ and any $p \in p^*$ it holds

$$\int \mathcal{N}_d(\mu, \Sigma) \mathrm{d}\,{}^{\bar{p}}d = \mathcal{N}_{{}^{p}d}\left({}^{p}\mu,\, {}^{p}\Sigma\right),$$

where

$${}^{p}\mu = {}^{p}S(\mu), \quad {}^{p}\Sigma = {}^{pp}S(\Sigma). \tag{4.68}$$

Algorithm 4.3.1 *(Approximation of ${}^{I}f(d)$ by a Gaussian mixture)*
Let $\mathring{p}$, ${}^{p}d$, ${}^{p}f\left({}^{p}d\right)$, ${}^{p}\alpha$ be given – see Section 4.1.

1. *Select parameters $\mathring{c}, I, K \in \mathbb{N}$, where*
   - *$\mathring{c}$ is a number of components in the approximations $\varphi_k(d)$,*
   - *I is a number of iterations of the generalized EM algorithm,*
   - *K is a number of iterations of the approximating step.* (4.34)
2. *For all $c \in c^*$, select parameters $\hat{\omega}_{c,0} > 0, \hat{\mu}_{c,0} \in \mathbb{R}^n, \hat{\Sigma}_{c,0} \in \mathbb{R}^{n,n}$ of the initial approximation $\varphi_0(d)$ of ${}^{I}f\,(d)$, i.e.,*
$$\varphi_0(d) = \sum_{c=1}^{\mathring{c}} \hat{\omega}_{c,0} \mathcal{N}_d\left(\hat{\mu}_{c,0}, \hat{\Sigma}_{c,0}\right).$$
3. *$k := 1$ (counter of approximating steps* (4.34)*). Note that the symbol $:=$ denotes an assignment.*
4. *For all $p \in p^*, c \in c^*$, set parameters*
$$\begin{aligned} {}^{p}\hat{\mu}_{c,k-1} &:= {}^{p}S\left(\hat{\mu}_{c,k-1}\right), \\ {}^{p}\hat{\Sigma}_{c,k-1} &:= {}^{pp}S\left(\hat{\Sigma}_{c,k-1}\right). \end{aligned}$$
5. *Set*
$$h_k(d) := A\varphi_{k-1} = \sum_{p=1}^{\mathring{p}} {}^{p}\alpha\, {}^{p}f\left({}^{p}d\right) \frac{\sum_{c=1}^{\mathring{c}} \hat{\omega}_{c,k-1} \mathcal{N}_d\left(\hat{\mu}_{c,k-1}, \hat{\Sigma}_{c,k-1}\right)}{\sum_{c=1}^{\mathring{c}} \hat{\omega}_{c,k-1} \mathcal{N}_{{}^{p}d}\left({}^{p}\hat{\mu}_{c,k-1},\, {}^{p}\hat{\Sigma}_{c,k-1}\right)}.$$

6. *For all $c \in c^*$, set parameters*

$$\begin{aligned}\hat{\omega}_{c,k}^{(0)} &:= \hat{\omega}_{c,k-1},\\ \hat{\mu}_{c,k}^{(0)} &:= \hat{\mu}_{c,k-1},\\ \hat{\Sigma}_{c,k}^{(0)} &:= \hat{\Sigma}_{c,k-1}\end{aligned}$$

of the initial approximation $\varphi_k^{(0)}(d)$ of the pdf $h_k(d)$, i.e., $\varphi_k^{(0)}(d) = \varphi_{k-1}(d)$.

7. *$i := 1$ (counter of EM iterations)*

8. *For all $c \in c^*$, set*

$$\xi_{c,k}^{(i-1)}(d) := h_k(d)\frac{\hat{\omega}_{c,k}^{(i-1)}\mathcal{N}_d\left(\hat{\mu}_{c,k}^{(i-1)}, \hat{\Sigma}_{c,k}^{(i-1)}\right)}{\sum_{\tilde{c}=1}^{\mathring{c}}\hat{\omega}_{\tilde{c},k}^{(i-1)}\mathcal{N}_d\left(\hat{\mu}_{\tilde{c},k}^{(i-1)}, \hat{\Sigma}_{\tilde{c},k}^{(i-1)}\right)}.$$

9. *For all $c \in c^*$, set parameters of the i-th approximation $\varphi_k^{(i)}(d)$ of the pdf $h_k(d)$:*

$$\begin{aligned}\hat{\omega}_{c,k}^{(i)} &:= \int \xi_{c,k}^{(i-1)}(d)\,\mathrm{d}d,\\ \hat{\mu}_{c,k}^{(i)} &:= \frac{1}{\hat{\omega}_{c,k}^{(i)}}\int \xi_{c,k}^{(i-1)}(d)\,d\,\mathrm{d}d,\\ \hat{\Sigma}_{c,k}^{(i)} &:= \frac{1}{\hat{\omega}_{c,k}^{(i)}}\int \xi_{c,k}^{(i-1)}(d)\left(d-\hat{\mu}_{c,k}^{(i)}\right)\left(d-\hat{\mu}_{c,k}^{(i)}\right)'\mathrm{d}d.\end{aligned}$$

10. *$i := i + 1$; if $i \leq I$, then go to step 8.*

11. *For all $c \in c^*$, set parameters*

$$\begin{aligned}\hat{\omega}_{c,k} &:= \hat{\omega}_{c,k}^{(I)},\\ \hat{\mu}_{c,k} &:= \hat{\mu}_{c,k}^{(I)},\\ \hat{\Sigma}_{c,k} &:= \hat{\Sigma}_{c,k}^{(I)}\end{aligned}$$

of the k-th approximation $\varphi_k(d)$ of ${}^{I}f\,(d)$.

12. *$k := k + 1$; If $k \leq K$, then go to step 4.*

The result of Algorithm 4.3.1 are the parameters $\hat{\omega}_{c,K}, \hat{\mu}_{c,K}, \hat{\Sigma}_{c,K}$ of the best achieved approximation

$$\varphi_K(d) = \sum_{c=1}^{\mathring{c}}\hat{\omega}_{c,K}\mathcal{N}_d\left(\hat{\mu}_{c,K}, \hat{\Sigma}_{c,K}\right)$$

of ${}^{I}f\,(d)$.

The above described basic version of the algorithm is far from being directly implementable. However, it clearly reflects the key point which consit in the combination of the iterative algorithm based on (4.15) and the generalized EM algorithm.

Comments on Algorithm 4.3.1:

- All functions introduced in the algorithm are fully characterized by finite-dimensional parameters, which is important for an implemntation of the algorithm.
- A numerical integration has to be employed in step 9. On that account, the applicability of the algorithm is limited by the dimension of the data d.

- In the presented basic version of the algorithm, the number of the approximating steps (4.34) K as well as the number of the iterations of the generalized EM algorithm I, are fixed. Instead, some kind of stopping rules for cycles over k and i should be employed.

- The number of components $\mathring{c}$ in the approximations $\varphi_k(d)$ also need not be given in advance. It should be rather selected dynamically, according to the evolution of the approximations $\varphi_k(d)$.

- To show that the approximations $\varphi_k(d)$ produced by Algorithm 4.3.1 converges monotonically, i.e., that $\mathcal{D}(\varphi_k) \leq \mathcal{D}(\varphi_{k-1})$, for all $k \in \{1, \ldots, K\}$, it remains to prove that $\varphi_1(d), \ldots, \varphi_K(d)$ satisfy the condition (4.34). Indeed, due to properties (2.17) and (2.12) of the Kerridge inaccuracy, it holds

$$\begin{aligned}
&\sum_{p=1}^{\mathring{p}} {}^{p}\alpha \mathsf{D}\left({}^{p}f\left({}^{p}d\right)\varphi_{k-1}({}^{\bar{p}}d|\,{}^{p}d)\middle|\middle|\varphi_k(d)\right) \\
&\qquad = -\left(\sum_{p=1}^{\mathring{p}} {}^{p}\alpha \mathsf{H}\left({}^{p}f\left({}^{p}d\right)\varphi_{k-1}({}^{\bar{p}}d|\,{}^{p}d)\right)\right) + \mathsf{K}\left(\sum_{p=1}^{\mathring{p}} {}^{p}\alpha\, {}^{p}f\left({}^{p}d\right)\varphi_{k-1}({}^{\bar{p}}d|\,{}^{p}d), \varphi_k(d)\right). \quad (4.69)
\end{aligned}$$

  The choice $\varphi_k^{(0)}(d) = \varphi_{k-1}(d)$ in the step (6) and repetitive use of Proposition 4.3.3 imply that

$$\mathsf{K}\left(\sum_{p=1}^{\mathring{p}} {}^{p}\alpha\, {}^{p}f\left({}^{p}d\right)\varphi_{k-1}({}^{\bar{p}}d|\,{}^{p}d), \varphi_k(d)\right) \leq \mathsf{K}\left(\sum_{p=1}^{\mathring{p}} {}^{p}\alpha\, {}^{p}f\left({}^{p}d\right)\varphi_{k-1}({}^{\bar{p}}d|\,{}^{p}d), \varphi_{k-1}(d)\right). \quad (4.70)$$

  Using (4.69), (4.70), and (2.17), we get the desired inequality

$$\sum_{p=1}^{\mathring{p}} {}^{p}\alpha \mathsf{D}\left({}^{p}f\left({}^{p}d\right)\varphi_{k-1}({}^{\bar{p}}d|\,{}^{p}d)\middle|\middle|\varphi_k(d)\right) \leq \sum_{p=1}^{\mathring{p}} {}^{p}\alpha \mathsf{D}\left({}^{p}f\left({}^{p}d\right)\varphi_{k-1}({}^{\bar{p}}d|\,{}^{p}d)\middle|\middle|\varphi_{k-1}(d)\right).$$

- Although it has been proved that, for all $k \in \{1, \ldots, K\}$, it holds $\mathcal{D}(\varphi_k) \leq \mathcal{D}(\varphi_{k-1})$, other convergence properties of Algorithm 4.3.1 are subject of further research and practical experiments.

- Selection of the parameters of the initial approximation $\varphi_0(d)$, in step 2, is also a subject of further experiments. The parameters selected so that $\varphi_0(d)$ is rather flat and its components are mutually sufficiently different, together with uniformly distributed weights $\hat{\omega}_{c,0}$, could serve as a suitable starting point.

A computationaly more efficient version of Algorithm 4.3.1 can be achieved by setting the number of iterations of the generalized EM algorithm $I := 1$, i.e., each application of the operator A is followed by one step of the generalized EM algorithm. Because $\hat{\omega}_{c,k}^{(0)} = \hat{\omega}_{c,k-1}$, $\hat{\mu}_{c,k}^{(0)} = \hat{\mu}_{c,k-1}$, and $\hat{\Sigma}_{c,k-1}^{(0)} = \hat{\Sigma}_{c,k}$ (step 6 of Algorithm 4.3.1), we get for $\xi_{c,k}^{(0)}(d)$ (step 8)

$$\xi_{c,k}^{(0)}(d) = \hat{\omega}_{c,k-1}\mathcal{N}_d\left(\hat{\mu}_{c,k-1}, \hat{\Sigma}_{c,k-1}\right) \sum_{p=1}^{\mathring{p}} {}^{p}\alpha \frac{{}^{p}f\left({}^{p}d\right)}{\sum_{c=1}^{\mathring{c}} \hat{\omega}_{c,k-1}\mathcal{N}_{{}^{p}d}\left({}^{p}\hat{\mu}_{c,k-1}, {}^{p}\hat{\Sigma}_{c,k-1}\right)}. \quad (4.71)$$

Notice, that each term in the sum in (4.71) depends only on ${}^{p}d$. This fact simplifies the integration in step 9 of Algorithm 4.3.1, because marginal and conditional pdfs of $\mathcal{N}_d\left(\hat{\mu}_{c,k-1}, \hat{\Sigma}_{c,k-1}\right)$ and their moments can be found analytically.

In order to evaluate parameters of conditional pdfs of a Gaussian pdf, we employ operators ${}^{\bar{p}}S$, ${}^{\bar{p}p}S$, ${}^{p\bar{p}}S$, and ${}^{\bar{p}\bar{p}}S$ in addition to the operators ${}^{p}S$ and ${}^{pp}S$ defined by (4.66) and (4.67). ${}^{\bar{p}}S : \mathbb{R}^n \to \mathbb{R}^{n-{}^{\mathring{p}_i}}$ is defined for $x \equiv (x_1, \ldots, x_n)$ by

$${}^{\bar{p}}S(x) = (x_i)_{i \in \{1,\ldots,n\} \setminus {}^{p}i^*}.$$

${}^{\bar{p}p}S : \mathbb{R}^{n,n} \to \mathbb{R}^{n-{}^{\mathring{p}_i}, {}^{\mathring{p}_i}}$ is defined for $X \equiv (x_{ij})_{i,j=1}^{n}$ by

$${}^{\bar{p}p}S(X) = (x_{i,j})_{i \in \{1,\ldots,n\} \setminus {}^{p}i^*, j \in {}^{p}i^*}.$$

Operators ${}^{p}\bar{p}S : \mathbb{R}^{n,n} \to \mathbb{R}^{{}^{p}_{i},n-{}^{p}_{i}}$ and ${}^{\bar{p}}p\bar{p}S : \mathbb{R}^{n,n} \to \mathbb{R}^{n-{}^{p}_{i},n-{}^{p}_{i}}$ are defined analogously to ${}^{\bar{p}}pS$. Using the introduced operators, we can easily express conditional pdfs of ${}^{\bar{p}}d$ given ${}^{p}d$ for $f(d) \equiv \mathcal{N}_d(\mu, \Sigma)$, see, e.g., [46]. For all $p \in p^*$ it holds

$$f\left({}^{\bar{p}}d \middle|\, {}^{p}d\right) = \mathcal{N}_{{}^{\bar{p}}d}\left({}^{\bar{p}}\mu + {}^{\bar{p}p}\Sigma\left({}^{p}\Sigma\right)^{-1}\left({}^{p}d - {}^{p}\mu\right), {}^{\bar{p}}\Sigma - {}^{\bar{p}p}\Sigma\left({}^{p}\Sigma\right)^{-1}\,{}^{p\bar{p}}\Sigma\right), \tag{4.72}$$

where

$${}^{\bar{p}}\mu = {}^{\bar{p}}S(\mu),\;\; {}^{\bar{p}p}\Sigma = {}^{\bar{p}p}S(\Sigma),\;\; {}^{\bar{p}}\Sigma = {}^{\bar{p}\bar{p}}S(\Sigma),\;\; {}^{p\bar{p}}\Sigma = \left({}^{\bar{p}p}\Sigma\right)'.$$

${}^{p}\mu$ and ${}^{p}\Sigma$ are already introduced by (4.68). For the mean, taken as a function of ${}^{p}d$, and the covariance matrix of (4.72) we use a short notation

$$\begin{aligned} {}^{\bar{p}|p}\mu({}^{p}d) &\equiv {}^{\bar{p}}\mu + {}^{\bar{p}p}\Sigma\left({}^{p}\Sigma\right)^{-1}\left({}^{p}d - {}^{p}\mu\right), \\ {}^{\bar{p}|p}\Sigma &\equiv {}^{\bar{p}}\Sigma - {}^{\bar{p}p}\Sigma\left({}^{p}\Sigma\right)^{-1}\,{}^{p\bar{p}}\Sigma. \end{aligned}$$

Denoting

$$\varrho_{p,k}({}^{p}d) \equiv \frac{{}^{p}f\left({}^{p}d\right)}{\sum_{c=1}^{\mathring{c}} \hat{\omega}_{c,k-1}\mathcal{N}_{{}^{p}d}\left({}^{p}\hat{\mu}_{c,k-1},\, {}^{p}\hat{\Sigma}_{c,k-1}\right)},$$

we get for the weights $\hat{\omega}_{c,k}$ the following relation:

$$\hat{\omega}_{c,k} = \hat{\omega}_{c,k-1}\int \mathcal{N}_d\left(\hat{\mu}_{c,k-1}, \hat{\Sigma}_{c,k-1}\right)\sum_{p=1}^{\mathring{p}} {}^{p}\alpha\varrho_{p,k}({}^{p}d)\,\mathrm{d}d = \hat{\omega}_{c,k-1}\sum_{p=1}^{\mathring{p}} {}^{p}\alpha\lambda_{c,p,k},$$

where

$$\lambda_{c,p,k} \equiv \int \mathcal{N}_{{}^{p}d}\left({}^{p}\hat{\mu}_{c,k-1},\, {}^{p}\hat{\Sigma}_{c,k-1}\right)\varrho_{p,k}({}^{p}d)\,\mathrm{d}\,{}^{p}d. \tag{4.73}$$

For the means $\hat{\mu}_{c,k}$ we get

$$\hat{\mu}_{c,k} = \frac{\hat{\omega}_{c,k-1}}{\hat{\omega}_{c,k}}\int \mathcal{N}_d\left(\hat{\mu}_{c,k-1}, \hat{\Sigma}_{c,k-1}\right)\sum_{p=1}^{\mathring{p}} {}^{p}\alpha\varrho_{p,k}({}^{p}d)\,d\,\mathrm{d}d = \frac{\hat{\omega}_{c,k-1}}{\hat{\omega}_{c,k}}\sum_{p=1}^{\mathring{p}} {}^{p}\alpha\nu_{c,p,k},$$

where for ${}^{p}\nu_{c,p,k} \equiv {}^{p}S(\nu_{c,p,k})$ and ${}^{\bar{p}}\nu_{c,p,k} \equiv {}^{\bar{p}}S(\nu_{c,p,k})$ it holds

$$\begin{aligned} {}^{p}\nu_{c,p,k} &\equiv \int \mathcal{N}_{{}^{p}d}\left({}^{p}\hat{\mu}_{c,k-1},\, {}^{p}\hat{\Sigma}_{c,k-1}\right)\varrho_{p,k}({}^{p}d)\;{}^{p}d\;\mathrm{d}\,{}^{p}d, \\ {}^{\bar{p}}\nu_{c,p,k} &\equiv \int \mathcal{N}_d\left(\hat{\mu}_{c,k-1}, \hat{\Sigma}_{c,k-1}\right)\varrho_{p,k}({}^{p}d)\;{}^{\bar{p}}d\;\mathrm{d}d = \\ &= \int \mathcal{N}_{{}^{p}d}\left({}^{p}\hat{\mu}_{c,k-1},\, {}^{p}\hat{\Sigma}_{c,k-1}\right)\varrho_{p,k}({}^{p}d)\,{}^{\bar{p}|p}\hat{\mu}_{c,k-1}({}^{p}d)\,\mathrm{d}\,{}^{p}d. \end{aligned}$$

Finally, new relations for the covariance matrices $\hat{\Sigma}_{c,k}$ are

$$\hat{\Sigma}_{c,k} = \frac{\hat{\omega}_{c,k-1}}{\hat{\omega}_{c,k}}\int \mathcal{N}_d\left(\hat{\mu}_{c,k-1}, \hat{\Sigma}_{c,k-1}\right)\sum_{p=1}^{\mathring{p}} {}^{p}\alpha\varrho_{p,k}({}^{p}d)\,(d-\hat{\mu}_{c,k})\,(d-\hat{\mu}_{c,k})'\,\mathrm{d}d = \frac{\hat{\omega}_{c,k-1}}{\hat{\omega}_{c,k}}\sum_{p=1}^{\mathring{p}} {}^{p}\alpha\Omega_{c,p,k},$$

where ${}^{p}\Omega_{c,k,p} \equiv {}^{pp}S\left(\Omega_{c,k,p}\right)$, ${}^{p\bar{p}}\Omega_{c,k,p} = \left({}^{\bar{p}p}\Omega_{c,k,p}\right)' \equiv {}^{p\bar{p}}S\left(\Omega_{c,k,p}\right)$, and ${}^{\bar{p}}\Omega_{c,k,p} \equiv {}^{\bar{p}\bar{p}}S\left(\Omega_{c,k,p}\right)$ are given by the following relations:

$$
\begin{aligned}
{}^{p}\Omega_{c,p,k} &= \int \mathcal{N}_{{}^{p}d}\left({}^{p}\hat{\mu}_{c,k-1},\, {}^{p}\hat{\Sigma}_{c,k-1}\right) \varrho_{p,k}({}^{p}d)\,({}^{p}d - {}^{p}\hat{\mu}_{c,k})\,({}^{p}d - {}^{p}\hat{\mu}_{c,k})'\,\mathrm{d}\,{}^{p}d \\
&= \int \mathcal{N}_{{}^{p}d}\left({}^{p}\hat{\mu}_{c,k-1},\, {}^{p}\hat{\Sigma}_{c,k-1}\right) \varrho_{p,k}({}^{p}d)\;{}^{p}d\;{}^{p}d'\,\mathrm{d}\,{}^{p}d - {}^{p}\nu_{c,p,k}\,{}^{p}\hat{\mu}'_{c,k} - {}^{p}\hat{\mu}_{c,k}\,{}^{p}\nu'_{c,p,k} + \lambda_{c,k,p}\,{}^{p}\hat{\mu}_{c,k}\,{}^{p}\hat{\mu}'_{c,k} \\
{}^{p\bar{p}}\Omega_{c,p,k} &= \int \mathcal{N}_{d}\left(\hat{\mu}_{c,k-1}, \hat{\Sigma}_{c,k-1}\right) \varrho_{p,k}({}^{p}d)\,({}^{p}d - {}^{p}\hat{\mu}_{c,k})\left(\bar{}^{p}d - \bar{}^{p}\hat{\mu}_{c,k}\right)'\,\mathrm{d}d \\
&= \int \mathcal{N}_{{}^{p}d}\left({}^{p}\hat{\mu}_{c,k-1},\, {}^{p}\hat{\Sigma}_{c,k-1}\right) \varrho_{p,k}({}^{p}d)\,({}^{p}d - {}^{p}\hat{\mu}_{c,k})\left({}^{\bar{p}|p}\hat{\mu}_{c,k-1}({}^{p}d) - \bar{}^{p}\hat{\mu}_{c,k}\right)'\,\mathrm{d}\,{}^{p}d \\
&= \int \mathcal{N}_{{}^{p}d}\left({}^{p}\hat{\mu}_{c,k-1},\, {}^{p}\hat{\Sigma}_{c,k-1}\right) \varrho_{p,k}({}^{p}d)\,({}^{p}d - {}^{p}\hat{\mu}_{c,k})\;{}^{\bar{p}|p}\hat{\mu}'_{c,k-1}({}^{p}d)\,\mathrm{d}\,{}^{p}d \\
&\quad -\left({}^{p}\nu_{c,p,k} - \lambda_{c,k,p}\,{}^{p}\hat{\mu}_{c,k}\right)\bar{}^{p}\hat{\mu}'_{c,k} \\
{}^{\bar{p}}\Omega_{c,p,k} &= \int \mathcal{N}_{d}\left(\hat{\mu}_{c,k-1}, \hat{\Sigma}_{c,k-1}\right) \varrho_{p,k}({}^{p}d)\left(\bar{}^{p}d - \bar{}^{p}\hat{\mu}_{c,k}\right)\left(\bar{}^{p}d - \bar{}^{p}\hat{\mu}_{c,k}\right)'\,\mathrm{d}d \\
&= \int \mathcal{N}_{{}^{p}d}\left({}^{p}\hat{\mu}_{c,k-1},\, {}^{p}\hat{\Sigma}_{c,k-1}\right) \varrho_{p,k}({}^{p}d)\left({}^{\bar{p}|p}\Sigma_{c,k-1} + {}^{\bar{p}|p}\hat{\mu}_{c,k-1}({}^{p}d)\,{}^{\bar{p}|p}\hat{\mu}'_{c,k-1}({}^{p}d)\right. \\
&\quad \left. - {}^{\bar{p}|p}\hat{\mu}_{c,k-1}({}^{p}d)\,\bar{}^{p}\hat{\mu}'_{c,k} - \bar{}^{p}\hat{\mu}_{c,k}\,{}^{\bar{p}|p}\hat{\mu}'_{c,k-1}({}^{p}d) + \bar{}^{p}\hat{\mu}_{c,k}\,\bar{}^{p}\hat{\mu}'_{c,k}\right)\mathrm{d}\,{}^{p}d \\
&= \int \mathcal{N}_{{}^{p}d}\left({}^{p}\hat{\mu}_{c,k-1},\, {}^{p}\hat{\Sigma}_{c,k-1}\right) \varrho_{p,k}({}^{p}d)\left({}^{\bar{p}|p}\hat{\mu}_{c,k-1}({}^{p}d)\,{}^{\bar{p}|p}\hat{\mu}'_{c,k-1}({}^{p}d) - {}^{\bar{p}|p}\hat{\mu}_{c,k-1}({}^{p}d)\,\bar{}^{p}\hat{\mu}'_{c,k}\right. \\
&\quad \left. - \bar{}^{p}\hat{\mu}_{c,k}\,{}^{\bar{p}|p}\hat{\mu}'_{c,k-1}({}^{p}d)\right)\mathrm{d}\,{}^{p}d + \lambda_{c,k,p}\left({}^{\bar{p}|p}\Sigma_{c,k-1} + \bar{}^{p}\hat{\mu}_{c,k}\,\bar{}^{p}\hat{\mu}'_{c,k}\right)
\end{aligned}
$$

The outlined modification is summarized by the following algorithm.

Algorithm 4.3.2 *(Approximation of ${}^{I}f(d)$ by a Gaussian mixture with one iteration of the generalized EM algorithm)*
Let $\mathring{p}$, ${}^{p}d$, ${}^{p}f\,({}^{p}d)$, ${}^{p}\alpha$ be given – see Section 4.1.

1. *Select parameters $\mathring{c}, K \in \mathbb{N}$, where*
   - *$\mathring{c}$ is a number of components in approximations $\varphi_k(d)$,*
   - *K is a number of iterations of the approximating step* (4.34).
2. *For all $c \in c^*$, select parameters $\hat{\omega}_{c,0} > 0, \hat{\mu}_{c,0} \in \mathbb{R}^n, \hat{\Sigma}_{c,0} \in \mathbb{R}^{n,n}$ of the initial approximation $\varphi_0(d)$ of ${}^{I}f\,(d)$, i.e.,*
$$\varphi_0(d) = \sum_{c=1}^{\mathring{c}} \hat{\omega}_{c,0}\mathcal{N}_d\left(\hat{\mu}_{c,0}, \hat{\Sigma}_{c,0}\right).$$
3. *$k := 1$ (counter of approximating steps* (4.34)*)*
4. *For all $p \in p^*, c \in c^*$, set parameters*
$$
\begin{aligned}
{}^{p}\hat{\mu}_{c,k-1} &:= {}^{p}S(\hat{\mu}_{c,k-1}), \\
\bar{}^{p}\hat{\mu}_{c,k-1} &:= \bar{}^{p}S\left(\hat{\mu}_{c,k-1}\right), \\
{}^{p}\hat{\Sigma}_{c,k-1} &:= {}^{pp}S\left(\hat{\Sigma}_{c,k-1}\right), \\
{}^{\bar{p}|p}\hat{\Sigma}_{c,k-1} &:= {}^{\bar{p}\bar{p}}S\left(\hat{\Sigma}_{c,k-1}\right) - {}^{\bar{p}p}S\left(\hat{\Sigma}_{c,k-1}\right)\,{}^{p}\hat{\Sigma}^{-1}_{c,k-1}\,{}^{p\bar{p}}S\left(\hat{\Sigma}_{c,k-1}\right),
\end{aligned}
$$
   and function ${}^{\bar{p}|p}\hat{\mu}_{c,k-1} : \mathbb{R}^{{}^{p}i} \to \mathbb{R}^{n-\bar{}^{p}i}$
$${}^{\bar{p}|p}\hat{\mu}_{c,k-1}({}^{p}d) := \bar{}^{p}\hat{\mu}_{c,k-1} + {}^{\bar{p}p}S\left(\hat{\Sigma}_{c,k-1}\right)\,{}^{p}\hat{\Sigma}^{-1}_{c,k-1}\left({}^{p}d - {}^{p}\hat{\mu}_{c,k-1}\right).$$

5. *For all* $p \in p^*$, *set*

$$\varrho_{p,k}({}^{p}d) := \frac{{}^{p}f\,({}^{p}d)}{\sum_{c=1}^{\mathring{c}} \hat{\omega}_{c,k-1}\mathcal{N}_{{}^{p}d}\left({}^{p}\hat{\mu}_{c,k-1},\, {}^{p}\hat{\Sigma}_{c,k-1}\right)}.$$

6. *For all* $c \in c^*, p \in p^*$, *set*

$$\begin{aligned}
\lambda_{c,p,k} &:= \int \mathcal{N}_{{}^{p}d}\left({}^{p}\hat{\mu}_{c,k-1},\, {}^{p}\hat{\Sigma}_{c,k-1}\right) \varrho_{p,k}({}^{p}d)\,\mathrm{d}\,{}^{p}d,\\
{}^{p}\nu_{c,p,k} &:= \int \mathcal{N}_{{}^{p}d}\left({}^{p}\hat{\mu}_{c,k-1},\, {}^{p}\hat{\Sigma}_{c,k-1}\right) \varrho_{p,k}({}^{p}d)\ {}^{p}d\ \mathrm{d}\,{}^{p}d,\\
{}^{\bar{p}}\nu_{c,p,k} &:= \int \mathcal{N}_{{}^{p}d}\left({}^{p}\hat{\mu}_{c,k-1},\, {}^{p}\hat{\Sigma}_{c,k-1}\right) \varrho_{p,k}({}^{p}d)\ {}^{\bar{p}|p}\hat{\mu}_{c,k-1}({}^{p}d)\,\mathrm{d}\,{}^{p}d.\\
{}^{p}\Omega_{c,p,k} &:= \int \mathcal{N}_{{}^{p}d}\left({}^{p}\hat{\mu}_{c,k-1},\, {}^{p}\hat{\Sigma}_{c,k-1}\right) \varrho_{p,k}({}^{p}d)\ {}^{p}d\ {}^{p}d'\,\mathrm{d}\,{}^{p}d\\
&\quad - {}^{p}\nu_{c,p,k}\,{}^{p}\hat{\mu}'_{c,k} - {}^{p}\hat{\mu}_{c,k}\,{}^{p}\nu'_{c,p,k} + \lambda_{c,k,p}\,{}^{p}\hat{\mu}_{c,k}\,{}^{p}\hat{\mu}'_{c,k}\\
{}^{p\bar{p}}\Omega_{c,p,k} &:= \int \mathcal{N}_{{}^{p}d}\left({}^{p}\hat{\mu}_{c,k-1},\, {}^{p}\hat{\Sigma}_{c,k-1}\right) \varrho_{p,k}({}^{p}d)\,({}^{p}d - {}^{p}\hat{\mu}_{c,k})\ {}^{\bar{p}|p}\hat{\mu}'_{c,k-1}({}^{p}d)\,\mathrm{d}\,{}^{p}d\\
&\quad - ({}^{p}\nu_{c,p,k} - \lambda_{c,k,p}\,{}^{p}\hat{\mu}_{c,k})\ {}^{\bar{p}}\hat{\mu}'_{c,k}\\
{}^{\bar{p}}\Omega_{c,p,k} &:= \int \mathcal{N}_{{}^{p}d}\left({}^{p}\hat{\mu}_{c,k-1},\, {}^{p}\hat{\Sigma}_{c,k-1}\right) \varrho_{p,k}({}^{p}d)\left({}^{\bar{p}|p}\hat{\mu}_{c,k-1}({}^{p}d)\,{}^{\bar{p}|p}\hat{\mu}'_{c,k-1}({}^{p}d) - {}^{\bar{p}|p}\hat{\mu}_{c,k-1}({}^{p}d)\,{}^{\bar{p}}\hat{\mu}'_{c,k}\right.\\
&\quad \left. - {}^{\bar{p}}\hat{\mu}_{c,k}\,{}^{\bar{p}|p}\hat{\mu}'_{c,k-1}({}^{p}d)\right)\mathrm{d}\,{}^{p}d + \lambda_{c,k,p}\left({}^{\bar{p}|p}\Sigma_{c,k-1} + {}^{\bar{p}}\hat{\mu}_{c,k}\,{}^{\bar{p}}\hat{\mu}'_{c,k}\right)
\end{aligned}$$

7. *For all* $c \in c^*, p \in p^*$, *set* $\nu_{c,p,k} \in \mathbb{R}^n$ *and* $\Omega_{c,p,k} \in \mathbb{R}^{n,n}$ *so that*

$${}^{p}S(\nu_{c,p,k}) = {}^{p}\nu_{c,p,k},\ \ {}^{\bar{p}}S(\nu_{c,p,k}) = {}^{\bar{p}}\nu_{c,p,k},$$

$${}^{pp}S(\Omega_{c,p,k}) = {}^{p}\Omega_{c,p,k},\ \ {}^{p\bar{p}}S(\Omega_{c,p,k}) = {}^{p\bar{p}}\Omega_{c,p,k},\ \ {}^{\bar{p}p}S(\Omega_{c,p,k}) = {}^{p\bar{p}}\Omega'_{c,p,k},\ \ {}^{\bar{p}\bar{p}}S(\Omega_{c,p,k}) = {}^{\bar{p}}\Omega_{c,p,k}.$$

8. *For all* $c \in c^*$, *set new parameter estimates*

$$\begin{aligned}
\hat{\omega}_{c,k} &:= \hat{\omega}_{c,k-1}\sum_{p=1}^{\mathring{p}} {}^{p}\alpha\lambda_{c,p,k},\\
\hat{\mu}_{c,k} &:= \frac{\hat{\omega}_{c,k-1}}{\hat{\omega}_{c,k}}\sum_{p=1}^{\mathring{p}} {}^{p}\alpha\nu_{c,p,k},\\
\hat{\Sigma}_{c,k} &:= \frac{\hat{\omega}_{c,k-1}}{\hat{\omega}_{c,k}}\sum_{p=1}^{\mathring{p}} {}^{p}\alpha\Omega_{c,p,k}.
\end{aligned}$$

9. $k := k + 1$; *If* $k \le K$, *then go to step* (4).

The result of Algorithm 4.3.2 are the parameters $\hat{\omega}_{c,K}, \hat{\mu}_{c,K}, \hat{\Sigma}_{c,K}$ of the best achieved approximation

$$\varphi_K(d) = \sum_{c=1}^{\mathring{c}} \hat{\omega}_{c,K}\mathcal{N}_d\left(\hat{\mu}_{c,K}, \hat{\Sigma}_{c,K}\right)$$

of ${}^{I}f\,(d)$ minimizing (4.1). The asset of this version of the algorithm is that the numerical integration (step 6) is performed over sets ${}^{p}d^*$ only. On that account, the applicability of Algorithm 4.3.2 is limited by the dimensions of quantities ${}^{p}d$ treated by the individual participants instead of the dimension of the complete system.

4.4 Application in Multiple Participant Decision making

Apart from constructing the global objectives, the proposed method can be applied also in the distributed multiple participant decision making. In this case, the method can be used by a pair of neighbours to adjust marginals of their ideal pdfs on common parts of their environments. Another possibility is to let a participant communicate with all its neighbours simultaneously and adjust the objectives on its complete environment. The former case leads to the problem addressed in Example 4.2.1, the analytical solution of which is known. In the latter case, the proposed iterative algorithm, or another suitable approximation, must be applied. In both cases the method is supposed to be applied repeatedly for various pairs or groups of participants. It is expected that the bigger groups of simultaneously communicating participants are, the smaller is a chance that the resulting adjusted ideal pdfs significantly differ from the corresponding marginals of the global ideal pdf. Note that in case of the distributed multiple participant decision making, the proposed method represents just a technical means for objective adjustment. A design of cooperation strategies, according to which the participants decides when, to whom, what to communicate, how to exploit the information acquired, etc., is still an open problem.

Chapter 5

Bayesian Knowledge Merging

In Section 3.5, we have sketched the role of the knowledge sharing in the multiple participant decision making. Although the design of a distributed solution has been left open, it is expected that it will rely on the knowledge sharing much like the global task. A distributed solution is to be reached by a cooperation among participants, which, due to limited computational resources, is supposed to be performed sequentially between pairs of neighbours. On that account, for the design of the cooperation methods a feasible procedure for knowledge sharing between pairs of neighbours is needed.

As the individual participants can employ completely different parametric models, the communication must be performed via quantities which are in common of the neighbours, i.e., common data. The relation (3.12) suggests that, on condition that the knowledge of individual participants can be equivalently expressed as sets of virtual observations, the information exchange could be provided by communicating these sets. However, even in this case it could be technically difficult to find these sets and to exploit them for an update of the prior pdfs. Moreover, this approach does not tackle the general case, in which the prior knowledge cannot be expressed as a set of virtual observations.

In this chapter, a method is proposed which allows a decision maker to exploit a knowledge represented by a joint pdf of data to update its prior pdf. In the multiple participant decision making the method could serve as a basis for a practically feasible, thought possibly approximate, procedure for knowledge exchange between participants. In such case, a joint pdf of data, or its approximation, can be acquired from participants' predictive pdfs and decision strategies.

In Section 5.1, the problem is stated and its solution for a special case is outlined. This motivates a heuristic solution of a general case, which is presented in Section 5.2. Sections 5.3 and 5.4 describe the proposed method tailored to parametric models from an exponential family and in a from of a probabilistic mixture, respectively. In Section 5.5, we briefly mention a case in which the information represented by a pdf of data is not specified for all quantities in the parametric model. Finally, the application of the proposed method in knowledge exchange between participants is discussed in Section 5.6.

5.1 Problem Formulation

Let us consider a time-invariant parametric model $f(d|\Theta)$, prior pdf $f(\Theta)$, and an information in a form of pdf $g(d)$. The aim is to design a method which exploits information represented by $g(d)$ to update the prior pdf $f(\Theta)$ so that $g(d)$ is taken with "limited relevance", i.e., $g(d)$ influences the prior pdf similarly as a finite number of observations. The following example motivates the proposed solution.

Let $\mathring{\tau} \in \mathbb{N}$ and a pdf $g(d)$ be in a form

$$g(d) = \frac{1}{\mathring{\tau}} \sum_{\tau=1}^{\mathring{\tau}} \delta(d - d_\tau),$$

for some $d_1, \ldots, d_{\mathring{\tau}} \in d^*$, i.e., $g(d)$ is a pdf corresponding to an empirical distribution of d from observations $d_1, \ldots, d_{\mathring{\tau}}$. In this case a suitable update of $f(\Theta)$ by $g(d)$, denoted by $f(\Theta|g(d), \mathring{\tau})$, suggests itself

as the posterior pdf of Θ given $d_1, \ldots, d_{\mathring{\tau}}$, i.e.,

$$f\left(\Theta|\, g(d), \mathring{\tau}\right) \equiv f\left(\Theta|\, d_1, \ldots, d_{\mathring{\tau}}\right). \tag{5.1}$$

The concept of the "limited relevance"of $g(d)$ has in (5.1) a clear meaning: it influences the prior pdf $f\left(\Theta\right)$ in the same way as $\mathring{\tau}$ observations. Note that the notation $f\left(\Theta|\, g(d), \mathring{\tau}\right)$ is merely a symbol and it has not a meaning of a conditional pdf.

In (5.1), the pdf $f\left(\Theta|\, g(d), \mathring{\tau}\right)$ is expressed using the sequence $d_1, \ldots, d_{\mathring{\tau}}$. Nevertheless, from $g(d)$ itself it is possible to acquire only values in the sequence $d_1, \ldots, d_{\mathring{\tau}}$ and their relative frequencies but not absolute numbers of occurrences. For example, sequences of observations $d_1, \ldots, d_{\mathring{\tau}}$ and $\tilde{d}_1, \ldots, \tilde{d}_{2\mathring{\tau}}$ such that, for all $\tau \in \{1, \ldots, \mathring{\tau}\}$, $\tilde{d}_\tau = \tilde{d}_{\mathring{\tau}+\tau} = d_\tau$ lead to the same empirical distributions. On that account, the parameter $\mathring{\tau}$, or its analogy expressing the "weight" of information $g(d)$, must be supplied externally together with the pdf $g(d)$.

5.2 Solution

In order to reach an analogy of (5.1) for an arbitrary pdf $g(d)$ and $\mathring{\tau} \geq 0$, we express the posterior pdf $f\left(\Theta|\, d_1, \ldots, d_{\mathring{\tau}}\right)$ using a relation involving the Kerridge inaccuracy (2.10) of the empirical pdf and the parametric model; see [38].

$$\begin{aligned} f\left(\Theta|\, d_1, \ldots, d_{\mathring{\tau}}\right) &\propto f\left(\Theta\right) \prod_{\tau=1}^{\mathring{\tau}} f\left(d_\tau|\, \Theta\right) = f\left(\Theta\right) \exp\left(\sum_{\tau=1}^{\mathring{\tau}} \ln f\left(d_\tau|\, \Theta\right)\right) = \\ &= f\left(\Theta\right) \exp\left(\int \sum_{\tau=1}^{\mathring{\tau}} \delta(d - d_\tau) \ln f\left(d|\, \Theta\right) \mathrm{d}d\right) = \\ &= f\left(\Theta\right) \exp\left(-\mathring{\tau} \mathsf{K}\left(r(d), f\left(d|\, \Theta\right)\right)\right) \end{aligned} \tag{5.2}$$

In (5.2), $\mathsf{K}\left(\cdot, \cdot\right)$ denotes the Kerridge inaccuracy (2.10) and $r(d)$ is an empirical pdf from data $d_1, \ldots, d_{\mathring{\tau}}$, i.e.,

$$r(d) = \frac{1}{\mathring{\tau}} \sum_{\tau=1}^{\mathring{\tau}} \delta(d - d_\tau).$$

Notice, that (5.2) depends on data $d_1, \ldots, d_{\mathring{\tau}}$ only through the empirical pdf $r(d)$ and the number of observations $\mathring{\tau}$. Furthermore, $\mathring{\tau}$ need not necessarily reflect the number of distinct values in $d_1, \ldots, d_{\mathring{\tau}}$. The elements in relation (5.2) can be interpreted so that the Kerridge inaccuracy $\mathsf{K}\left(r(d), f\left(d|\, \Theta\right)\right)$ expresses a quality of approximation of the pdf $r(d)$ by $f\left(d|\, \Theta\right)$ as a function of $\Theta \in \Theta^*$, and $\mathring{\tau}$ quantifies a relevance of the information represented by the pdf $r(d)$. It is important that the facts that $\mathring{\tau} \in \mathbb{N}$ and $r(d)$ is an empirical pdf play no role in (5.2) – it is well defined for any $\mathring{\tau} \in \mathbb{R}$ and any pdf $r(d)$ such that the Kerridge inaccuracy in (5.2) is finite for some Θ. Therefore, for an arbitrary pdf $g(d)$ and $\upsilon > 0$, the prior pdf $f\left(\Theta\right)$ updated by pdf $g(d)$ with the weight υ we establish as

$$f\left(\Theta|\, g(d), \upsilon\right) \propto f\left(\Theta\right) \exp\left(-\upsilon \mathsf{K}\left(g(d), f\left(d|\, \Theta\right)\right)\right). \tag{5.3}$$

Nevertheless, it should be stressed that in the current state of development the relation (5.3) is completely heuristic.

Generalization of (5.3) for a parametric model with a nonempty regression vector is straightforward. Consider a parametric model in a form of a conditional pdf $f\left(\Delta|\, \phi, \Theta\right)$. At this point, actions need not be explicitly distinguished, thus we can suppose that they are part of the state vector ϕ. The reasoning that leads to (5.2) indicates that for the parametric model $f\left(\Delta|\, \phi, \Theta\right)$, joint pdf $g(\Delta, \phi)$, and $\upsilon > 0$ the prior pdf $f\left(\Theta\right)$ updated by $g(\Delta, \phi)$ with weight υ has a form

$$f\left(\Theta|\, g(\Delta, \phi), \upsilon\right) \propto f\left(\Theta\right) \exp\left(-\upsilon \mathsf{K}\left(g(\Delta, \phi), f\left(\Delta|\, \phi, \Theta\right)\right)\right). \tag{5.4}$$

Note that from the analogy with the empirical pdf $r(\Delta, \phi)$ it follows that $g(\Delta, \phi)$ must be a joint pdf although the parametric model $f\left(\Delta|\, \phi, \Theta\right)$ is not. Roughly speaking, a conditional pdf $g(\Delta|\phi)$ gives

information about a frequency of observations of individual values of Δ after a particular regression vector ϕ has been observed. However, from $g(\Delta|\phi)$ it is impossible to deduce a weight of such information – it gives no information about "how often" a particular value of ϕ arises or if it arises at all.

5.3 Application to an Exponential Family

The proposed method can be easily tailored to parametric models from an exponential family (2.38). Consider a parametric model in a form

$$f\left(\Delta|\,\phi,\Theta\right) = A(\Theta)\exp\left(\langle B(\Delta,\phi),C(\Theta)\rangle\right),$$

see Section 2.3.3, and a joint pdf $g(\Delta,\phi)$. For such parametric model, the Kerridge inaccuracy in (5.4) has a form

$$\mathsf{K}\left(g(\Delta,\phi),f\left(\Delta|\,\phi,\Theta\right)\right) = -\ln A(\Theta) - \int g(\Delta,\phi)\langle B(\Delta,\phi),C(\Theta)\rangle \mathrm{d}\Delta\mathrm{d}\phi. \tag{5.5}$$

Substituting (5.5) into (5.4) we get

$$f\left(\Theta|\,g(\Delta,\phi),\upsilon\right) \propto f\left(\Theta\right)A^{\upsilon}(\Theta)\exp\left(\left\langle \upsilon\int g(\Delta,\phi)B(\Delta,\phi)\mathrm{d}\Delta\mathrm{d}\phi, C(\Theta)\right\rangle\right). \tag{5.6}$$

For a prior pdf in a conjugate form

$$f\left(\Theta\right) \propto A^{\nu_0}(\Theta)\exp\left(\langle V_0,C(\Theta)\rangle\right),$$

the pdf $f\left(\Theta|\,g(\Delta,\phi),\upsilon\right)$ can be expressed as

$$f\left(\Theta|\,g(\Delta,\phi),\upsilon\right) \propto A^{\nu}(\Theta)\exp\left(\langle V,C(\Theta)\rangle\right), \tag{5.7}$$

where

$$\nu = \nu_0 + \upsilon, \tag{5.8}$$

$$V = V_0 + \upsilon\int g(\Delta,\phi)B(\Delta,\phi)\mathrm{d}\Delta\mathrm{d}\phi. \tag{5.9}$$

In other words, for parametric models from the exponential family and conjugate prior pdfs, the evaluation of (5.4) reduces to updating the parameters V_0,ν_0 according to (5.8) and (5.9). Compared to (2.40), V_0 in (5.9) is updated by the ν-multiple of the expected value of $B(\Delta,\phi)$ with respect to $g(\Delta,\phi)$.

5.4 Quasi-Bayes Algorithm

For parametric models in a form of a probabilistic mixture (2.41), the updated prior pdf (5.4) can be approximately evaluated using a slightly modified quasi-Bayes algorithm; see Section 2.3.3.

Assume a prior pdf in a form analogous to (2.42)

$$f\left(\Theta\right) = Di_{\alpha}(\kappa)\prod_{c=1}^{\mathring{c}} f\left(\Theta_c\right),$$

where $\Theta \equiv (\alpha,\Theta_1,\ldots,\Theta_{\mathring{c}})$, $\alpha \equiv (\alpha_1,\ldots,\alpha_{\mathring{c}})$, $\alpha_c \geq 0$, for all $c \in c^*$, $\sum_{c=1}^{\mathring{c}}\alpha_c = 1$, and $\kappa \equiv (\kappa_1,\ldots,\kappa_{\mathring{c}})$ with $\kappa_c \geq 0$, for all $c \in c^*$. For a joint pdf $g(\Delta,\phi)$ and $\upsilon > 0$, the modified quasi-Bayes algorithm leads to the updated prior pdf $f\left(\Theta|\,g(\Delta,\phi),\upsilon\right)$ in a form

$$f\left(\Theta|\,g(\Delta,\phi),\upsilon\right) = Di_{\alpha}(\tilde{\kappa})\prod_{c=1}^{\mathring{c}}\tilde{f}(\Theta_c),$$

where, for all $c \in \{1, \ldots, \mathring{c}\}$,

$$\begin{aligned} \tilde{\kappa}_c &= \kappa_c + \upsilon \int w_c(\Delta, \phi) g(\Delta, \phi) \mathrm{d}\Delta \mathrm{d}\phi, \\ \tilde{f}(\Theta_c) &\propto f(\Theta_c) \exp\left(\upsilon \int w_c(\Delta, \phi) g(\Delta, \phi) \ln f(\Delta | \phi, \Theta_c) \, \mathrm{d}\Delta \mathrm{d}\phi \right), \\ w_c(\Delta, \phi) &= \frac{\kappa_c \int f(\Delta | \phi, \Theta_c) f(\Theta_c) \, \mathrm{d}\Theta_c}{\sum_{\tilde{c} \in c^*} \kappa_{\tilde{c}} \int f(\Delta | \phi, \Theta_{\tilde{c}}) f(\Theta_{\tilde{c}}) \, \mathrm{d}\Theta_{\tilde{c}}}. \end{aligned}$$

In the modified quasi-Bayes algorithm the information represented by $g(\Delta, \phi)$ and υ is processed all at once. Thus, contrary to the quasi-Bayes algorithm described in Section 2.3.3, the modified algorithm is not iterative.

5.5 Partial Information

In this paragraph we briefly discuss a case in which the pdf $g(\Delta, \phi)$ in (5.4) is not specified for all quantities in the parametric model. Assume that observation Δ is a vector random quantity. Let us split it into a pair of random quantities $\Delta \equiv (\Delta_1, \Delta_2)$. Then, the parametric model has a form

$$f(\Delta_1, \Delta_2 | \phi, \Theta). \tag{5.10}$$

First, we assume that only the marginal pdf $g(\Delta_1, \phi)$ is available. The analogy with an empirical pdf and $f(\Theta | g(\Delta_1, \phi), \upsilon)$ being a posterior pdf immediately leads to the conclusion that the updated prior pdf has a form

$$f(\Theta | g(\Delta_1, \phi), \upsilon) \propto f(\Theta) \exp\left(\upsilon \mathsf{K}\left(g(\Delta_1, \phi), f(\Delta_1 | \phi, \Theta) \right) \right), \tag{5.11}$$

where $f(\Delta_1 | \phi, \Theta)$ is a marginal pdf of the parametric model (5.10). However, even if the parametric model (5.10) is in an exponential family and the prior pdf $f(\Theta)$ is a conjugate one, the form of the resulting pdf (5.11) differs from (5.7), which can make its evaluation technically difficult.

Now, assume that the available information has a form $g(\Delta_1, \Delta_2)$. To evaluate $f(\Theta | g(\Delta_1, \Delta_2), \upsilon)$ we would need a pdf $f(\Delta_1, \Delta_2 | \Theta)$. Nevertheless, $f(\Delta_1, \Delta_2 | \phi, \Theta)$ cannot be, in general, derived from the parametric model (5.10) because a pdf $f(\phi | \Theta)$ is not available. This problem is exactly the same as that discussed in Section 3.5.2 in connection with the construction of a global prior pdf. The conclusions from Section 3.5.2 are thus valid here too.

5.6 Application in Multiple Participant Decision Making

This section is focused on a potential application of the proposed method on knowledge sharing between participants. The basic idea is simple: Participant's knowledge in a form of a prior (posterior) pdf to be shared is transformed to a joint pdf of data in common of the cooperating participants. This pdf is communicated to a neighbour, which exploits it using the relation (5.4). A suitable transformation of the prior pdf to the joint pdf of data can be acquired using the predictive pdf (2.30) and, eventually, the decision strategy. However, in a general case, which is shortly discussed at the end of this section, such a transformation is not so straightforward.

In what follows, we illustrate some features of the knowledge sharing based on the communication of the predictive pdfs and application of (5.4). As the method is partially based on heuristics, and suffers from a lack of the underlying theory, we use a particular example for this purpose. The example employs parametric models from an exponential family. For simplicity, we use a static parametric models in a natural parameterization, which is slightly different from the more general one introduced by (2.38). Before we approach to the example itself, let us recall some basic facts related to estimation of parametric models from an exponential family [18].

Let us consider a parametric model

$$f(d | \Theta) = \exp\left(\sum_{k=1}^{\mathring{k}} B_k(d) \Theta_k - A(\Theta) \right), \tag{5.12}$$

where $\mathring{k} \in \mathbb{N}, B_k : d^* \to \mathbb{R}$, for all $k \in k^*$, $\Theta \equiv (\Theta_1, \ldots, \Theta_{\mathring{k}}) \in \Theta^* \subset \mathbb{R}^{\mathring{k}}$, and $A : \Theta^* \to \mathbb{R}$. Θ_k in (5.12) are called natural parameters and $B(d) \equiv \big(B_1(d), \ldots, B_{\mathring{k}}(d)\big)$ is a sufficient statistic of the model. From the fact that $\int f\,(d|\,\Theta)\,\mathrm{d}d = 1$ for all $\Theta \in \Theta^*$ it follows that for $A(\Theta)$ it holds

$$A(\Theta) = \ln \int \exp\left(\sum_{k=1}^{\mathring{k}} B_k(d)\Theta_k\right) \mathrm{d}d.$$

Suppose that

$$\Theta^* = \left\{ \Theta \in \mathbb{R}^{\mathring{k}} \,\middle|\, A(\Theta) < +\infty \right\},$$

and Θ^* is an open set.

Now, consider a prior pdf in a conjugate form

$$f\,(\Theta) \propto \exp\left(n \sum_{k=1}^{\mathring{k}} V_k\Theta_k - nA(\Theta)\right), \tag{5.13}$$

where $n > 0$ and $V_k \in \mathbb{R}$, for all $k \in k^*$.

For the pdfs (5.12) and (5.13) it holds, see [18],

$$\mathsf{E}\left[B_k(d)|\,\Theta\right] = \int B_k(d) f\,(d|\,\Theta)\,\mathrm{d}d = \frac{\partial A(\Theta)}{\partial \Theta_k}, \tag{5.14}$$

$$\mathsf{E}\left[\frac{\partial A(\Theta)}{\partial \Theta_k}\right] = \int \frac{\partial A(\Theta)}{\partial \Theta_k} f\,(\Theta)\,\mathrm{d}\Theta = V_k, \tag{5.15}$$

for all $k \in k^*$. From (5.14) and (5.15) we get an important equality for the expectation of the sufficient statistic with respect to the predictive pdf. For all $k \in k^* \equiv \{1, \ldots, \mathring{k}\}$, it holds

$$\int B_k(d) f\,(d|\,\Theta)\, f\,(\Theta)\,\mathrm{d}\Theta\mathrm{d}d = V_k. \tag{5.16}$$

The following example extensively employs prior/posterior pdfs in a conjugate form. In the example we use for the conjugate pdfs a notation with explicitly indicated parameters of the pdfs. For example, the conjugate pdf (5.13) is denoted as $f\,(\Theta|\,n, V)$, where $V \equiv \big(V_1, \ldots, V_{\mathring{k}}\big)$.

Example 5.6.1 *Consider a pair of participants dealing with the same data $d \equiv {}^1d \equiv {}^2d$ and having parametric models from an exponential family with natural parameterization* (5.12)*, i.e., for $p \in \{1, 2\}$,*

$${}^pf\,(d|\,{}^p\Theta) = \exp\left(\sum_{{}^pk=1}^{{}^p\mathring{k}} {}^pB_{{}^pk}(d)\,{}^p\Theta_{{}^pk} - {}^pA\,({}^p\Theta)\right), \tag{5.17}$$

where ${}^p\mathring{k}$, ${}^pB_{{}^pk}$, ${}^p\Theta$, and ${}^pA$ are defined analogously to $\mathring{k}, B_k, \Theta$, and A in (5.12)*. Let ${}^p\Theta^*$ be open sets and satisfy*

$${}^p\Theta^* = \left\{ {}^p\Theta \in \mathbb{R}^{{}^p\mathring{k}} \,\middle|\, {}^pA({}^p\Theta) < +\infty \right\}.$$

Suppose that the participants have conjugate prior pdfs, i.e., for $p \in \{1, 2\}$, the prior pdfs are in a form

$${}^pf\,({}^p\Theta|\,{}^pn, {}^pV) \propto \exp\left({}^pn \sum_{{}^pk=1}^{{}^p\mathring{k}} {}^pV_{{}^pk}\,{}^p\Theta_{{}^pk} - {}^pn\,{}^pA\,({}^p\Theta)\right), \tag{5.18}$$

where ${}^pn > 0$ and ${}^pV \equiv \big({}^pV_1, \ldots, {}^pV_{{}^p\mathring{k}}\big) \in \mathbb{R}^{{}^p\mathring{k}}$. Furthermore, suppose that there is a sequence of observations $d^{1:\mathring{\tau}}$, for some $\mathring{\tau} \in \mathbb{N}$, which is available to the first participant. Thus, its actual knowledge is represented by the posterior pdf ${}^1f\left({}^1\Theta \middle| d^{1:\mathring{\tau}}\right)$, which is also conjugate one. Denoting, for all ${}^1k \in {}^1k^*$,

$${}^1U_{{}^1k} \equiv \frac{1}{\mathring{\tau}} \sum_{\tau=1}^{\mathring{\tau}} {}^1B_{{}^1k}(d_\tau),$$

the posterior pdf can be expressed as

$$ {}^1f\left({}^1\Theta\middle|\, d^{1:\mathring{\tau}}\right) = {}^1f\left({}^1\Theta\middle|\, {}^1\bar{n},\, {}^1\bar{V}\right), \tag{5.19} $$

where ${}^1\bar{V} \equiv \left({}^1\bar{V}_1, \ldots, {}^1\bar{V}_{{}^1\mathring{k}}\right)$ *and*

$$ \begin{aligned} {}^1\bar{n} &= {}^1n + \mathring{\tau}, \\ {}^1\bar{V}_{{}^1k} &= \frac{{}^1n}{{}^1n + \mathring{\tau}}\, {}^1V_{{}^1k} + \frac{\mathring{\tau}}{{}^1n + \mathring{\tau}}\, {}^1U_{{}^1k}, \end{aligned} $$

for all ${}^1k \in {}^1k^*$.

Now, observe what happens if the first participant provides a predictive pdf with a corresponding weight to the second participant, which exploits this information using the relation (5.4)*. In what follows, two particular cases of the predictive pdfs are considered:*

1. *The first participant provides the predictive pdf* ${}^1f\left(d\right)$ *based on the prior pdf* ${}^1f\left({}^1\Theta\middle|\, {}^1n,\, {}^1V\right)$*, i.e.,*

$$ {}^1f\left(d\right) = \int {}^1f\left(d\middle|\, {}^1\Theta\right)\, {}^1f\left({}^1\Theta\middle|\, {}^1n,\, {}^1V\right) \mathrm{d}\, {}^1\Theta, $$

   with the weight ${}^1n$*.*

2. *The first participant provides the predictive pdf* ${}^1f\left(d\middle|\, d^{1:\mathring{\tau}}\right)$ *based on the posterior pdf* ${}^1f\left({}^1\Theta\middle|\, {}^1\bar{n},\, {}^1\bar{V}\right)$*, i.e.,*

$$ {}^1f\left(d\middle|\, d^{1:\mathring{\tau}}\right) = \int {}^1f\left(d\middle|\, {}^1\Theta\right)\, {}^1f\left({}^1\Theta\middle|\, {}^1\bar{n},\, {}^1\bar{V}\right) \mathrm{d}\, {}^1\Theta, $$

   with the weight ${}^1\bar{n}$*.*

In the first case, the updated prior pdf of the second participant has a form

$$ {}^2f\left({}^2\Theta\middle|\, {}^1f\left(d\right), {}^1n\right) \propto {}^2f\left({}^2\Theta\right) \exp\left(-\, {}^1n \mathsf{K}\left({}^1f\left(d\right), {}^2f\left(d\middle|\, {}^2\Theta\right)\right)\right), $$

from which we get, using (5.17)*,* (5.18)*, and the definition* (2.10) *of the Kerridge inaccuracy,*

$$ \begin{aligned} &{}^2f\left({}^2\Theta\middle|\, {}^1f\left(d\right), {}^1n\right) \\ &\propto \exp\left(\left({}^2n + {}^1n\right) \sum_{{}^2k=1}^{{}^2\mathring{k}} \left(\frac{{}^2n}{{}^2n + {}^1n}\, {}^2V_{{}^2k} + \frac{{}^1n}{{}^2n + {}^1n} \int {}^1f\left(d\right)\, {}^2B_{{}^2k}(d)\mathrm{d}d\right) {}^2\Theta_{{}^2k} - \left({}^2n + {}^1n\right)\, {}^2A\left({}^2\Theta\right)\right). \end{aligned} $$

It means that the updated prior pdf ${}^2f\left({}^2\Theta\middle|\, {}^1f\left(d\right), {}^1n\right)$ *remains in the conjugate form and, using the notation introduced in* (5.18)*, it can be expressed as*

$$ {}^2f\left({}^2\Theta\middle|\, {}^1f\left(d\right), {}^1n\right) = {}^2f\left({}^2\Theta\middle|\, {}^2n^{(Pr)},\, {}^2V^{(Pr)}\right), \tag{5.20} $$

where ${}^2V^{(Pr)} \equiv \left({}^2V_1^{(Pr)}, \ldots, {}^2V_{{}^2\mathring{k}}^{(Pr)}\right)$ *and*

$$ {}^2n^{(Pr)} = {}^2n + {}^1n, \tag{5.21} $$

$$ {}^2V_{{}^2k}^{(Pr)} = \frac{{}^2n}{{}^2n + {}^1n}\, {}^2V_{{}^2k} + \frac{{}^1n}{{}^2n + {}^1n} \int {}^1f\left(d\right)\, {}^2B_{{}^2k}(d)\mathrm{d}d, \tag{5.22} $$

for all ${}^2k \in {}^2k^*$*. Analogously, for the predictive pdf* ${}^1f\left(d\middle|\, d^{1:\mathring{\tau}}\right)$*, the updated prior pdf of the second participant can be written as*

$$ {}^2f\left({}^2\Theta\middle|\, {}^1f\left(d\middle|\, d^{1:\mathring{\tau}}\right), {}^1\bar{n}\right) = {}^2f\left({}^2\Theta\middle|\, {}^2n^{(Po)},\, {}^2V^{(Po)}\right), \tag{5.23} $$

where $^2V^{(Po)} \equiv \left({}^2V_1^{(Po)}, \ldots, {}^2V_{{}^2\mathring{k}}^{(Po)} \right)$ *and*

$$\begin{aligned}
{}^2n^{(Po)} &= {}^2n + {}^1\bar{n} = {}^2n + {}^1n + \mathring{\tau}, \\
{}^2V_{{}^2k}^{(Po)} &= \frac{{}^2n}{{}^2n + {}^1\bar{n}} {}^2V_{{}^2k} + \frac{{}^1\bar{n}}{{}^2n + {}^1\bar{n}} \int {}^1f\left(d \middle| d^{1:\mathring{\tau}}\right) {}^2B_{{}^2k}(d)\mathrm{d}d \\
&= \frac{{}^2n}{{}^2n + {}^1n + \mathring{\tau}} {}^2V_{{}^2k} + \frac{{}^1n + \mathring{\tau}}{{}^2n + {}^1n + \mathring{\tau}} \int {}^1f\left(d \middle| d^{1:\mathring{\tau}}\right) {}^2B_{{}^2k}(d)\mathrm{d}d \qquad (5.24)
\end{aligned}$$

for all ${}^2k \in {}^2k^*$.

If, for some ${}^1k \in {}^1k^*$ *and* ${}^2k \in {}^2k^*$, *it holds* ${}^1B_{{}^1k} = {}^2B_{{}^2k}$, *then, according to* (5.16), *the relations* (5.22) *and* (5.24) *can be expressed as*

$${}^2V_{{}^2k}^{(Pr)} = \frac{{}^2n}{{}^2n + {}^1n} {}^2V_{{}^2k} + \frac{{}^1n}{{}^2n + {}^1n} {}^1V_{{}^1k} \qquad (5.25)$$

and

$$\begin{aligned}
{}^2V_{{}^2k}^{(Po)} &= \frac{{}^2n}{{}^2n + {}^1\bar{n}} {}^2V_{{}^2k} + \frac{{}^1\bar{n}}{{}^2n + {}^1\bar{n}} {}^1\bar{V}_{{}^1k} \\
&= \frac{{}^2n}{{}^2n + {}^1n + \mathring{\tau}} {}^2V_{{}^2k} + \frac{{}^1n}{{}^2n + {}^1n + \mathring{\tau}} {}^1V_{{}^1k} + \frac{\mathring{\tau}}{{}^2n + {}^1n + \mathring{\tau}} {}^1U_{{}^1k}, \qquad (5.26)
\end{aligned}$$

respectively.

The example illustrates some interesting features of the proposed method:

- The knowledge sharing based on the communication of the predictive pdf and application of the rlation (5.4) is able to provide accurate knowledge sharing in some cases. To clarify it, we evaluate a posterior pdf
$$ {}^2f\left({}^2\Theta \middle| {}^1f\left(d\right), {}^1n, d^{1:\mathring{\tau}} \right) \propto {}^2f\left({}^2\Theta \middle| {}^1f\left(d\right), {}^1n \right) {}^2f\left(d^{1:\mathring{\tau}} \middle| {}^2\Theta \right) \qquad (5.27)$$
of the second participant corresponding to the prior pdf (5.20) updated by the data $d^{1:\mathring{\tau}}$. Denoting, for all ${}^2k \in {}^2k^*$,
$$ {}^2U_{{}^2k} \equiv \frac{1}{\mathring{\tau}} \sum_{\tau=1}^{\mathring{\tau}} {}^2B_{{}^2k}(d_\tau), $$
the posterior pdf can be expressed as
$$ {}^2f\left({}^2\Theta \middle| {}^1f\left(d\right), {}^1n, d^{1:\mathring{\tau}} \right) = {}^2f\left({}^2\Theta \middle| {}^2n^{(Co)}, {}^2V^{(Co)} \right), \qquad (5.28)$$
where $^2V^{(Co)} \equiv \left({}^2V_1^{(Co)}, \ldots, {}^2V_{{}^2\mathring{k}}^{(Co)} \right)$ and
$$\begin{aligned}
{}^2n^{(Co)} &= {}^2n^{(Pr)} + \mathring{\tau} = {}^2n + {}^1n + \mathring{\tau}, \\
{}^2V_{{}^2k}^{(Co)} &= \frac{{}^2n^{(Pr)}}{{}^2n^{(Pr)} + \mathring{\tau}} {}^2V_{{}^2k}^{(Pr)} + \frac{\mathring{\tau}}{{}^2n^{(Pr)} + \mathring{\tau}} {}^2U_{{}^2k} = \\
&= \frac{{}^2n}{{}^2n + {}^1n + \mathring{\tau}} {}^2V_{{}^2k} + \frac{{}^1n}{{}^2n + {}^1n + \mathring{\tau}} \int {}^1f\left(d\right) {}^2B_{{}^2k}(d)\mathrm{d}d + \frac{\mathring{\tau}}{{}^2n + {}^1n + \mathring{\tau}} {}^2U_{{}^2k},
\end{aligned}$$
for all ${}^2k \in {}^2k^*$. Note that the superscript (Co) used here refers to a combined case. The parameters ${}^2n^{(Po)}$, $^2V^{(Po)}$ and ${}^2n^{(Co)}$, $^2V^{(Co)}$ of the updated prior pdf (5.20) and the posterior pdf (5.28) have the following properties:

1. ${}^{2}n^{(Po)} = {}^{2}n^{(Co)}$
2. If, for some ${}^{1}k \in {}^{1}k^{*}$ and ${}^{2}k \in {}^{2}k^{*}$, it holds that ${}^{1}B_{{}^{1}k} = {}^{2}B_{{}^{2}k}$, then from (5.16) and the equality of statistics ${}^{1}U_{{}^{1}k} = {}^{2}U_{{}^{2}k}$ it follows that ${}^{2}V_{{}^{2}k}^{(Po)} = {}^{2}V_{{}^{2}k}^{(Co)}$.

These properties can be interpreted so that the piece of knowledge from the data $d^{1:\mathring{\tau}}$ which is related to the parameter ${}^{2}\Theta_{{}^{2}k}$ is transferred accurately. Note that the equality ${}^{1}B_{{}^{1}k} = {}^{2}B_{{}^{2}k}$ is not a necessary condition for ${}^{2}V_{{}^{2}k}^{(Po)} = {}^{2}V_{{}^{2}k}^{(Co)}$. Due to the linearity of expectation, it is sufficient to require that

$$ {}^{2}B_{{}^{2}k} = \sum_{{}^{1}k=1}^{{}^{1}\mathring{k}} \beta_{{}^{1}k}\, {}^{1}B_{{}^{1}k} \tag{5.29} $$

for some $\beta_1, \ldots, \beta_{{}^{1}\mathring{k}} \in \mathbb{R}$. Notice also, that if ${}^{1}B_{{}^{1}k} = {}^{2}B_{{}^{2}k}$ holds for some ${}^{1}k \in {}^{1}k^{*}$ and ${}^{2}k \in {}^{2}k^{*}$, then the resulting value of the parameter ${}^{2}V_{{}^{2}k}^{(Po)}$ is independent of the "rest" of the parametric model ${}^{1}f\left(d\middle|\, {}^{1}\Theta\right)$, i.e., of the statistics ${}^{1}B_k(d)$ for $k \in {}^{1}k^{*}$ such that $k \neq {}^{1}k$. This result can be straightly generalized to the case in which ${}^{2}B_{{}^{2}k}$ satisfies the condition (5.29). ${}^{2}V_{{}^{2}k}^{(Po)}$ is then independent of the statistics ${}^{1}B_{{}^{1}k}(d)$ for ${}^{1}k \in {}^{1}k^{*}$ such that $\beta_{{}^{1}k} = 0$.

- In a general case, i.e., for ${}^{2}k \in {}^{2}k^{*}$ such that the condition (5.29) is not satisfied, the value of ${}^{2}V_{{}^{2}k}^{(Po)}$ is different from ${}^{2}V_{{}^{2}k}^{(Co)}$. Whereas the parameter ${}^{2}V_{{}^{2}k}^{(Co)}$ depends on data $d^{1:\mathring{\tau}}$ directly through the statistic ${}^{2}U_{{}^{2}k}$, the parameter ${}^{2}V_{{}^{2}k}^{(Po)}$ is influenced by the data through the expectation of ${}^{2}B_{{}^{2}k}(d)$ with respect to the predictive pdf $f\left(d\middle|\, d^{1:\mathring{\tau}}\right)$. Note, that, contrary to the above discussed case, the relation (5.24) makes the parameter ${}^{2}V_{{}^{2}k}^{(Po)}$ dependent on the complete parametric model ${}^{1}f\left(d\middle|\, {}^{1}\Theta\right)$, i.e., on all statistics ${}^{1}B_{{}^{1}k}(d)$ for ${}^{1}k \in {}^{1}k^{*}$, as well as on the entire parameter $\bar{V}$ of the posterior pdf (5.19). In this sense the proposed method provides only approximate knowledge sharing.

- The following feature of the proposed method is not strictly related to the parametric models from an exponential family but holds generally. The relation (5.4) for updating the prior pdf $f(\Theta)$ by the information in the form $g(\Delta, \phi), \nu$ is "reversible" in the sense that knowing $g(\Delta, \phi)$ and ν the prior pdf $f(\Theta)$ could be reached from $f(\Theta|\, g(\Delta, \phi), \nu)$ by the relation

$$ f(\Theta) \propto f(\Theta|\, g(\Delta, \phi), \upsilon) \exp\left(\upsilon \mathsf{K}\left(g(\Delta, \phi), f(\Delta|\, \phi, \Theta)\right)\right). \tag{5.30} $$

This property is important for practical applications as it is expected that the particpants will share the knowledge repeatedly in time. Assume, for instance, that the first participant in Example 5.6.1 initially provides the predictive pdf ${}^{1}f(d)$ to the second participant. Later, after it acquires data $d^{1:\mathring{\tau}}$, it provides also the predictive pdf ${}^{1}f\left(d\middle|\, d^{1:\mathring{\tau}}\right)$. If the second participant exploits both these information pieces via (5.4), its updated prior pdf

$$ {}^{2}f\left({}^{2}\Theta\middle|\, {}^{1}f(d), {}^{1}n, {}^{1}f\left(d\middle|\, d^{1:\mathring{\tau}}\right), {}^{1}\bar{n}\right) \propto {}^{2}f\left({}^{2}\Theta\middle|\, {}^{1}f(d), {}^{1}n\right) \exp\left(-{}^{1}\bar{n}\mathsf{K}\left({}^{1}f\left(d\middle|\, d^{1:\mathring{\tau}}\right), {}^{2}f\left(d\middle|\, {}^{2}\Theta\right)\right)\right) $$

incorporates the information from ${}^{1}f(\Theta)$ twice – through ${}^{1}f(d)$ and through ${}^{1}f\left(d\middle|\, d^{1:\mathring{\tau}}\right)$. Knowing ${}^{1}f(d)$ and ${}^{1}n$, the relation (5.30) allows to update ${}^{2}f\left({}^{2}\Theta\middle|\, {}^{1}f(d), {}^{1}n\right)$ to ${}^{2}f\left({}^{2}\Theta\middle|\, {}^{1}f\left(d\middle|\, d^{1:\mathring{\tau}}\right), {}^{1}\bar{n}\right)$ correctly by

$$ \begin{aligned} &{}^{2}f\left({}^{2}\Theta\middle|\, {}^{1}f\left(d\middle|\, d^{1:\mathring{\tau}}\right), {}^{1}\bar{n}\right) \\ &\quad \propto {}^{2}f\left({}^{2}\Theta\middle|\, {}^{1}f(d), {}^{1}n\right) \exp\left({}^{1}n\mathsf{K}\left({}^{1}f(d), {}^{2}f\left(d\middle|\, {}^{2}\Theta\right)\right) - {}^{1}\bar{n}\mathsf{K}\left({}^{1}f\left(d\middle|\, d^{1:\mathring{\tau}}\right), {}^{2}f\left(d\middle|\, {}^{2}\Theta\right)\right)\right). \end{aligned} $$

In this way, the presented method could be used not only to share a complete participant's knowledge but also to communicate and exploit its increments or changes caused, e.g., by a redesign of a decision strategy.

In Example 5.6.1 we have supposed that the parametric models are static ones, i.e., their state vectors are empty. In such case, the prior pdf can be easily transformed to the joint pdf of the data in common of the participants as a marginal from the predictive pdf. However, in a general case a suitable

transformation is not so straightforward. Assume, for an illustration, that the participant sharing its knowledge has a parametric model ${}^1f\left({}^1\Delta_t\middle|\,{}^1a_t,\,{}^1\phi_{t-1},\,{}^1\Theta\right)$, decision strategy ${}^1f\left({}^1a_t\middle|\,{}^1d^{1:t-1}\right)$, and prior pdf ${}^1f\left({}^1\Theta\right)$. For simplicity, we can suppose that the quantities common with the participant's neighbour are $\left({}^1\Delta_t,\,{}^1a_t,\,{}^1\phi_{t-1}\right)$. Then, the participant needs a joint pdf ${}^1f\left({}^1\Delta_t,\,{}^1a_t,\,{}^1\phi_{t-1}\right)$. Of course, this joint pdf cannot be derived form the parametric model, decision strategy, and prior pdf. Suppose, for a while, that the decision strategy ${}^1f\left({}^1a_t\middle|\,{}^1d^{1:t-1}\right)$ is time invariant. In this case, the actions ${}^1a_t$ necessarily depends only on some finite dimensional subvector of ${}^1d^{1:t-1}$ of a fixed structure. For simplicity, we can assume that this subvector is ${}^1d^{t-\,{}^1T:t-1}$, where ${}^1T$ is the length of the state vector ${}^1\phi_{t-1}$; see Section 3.2. The participant is then able to construct a conditional pdf

$$ {}^1f\left({}^1\Delta_t,\,{}^1a_t\middle|\,{}^1d^{t-\,{}^1T:t-1}\right)={}^1f\left({}^1a_t\middle|\,{}^1d^{t-\,{}^1T:t-1}\right)\int{}^1f\left({}^1\Delta_t\middle|\,{}^1a_t,\,{}^1\phi_{t-1},\,{}^1\Theta\right)\,{}^1f\left({}^1\Theta\right)\mathrm{d}\,{}^1\Theta, $$

from which a stationary pdf ${}^1\tilde{f}\left({}^1d^{t-\,{}^1T:t-1}\right)$ could be derived, if it exists. Finally, a suitable approximation of ${}^1f\left({}^1\Delta_t,\,{}^1a_t,\,{}^1\phi_{t-1}\right)$ can be acquired as a marginal pdf from

$$ {}^1f\left({}^1\Delta_t,\,{}^1a_t\middle|\,{}^1d^{t-\,{}^1T:t-1}\right)\,{}^1\tilde{f}\left({}^1d^{t-\,{}^1T:t-1}\right). $$

Nevertheless, in practice the decision strategy ${}^1f\left({}^1a_t\middle|\,{}^1d^{1:t-1}\right)$ is evolving in time and thus cannot be taken as time invariant. In such cases a more detailed modelling must be employed to reach a suitable approximation of ${}^1f\left({}^1\Delta_t,\,{}^1a_t,\,{}^1\phi_{t-1}\right)$.

Note that a byproduct of the proposed method is that the relation (5.4) provides a means for exploiting an expert information in a form of an arbitrary joint pdf $g(\Delta,a,\phi)$ in case that this pdf can be taken as a limit, or an approximation, of an empirical pdf and the information has a relevance analogous to a finite number of observations. This approach has been used, e.g., in [28], [36], and [37].

Chapter 6

Summary and Conclusions

The objective of the thesis is to design practically feasible methods that allow a group of Bayesian decision makers acting in the same system to communicate information pieces on their knowledge and objectives and use them to enhance the quality of the decision making. The main part of the work could be split into two parts:

- Chapter 3, in which the decision making with multiple participants is introduced and studied.
- Chapters 4 and 5, in which the communication methods themselves are proposed.

In **Chapter 3**, a discussion of a possible extension of the single participant Bayesian decision making towards the multiple participant one is presented. Due to the complexity of the addressed problem, a case study has been employed for this purpose. The key assumptions, which determine our approach, are that no consistencies of parametric models, prior pdfs, and objectives of individual participants are a priori required and that all computations are to be performed by the participants themselves, whereas their computational resources are limited, see Section 3.3. Although our primary aim is the multiple participant decision making in a distributed form, it comes out that some kind of a centralized decision making, called here a global task, must be inevitably considered, Section 3.4. The global task serves as a criterion which enables to compare $\mathring{p}$-tuples of decision strategies of the individual participants. We have attempted to construct the global task as a Bayesian one. However, there is no unique way to treat the multiplicity of uncertainty assessments and objectives of individual participants. On that account, we have adopted additional conditions that specify a particular form of the global task, Section 3.4. In Section 3.5, we have outlined partial problems related to the construction of the global parametric model, prior pdf, and loss function. At this point we have employed assumptions that allow to model the sources of inconsistencies of parametric models and prior pdfs of individual participants.

The results acquired in this chapter lead to the following conclusions:

- Any normative distributed multiple participant decision making induces certain kind of the global task, which provides a criterion according to that the $\mathring{p}$-tuples of decision strategies of the individual participants are compared, Section 3.4. However trivial this observation is, it plays a crucial role in the design of a distributed decision making. It implies that any attempt to design the distributed solution directly from the local tasks, without considering a related global task, is just a simplification for which it is payed by a loss of the normativeness.
- The participants themselves do not provide enough information for the global task to be constructed as a Bayesian one in the sense of Section 2.3; see Section 3.5.2. Any procedure that maps the information pieces provided by the individual participants to the global parametric model and prior pdf must, more or less explicitly, model the relation between participants' information and the "true" model of the system. This information model should reflect possible reasons of inconsistency, measure of overlapping, and eventually other aspects related to the multiplicity of information sources. However, the information acquired from the participants do not provide any evidence on which the choice of a suitable information model could be based. In Section 3.5.2 the lack of evidence has been compensated by the assumptions that the inconsistency of local parametric

models is solely caused by technical limitations of the participants and that the local prior pdfs represent information pieces from independent sources.

This issue seems to be the most serious open problem of the proposed treatment of the multiple participant decision making. On one hand, it is clear that the information model cannot be left completely unspecified because in such case there would not be any known relation between participants' information pieces and the "true" model of the system. On the other hand, the assumption that the information model is fully specified is rather unrealistic, especially if the distributed form of the decision making is aimed at. A compromise between the two extreme cases could potentially provide a way out: Whenever some kind of parametric information model could be assumed to be generally acceptable for some class of applications, the complete uncertainty about the information model reduces to the uncertainty about its unknown parameter. The parametric information models need not be necessarily represented merely by conditional pdfs, but could be possibly of a more general form. The relation between the local and global parametric models and the local and global prior pdfs based on the assumption stated in Section 3.5.2 and extended by some model of dependency of participants' information could serve as an example of such parametric information model. In this case, the parameter of the information model coincides with the parameter of the dependency model. Another type of the information model could be based, e.g., on the assumption that the participants information correspond to the "true" model of the system affected by some error of known type but with unknown parameters. Modeling of the dependency among information sources is widely addressed in the literature, see, e.g., [26] or [51].

The uncertainty about the unknown, ideally finite dimensional, parameter of the information model can be treated via standard procedures of decision making under uncertainty. Nevertheless, it should be stressed that the parameter relates to aspects of the information model for which no prior information can be acquired from the individual participants. In other words, an availability of any "informative" prior pdf of the parameter cannot be supposed.

At this point several scenarios of further development can be considered. We briefly mention two of them:

1. The prior pdf is left unspecified and the partial order of global decision strategies induced by their dominance is considered as the criterion according to which the strategies are compared.
2. The prior pdf is selected as a non-informative one, and the global strategies are compared according to the expected global loss. However, it is not clear what kind of non-informative prior pdfs should be selected for the parameters of information models. Moreover, this approach should be ideally supported by learning of the parameter, which is possible only if the participants communicate sequentially. In this case it would be also necessary to extend the participants' parametric models and prior pdfs so that they are able to reflect the development of knowledge about the information model.

Undoubtedly, this issue should be a matter of further discussion.

- The assumption on limited computational resources of the participants is insufficiently formalized, which is in a sharp contradiction with its significance. For example, the choice of participant's parametric model is mostly a compromise between its prior knowledge about the system and its technical limitations; see Section 3.5.2. From the parametric model itself it is impossible to abstract these two aspects separately, which heavily complicates the construction of the global task. In Section 3.5.2 this difficulty has been overcome by the assumption that the inconsistency of participants' parametric models is solely caused by the technical limitations.

- The special case in which the local prior pdfs are in a conjugate form enables to represent the knowledge of individual participants as sets of observations. If, in addition, all local parametric models have the same regression vector, the global prior pdf can be evaluated as a posterior one; see Section 3.5.2. This approach allows to combine the participants' knowledge in a natural way, even if the participants employ different parametric models. We believe that this special case may serve as a suitable starting point for a treatment of more general ones.

In **Chapter 4**, a method is proposed which enables to establish the global objective in case that the participants employ the fully probabilistic design. The global objective, represented by the global ideal pdf, is selected so that it is in some sense close to the ideal pdfs of the individual participants. The measure of proximity is selected as a weighted sum (4.1) of the Kullback-Leibler divergences of the participants' ideal pdfs to the corresponding marginal pdfs of the global ideal pdf, Section 4.1. The global ideal pdf is then searched as a minimizer of this function. The major part of Chapter 4 is focused on an analysis of the proposed optimization task and finding its approximate solution.

The main results of this chapter are as follows:

- A necessary condition for a global ideal pdf to be the optimal one is stated in Proposition 4.2.2. The condition is given by the equation (4.11), which is analytically intractable, except for a few special cases.

- Proposition 4.2.3 states that the equation (4.11) is, under an additional assumption, also a sufficient condition for a global ideal pdf to be the optimal one.

- An approximate solution of the optimization task defining the global ideal pdf can be reached using an iterative algorithm based on repetitive applications of the operator (4.9). Its convergence is, under a relatively mild assumption, guaranteed by Proposition 4.3.1.

- The iterative algorithm cannot be directly implemented for continuous quantities because of the intractable form of the pdfs approximating the global ideal pdf. To overcome this difficulty, a generalized EM algorithm is introduced in Section 4.3.2. The generalized EM algorithm can be used for approximation of a given pdf by a probabilistic mixture. Its monotone convergence is guaranteed by Proposition 4.3.3. Combining the two algorithms we get Algorithm 4.3.1 and its optimized version – Algorithm 4.3.2, which allow to find an approximate global ideal pdf in the form of a finite Gaussian mixture.

In **Chapter 5**, we present a method which allows a Bayesian decision maker to exploit a knowledge represented by a joint pdf of data. In multiple participant decision making this method can be used as an approximate, but practically feasible, means for knowledge sharing between participants. The core of the proposed method is the relation (5.4). Its practical feasibility stems from the fact that for a parametric model from an exponential family and a conjugate prior pdf the updated prior pdf remains in a conjugate form; see Section 5.3. The method can be also used with parametric models in a form of a finite mixture. In this case the updated prior pdf can be evaluated using a modified quasi-Bayes algorithm; see Section 5.4. Knowledge sharing between participants based on communication of a predictive pdf and application of the relation (5.4) is discussed in Section 5.6. It is also illustrated that, under some assumptions on the parametric models of the communicating participants, the proposed method is able to provide accurate knowledge sharing. However, in a general case the transferred knowledge can partially depend on the participants' parametric models. The applicability of the proposed method is limited by the requirement to specify the shared knowledge as a joint pdf of data in common of the neighbours, which can be often done only approximately. This property is just a consequence of the fact that the participants cannot provide enough information for the global task to be constructed as a Bayesian one; see Section 3.5.2.

6.1 Contributions

The main contributions of the thesis can be summarized as follows:

- The case study in Section 3.5 allows a better understanding of the bottlenecks faced in the Bayesian decision making with multiple participants.

- A feasible and theoretically supported algorithmic solution for the minimization of the weighted sum of the Kullback-Leibler divergences (4.2) has been reached for both discrete and continuous random quantities. The proposed algorithms allow to find common objectives of the participants employing the fully probabilistic design.

- A method has been proposed which allows a Bayesian decision maker to exploit an information represented by a joint pdf of data to update its prior pdf. In the multiple participant decision making this method provides a practically feasible means for knowledge sharing between participants.

Apart from the above listed main contributions also several minor ones have been acquired:

- The generalized EM algorithm proposed in Section 4.3.2 provides a means for approximation of complex pdfs by finite probabilistic mixtures.
- The method for knowledge sharing between participants proposed in Section 5.2 can be used for exploiting expert information in a form of a joint pdf of data.
- The comments on the fully probabilistic design in Appendix B point out its limitations and open a critical discussion on this kind of design.

6.2 Open Problems

- In the current state of development, the formulation of the multiple participant decision making is still inadequate. Especially, both the basic assumptions and general aims are insufficiently specified. For a future development the following issues should be addressed:
  - The assumptions on limited computational resources have to be specified in more details and in an appropriate form.
  - An extent of information that could be possibly acquired from an external source, i.e., not from the participants themselves, has to be specified.
  - The aims that are to be followed by the multiple participant decision making must be arranged.
- Even a simple instance of the multiple participant decision making is a technically demanding task and requires suitable approximating procedures.
  - The construction of the global prior pdf in Section 3.5 as well as the updating of prior pdfs of individual participants in Section 5.2 is typically done using an information which is specified only partially, i.e., not for all quantities in the corresponding parametric models. In such cases, the prior pdfs can easily became intractable even for parametric models from an exponential family. Similarly, the predictive pdfs representing participants' knowledge, see Section 5.6, can be hardly tractable. On that account, suitable approximation methods for handling such pdfs must be found.
  - In the multiple participant decision making a specific kind of problems stems from the inconsistency of the individual parametric models. The construction of the global parametric model in Section 3.5.2 is one of them. Although a general treatment of the problem has been sketched, it is technically entirely deficient. We hope that the apparatus of the information geometry [2] could provide means to acquire a better insight into this kind of problems.
- A design of a distributed form of the multiple participant decision making is still an open problem. The methods proposed is Chapters 4 and 5 provide means for communication of the participants, but a design of particular communication strategies, according to which the participants decide when, with whom, and what to communicate, remains unsolved.
- The relation (5.4), which allows a Bayesian decision maker to update its prior pdf by an information in a form of a joint pdf of data, is partially based on heuristics. This limits its further development as a procedure for exploiting expert information.

Appendix A

Binary Relations and Orders

Decision maker's preferences are often characterized in terms of binary relations. In this appendix, definitions and properties of several types of binary relations are summarized. They are taken from [20].

A binary relation $\mathcal{R}$ on a set M is

- *reflexive* if $x\mathcal{R}x$, for all $x \in M$,
- *irreflexive* if not $x\mathcal{R}x$, for all $x \in M$,
- *symmetric* if $x\mathcal{R}y \Rightarrow y\mathcal{R}x$, for all $x, y \in M$,
- *asymmetric* if $x\mathcal{R}y \Rightarrow$ not $y\mathcal{R}x$, for all $x, y \in M$,
- *antisymmetric* if $(x\mathcal{R}y, y\mathcal{R}x) \Rightarrow x = y$, for all $x, y \in M$,
- *transitive* if $(x\mathcal{R}y, y\mathcal{R}z) \Rightarrow x\mathcal{R}z$, for all $x, y, z \in M$,
- *negatively transitive* if $(\text{not } x\mathcal{R}y, \text{not } y\mathcal{R}z) \Rightarrow \text{not } x\mathcal{R}z$, for all $x, y, z \in M$,
- *connected* if $x\mathcal{R}y$ or $y\mathcal{R}x$, for all $x, y \in M$,
- *weakly connected* if $x \neq y \Rightarrow (x\mathcal{R}y \text{ or } y\mathcal{R}x)$, for all $x, y \in M$,
- a *partial order* if $\mathcal{R}$ on M is reflexive, antisymmetric, and transitive,
- a *strict partial order* if $\mathcal{R}$ on M is irreflexive and transitive,
- a *weak order* if $\mathcal{R}$ on M is asymmetric and negatively transitive,
- a *strict order* if $\mathcal{R}$ on M is a weakly connected weak order,
- an *equivalence* if $\mathcal{R}$ on M is reflexive, symmetric, and transitive.

For a binary relation $\prec$ on a set M representing preferences about elements of M, i.e., $x \prec y$ means that x is less preferred then y, we define a relation of indifference $\sim$:

$$x \sim y \Leftrightarrow (\text{not } x \prec y,\ \text{not } y \prec x)\,. \tag{A.1}$$

The preference-indifference relation $\preceq$ is defined as a union of $\prec$ and $\sim$:

$$x \preceq y \Leftrightarrow x \prec y \text{ or } x \sim y. \tag{A.2}$$

The following theorem summarizes basic properties of a weak order.

Theorem A.0.1 *Suppose that $\prec$ is a weak order on a set M, and the relations $\sim$ and $\preceq$ are defined by* (A.1) *and* (A.2)*, respectively. Then*

- *exactly one of* $x \prec y, y \prec x, x \sim y$ *holds, for all* $x, y \in M$,
- $\prec$ *is transitive,*
- $\sim$ *is an equivalence,*
- $(x \prec y, y \sim z) \Rightarrow x \prec z$, *and* $(x \sim y, y \prec z) \Rightarrow x \prec z$, *for all* $x, y, z \in M$,
- $\preceq$ *is transitive and connected,*
- *with* $\prec'$ *on the set* $M/\sim$, *being the set of equivalence classes of* M *under* $\sim$, *defined by*

$$a \prec' b \Leftrightarrow x \prec y \text{ for some } x \in a, y \in b,$$

  $\prec'$ *on* $M/\sim$ *is a strict order.*

Appendix B

Critical Comments on FPD

The definition of the optimal decision strategy in the sense of (2.48) has certain drawbacks of both theoretical and practical nature. They are briefly discussed here. In the discussion, we use a simple decision making task involving

- action a and observation Δ,
- loss function $L(a, \Delta)$,
- outer model of the system $f(\Delta|a)$.

The fundamental defect of the FPD is that it cannot be taken as an extension of a decision making on non-randomized actions to a decision making on randomized actions. Within the Bayesian framework, see Section 2.3, the preference ordering on randomized actions $f(a)$ is induced by the expected loss

$$\mathcal{L}_r(f(a)) \equiv \mathsf{E}\left[L(a,\Delta)\right] = \int L(a,\Delta) f(\Delta|a)\, f(a)\, \mathrm{d}\Delta \mathrm{d}a. \tag{B.1}$$

Its restriction to non-randomized actions, represented by pdfs $f(a) = \delta(a - \tilde{a})$, for $\tilde{a} \in a^*$, where $\delta(\cdot)$ denotes the Dirac delta function, corresponds to the expected utility

$$\mathcal{L}_n(a) \equiv \mathsf{E}\left[L(a,\Delta)|\, a\right] = \int L(a,\Delta) f(\Delta|a)\, \mathrm{d}\Delta$$

defined for non-randomized actions $a \in a^*$. In the case of the FPD, preferences among randomized actions are given by a function $\mathcal{L}_f : \mathcal{F}(a) \to \mathbb{R}$,

$$\mathcal{L}_f(f(a)) \equiv \mathsf{D}\left(f(\Delta|a)\, f(a) \,\middle|\middle|\, {}^I f(a,\Delta)\right), \tag{B.2}$$

for some ideal pdf ${}^I f(a,\Delta) \in \mathcal{F}(a,\Delta)$. For ${}^I f(a,\Delta)$ being a pdf of an absolutely continuous distribution, it holds, for all $\tilde{a} \in a^*$, $\mathcal{L}_f(\delta(a-\tilde{a})) = +\infty$, which implies that the function $\mathcal{L}_f$ does not induce any non-trivial ordering on non-randomized actions.

Often it is argued that the FPD represents a regular Bayesian decision making with the loss function $\ln \frac{f(a,\Delta)}{{}^I f(a,\Delta)}$. However, such "loss function" is not a function on the set $a^* \times \Delta^*$ but on the set $\mathcal{F}(a) \times a^* \times \Delta^*$, and thus it does not determine preference ordering on the set $a^* \times \Delta^*$.

Representing objectives by ideal pdfs leads also to some practical difficulties. Namely, analytical solution (2.49) of the FPD requires the ideal pdf to be specified as a joint pdf of a and Δ. In this way, a decision maker is indirectly pushed to specify a conditional pdf ${}^I f(\Delta|a)$. In general, this point seems to be irrational as the decision maker has no possibility to influence the resulting conditional pdf of Δ given a, which is fixed and given by the outer model of the system. Nevertheless, the choice of ${}^I f(\Delta|a)$ has an impact on the resulting optimal $f(a)$. An interesting alternative would be to specify only marginal ideal pdfs ${}^I f(a)$ and ${}^I f(\Delta)$. However, this approach does not fit into the FPD framework, as it would require simultaneous minimization of the Kullback-Leibler divergences

$$\mathsf{D}\left(f(a) \,\middle|\middle|\, {}^I f(a)\right) \text{ and } \mathsf{D}\left(\int f(\Delta|a)\, f(a)\, \mathrm{d}a \,\middle|\middle|\, {}^I f(\Delta)\right), \tag{B.3}$$

which are, in general, contradicting requirements. To overcome this difficulty, it is recommended to specify the joint ideal pdf in a form

$$ {}^{I}f(a,\Delta) = {}^{I}f(a)\,{}^{I}f(\Delta) \tag{B.4} $$

and use the regular FPD (2.48). Apparently, this approach is not suitable as (B.4) claims just that $\forall a \in a^*$, ${}^{I}f(\Delta|a) = {}^{I}f(\Delta)$. For example, in case that ${}^{I}f(\Delta) = \int f(\Delta|a)\,{}^{I}f(a)\,\mathrm{d}a$, the optimal solution based on minimization of the Kullback-Leibler divergences (B.3) is $f(a) = {}^{I}f(a)$, but the FPD with the ideal pdf in form (B.4) leads to

$$ f(a) \propto {}^{I}f(a) \exp\left(-\int f(\Delta|a) \ln \frac{f(\Delta|a)}{{}^{I}f(\Delta)} \mathrm{d}\Delta\right). \tag{B.5} $$

Obviously, $f(a)$ in (B.5) is, in general, not equal to ${}^{I}f(a)$.

In practice, it would be also interesting to specify objectives in a form ${}^{I}f(\Delta)$ only, and find an optimal randomized action so that

$$ {}^{O}f(a) \in \underset{f(a)\in\mathcal{F}(a)}{\arg\min}\, \mathsf{D}\left(\int f(\Delta|a)\, f(a)\,\mathrm{d}a \,\middle\|\, {}^{I}f(\Delta)\right). \tag{B.6} $$

However, to find an analytical solution of (B.6) is a hard, if not insolvable, task.

The FPD also does not allow to formulate some important types of preferences which could be easily described by loss functions (2.21) on data. This negative feature of the FPD is illustrated by the following examples.

In the first example, it is shown that by means of the FPD it is impossible to describe equality of preferences equivalent to a loss function constant on some set.

Example B.0.1 *In this example, we consider an oversimplified decision making task with only one quantity – action a. Let $[0,1] \subset a^*$, and a loss function $L : a^* \to \mathbb{R}$ be defined by*

$$ L(a) = \begin{cases} 0 & \text{for } a \in [0,1] \\ 1 & \text{otherwise} \end{cases}. $$

Obviously, this loss function, or its expectation, is minimized by any action $a \in [0,1]$, or any randomized action $f(a)$ such that $\int_0^1 f(a)\,\mathrm{d}a = 1$. It is impossible to reach such results within the FPD framework. To demonstrate it, suppose that ${}^{I}f(a) \in \mathcal{F}(a)$ is an ideal pdf such that $\mathsf{D}\left(f(a)\,\middle\|\,{}^{I}f(a)\right)$ is minimized iff $f(a) \in \mathcal{F}_0(a)$, where

$$ \mathcal{F}_0(a) \equiv \left\{ f(a) \in \mathcal{F}(a) \,\middle|\, \int_0^1 f(a)\,\mathrm{d}a = 1 \right\}. $$

Then, for some $C \in \mathbb{R}$, it holds

$$ \mathsf{D}\left(f(a)\,\middle\|\,{}^{I}f(a)\right) = C \tag{B.7} $$

iff $f(a) \in \mathcal{F}_0(a)$. Let $A, B \subset [0,1]$ be disjoint measurable sets such that $\int_A \mathrm{d}a = \int_B \mathrm{d}a > 0$, and denote $m \equiv \int_A \mathrm{d}a$. Consider a pair of pdfs $f_1(a), f_2(a) \in \mathcal{F}_0(a)$ defined by

$$ f_1(a) = \frac{1}{m} I_A(a), \quad f_2(a) = \frac{1}{m} I_B(a). $$

Because $\int f_1(a) \ln f_1(a) \mathrm{d}a = \int f_2(a) \ln f_2(a) \mathrm{d}a$, it follows from (B.7) that

$$ \int_A \ln {}^{I}f(a) \mathrm{d}a = \int_B \ln {}^{I}f(a) \mathrm{d}a. $$

Let us denote $q \equiv \int_A \ln {}^{I}f(a) \mathrm{d}a$. Now, consider another pair of pdfs $f_3(a), f_4(a) \in \mathcal{F}_0(a)$ defined by

$$ f_3(a) = \frac{u}{m} I_A(a) + \frac{1-u}{m} I_B(a), \quad f_4(a) = \frac{v}{m} I_A(a) + \frac{1-v}{m} I_B(a), $$

for arbitrary $u, v \in (0, 0.5)$ *such that* $u \neq v$. *For* $f_3(a), f_4(a)$ *it holds*

$$\mathsf{D}\left(f_3(a)\middle|\middle|\,{}^{I}f(a)\right) = u \ln \frac{u}{m} + (1-u) \ln \frac{1-u}{m} - q, \tag{B.8}$$

$$\mathsf{D}\left(f_4(a)\middle|\middle|\,{}^{I}f(a)\right) = v \ln \frac{v}{m} + (1-v) \ln \frac{1-v}{m} - q. \tag{B.9}$$

From (B.7), (B.8), *and* (B.9) *it follows that*

$$u \ln u + (1-u)\ln(1-u) = v \ln v + (1-v)\ln(1-v),$$

which is in a contradiction with the assumption $u \neq v$ *as the function* $x \ln x + (1-x)\ln(1-x)$ *is injective on the interval* $(0, 0.5)$.

The FPD can be also hardly used if the objectives are to maximize the expected value of some of the involved quantities. Such objectives could be suitably expressed by a strictly decreasing loss function. In case of the FPD, it is argued that for this purpose it is sufficient to select an ideal pdf concentrated close to an unreachable (high) value of the considered quantity. This approach has indeed some defects. First of all, nothing in the formulation of the FPD ensures that such unreachable value must exist. Furthermore, the unreachability of some value depends on the outer model of the system. For different outer models, the same objectives would be described by different ideal pdfs. Finally, this approach does not work in general, as it is illustrated in the following example.

Example B.0.2 *Let* $\Delta^* \equiv \mathbb{R}, a^* \equiv [0, 1]$, *and the outer model of the system be*

$$f(\Delta|\, a) \equiv \delta(\Delta - a). \tag{B.10}$$

The objective is to maximize the expected value of Δ.

In order to avoid the above discussed difficulties with specifying ${}^{I}f(\Delta|\, a)$, *we use here the version of the FPD based on* (B.6). *In this case, for the optimal randomized action* ${}^{O}f(a)$ *it must hold*

$${}^{O}f(a) \in \underset{f(a)\in f^*(a)}{\arg\min}\, \mathsf{D}\left(\int f(\Delta|\, a)\, f(a)\, \mathrm{d}a \middle|\middle|\, {}^{I}f(\Delta)\right), \tag{B.11}$$

where

$$f^*(a) = \left\{ f(a) \in \mathcal{F}(a) \,\middle|\, \int_0^1 f(a)\, \mathrm{d}a = 1 \right\}.$$

As for the outer model (B.10) *it holds*

$$\int f(\Delta|\, a)\, f(a)\, \mathrm{d}a = f(a)\,|_{a=\Delta},$$

the Kullback-Leibler divergence in (B.11) *has a form*

$$\mathsf{D}\left(\int f(\Delta|\, a)\, f(a)\, \mathrm{d}a \middle|\middle|\, {}^{I}f(\Delta)\right)$$
$$= \int_0^1 f(a)\,|_{a=\Delta} \ln \frac{f(a)\,|_{a=\Delta}}{{}^{I}f(\Delta)} \mathrm{d}\Delta = -\ln K + \mathsf{D}\left(f(a)\,|_{a=\Delta} \middle|\middle| \frac{{}^{I}f(\Delta)}{K} I_{[0,1]}(\Delta)\right),$$

where

$$K \equiv \int_0^1 {}^{I}f(\Delta)\, \mathrm{d}\Delta.$$

Then, due to (2.3), *it holds that*

$${}^{O}f(a) = \frac{{}^{I}f(\Delta)\,|_{\Delta=a}}{K} I_{[0,1]}(a). \tag{B.12}$$

From (B.12) *it is clear that concentrating the ideal pdf above the upper bound of* Δ *has no impact on the resulting optimal randomized action* ${}^{O}f(a)$. *In this case, the only property of the ideal pdf* ${}^{I}f(\Delta)$ *which influences the resulting randomized action is its shape on* $[0, 1]$.

Some of the discussed weaknesses of the FPD can be eliminated by a generalization of the FPD in which a set of ideal pdfs ${}^{I}\mathcal{F}(\Delta, a)$ is used instead of a single one. The optimal action is then defined by

$${}^{O}f\left(a\right) \in \arg\min_{f(a)\in\mathcal{F}(a)} \left(\min_{{}^{I}f(a,\Delta)\in {}^{I}\mathcal{F}(a,\Delta)} \mathsf{D}\left(f\left(\Delta|\, a\right) f\left(a\right) \,||\, {}^{I}f\left(a,\Delta\right)\right)\right). \tag{B.13}$$

For example, equality of preferences, discussed in Example B.0.1, can be expressed by

$${}^{I}\mathcal{F}(a) \equiv \left\{ f\left(a\right) \in \mathcal{F}(a) \,\middle|\, \int_0^1 f\left(a\right) \mathrm{d}a = 1 \right\}.$$

On the other hand, it seems that the problem discussed in Example B.0.2 cannot be satisfactorily solved even with this generalization. Moreover, contrary to the FPD using a single ideal pdf, the generalization (B.13) do not have an analytical solution.

As discussed in the beginning of this paragraph, the incapability of the FPD to formulate certain kind of objectives that can be easily expressed in terms of loss functions is a consequence of the fact that the FPD is not an extension of a decision making on data. Furthermore, in the FPD the ordering of preferences of randomized actions is given by a specific function on randomized actions – the Kullback-Leibler divergence (B.2) from an ideal pdf. This function is non-linear for any ideal pdf. Contrary to that, in the Bayesian decision making the ordering of preferences is given by the expected loss (B.1), which is always a linear function on randomized actions. From this point of view, the FPD and the Bayesian decision making form completely different approaches.

Bibliography

[1] M. Abramowitz and I.A. Stegun. *Handbook of Mathematical Functions*. Dover Publications, New York, 1972.

[2] Shun-Ichi Amari and Hiroshi Nagaoka. *Methods of Information Geometry*. American Mathematical Society, Providence, 2007.

[3] J. Andrýsek. Approximate recursive Bayesian estimation of dynamic probabilistic mixtures. In J. Andrýsek, M. Kárný, and J. Kracík, editors, *Multiple Participant Decision Making*, pages 39–54, Adelaide, May 2004. Advanced Knowledge International.

[4] J. Andrýsek. Projection based algorithms for estimation of complex models. In *Proceedings of the 5th International PhD Workshop on Systems and Control - a Young Generation Viewpoint*, pages 5–10, Budapest, September 2004. Hungarian Academy of Sciences.

[5] J. Andrýsek. Projection Based Estimation of Dynamic Probabilistic Mixtures. Technical Report 2098, ÚTIA AV ČR, Praha, 2004.

[6] K.J. Arrow. *Social Choice and Individual Values*. New Haven: Yale University Press, 1995. 2nd. ed.

[7] R. Bellman. *Introduction to the Mathematical Theory of Control Processes*. Academic Press, New York, 1967.

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

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

[10] J.M. Bernardo and A.F.M. Smith. *Bayesian Theory*. John Wiley & Sons, Chichester, New York, Brisbane, Toronto, Singapore, 2 edition, 1997.

[11] Robert F. Bordley. A multiplicative formula for aggregating probability assessments. *Management Science*, 28(10):1137–1148, 1982.

[12] Robert F. Bordley and Ronald W. Wolff. On the aggregation of individual probability estimates. *Management Science*, 27(8):959–964, 1981.

[13] Robert T. Clemen. Combining overlapping information. *Management Science*, 33(3):373–380, 1987.

[14] Robert T. Clemen and Robert L. Winkler. Limits for the precision and value of information from dependent sources. *Operations Research*, 33(2):427–442, 1985.

[15] Robert T. Clemen and Robert L. Winkler. Aggregating point estimates: A flexible modeling approach. *Management Science*, 39(4):501–515, 1993.

[16] M.H. DeGroot. *Optimal Statistical Decisions*. McGraw-Hill, New York, 1970.

[17] A. P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

[18] Persi Diaconis and Donald Ylvisaker. Conjugate priors for exponential families. *The Annals of Statistics*, 7(2):269–281, 1979.

[19] P. Ettler, M. Kárný, and T. V. Guy. Bayes for rolling mills: From parameter estimation to decision support. In P. Horáček, M. Šimandl, and P. Zítek, editors, *Preprints of the 16th World Congress of the International Federation of Automatic Control*, pages 1–6, Prague, July 2005. IFAC.

[20] P.C. Fishburn. *Utility Theory for Decision Making*. J. Wiley, New York, London, Sydney, Toronto, 1970.

[21] Christian Genest and Mark J. Schervish. Modeling expert judgments for bayesian updating. *Annals of Statistics*, 13(3):1198–1212., 1985.

[22] Christian Genest and James V. Zidek. Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, 1(1):114–148, 1986.

[23] J. Grim. On numerical evaluation of maximum likelihood estimates for finite mixtures of distributions. *Kybernetika*, 18:173–190, 19982.

[24] J. Heřmanská and L. Jirsa. Improved planning of radioiodine therapy for thyroid cancer — Reply. *Journal of Nuclear Medicine*, 43(5):714–714, May 2002. Reply to letters to editor.

[25] E.T. Jaynes. *Probability Theory - The Logic of Science*. Cambridge University Press, Cambridge, United Kingdom, 2003.

[26] Mohamed N. Jouini and Robert T. Clemen. Copula models for aggregating expert opinions. *Operations Research*, 44(3):444–457, 1996.

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

[28] M. Kárný, J. Andrýsek, A. Bodini, T. V. Guy, J. Kracík, and F. Ruggeri. How to exploit external model of data for parameter estimation? *International Journal of Adaptive Control and Signal Processing*, 20(1):41–50, 2006.

[29] M. Kárný, J. Böhm, T. V. Guy, L. Jirsa, I. Nagy, P. Nedoma, and L. Tesař. *Optimized Bayesian Dynamic Advising: Theory and Algorithms*. Springer, London, 2005.

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

[31] M. Kárný and J. Kracík. A normative probabilistic design of a fair governmental decision strategy. *Journal of Multi-Criteria Decision Analysis*, 12(2-3):1–15, 2004.

[32] R.L. Keeney. A group preference axiomatization with cardinal utility. *Management Science*, 23(2):140–145, 1976.

[33] R.L. Keeney and C.W. Kirkwood. Group decision making using cardinal social welfare functions. *Management Science*, 22(4):430–437, 1975.

[34] R.L. Keeny and H. Raiffa. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. J. Wiley, New York, 1978.

[35] D.F. Kerridge. Inaccuracy and inference. *Journal of Royal Statistical Society*, B 23:284–294, 1961.

[36] J. Kracík. Processing of expert information in bayesian parameter estimation. In *Proceedings of the 6th International PhD Workshop on Systems and Control - a Young Generation Viewpoint*, page 4, Ljubljana, December 2005. Josef Stefan Institute.

[37] J. Kracík and M. Kárný. Merging of data knowledge in Bayesian estimation. In J. Filipe, J. A. Cetto, and J. L. Ferrier, editors, *Proceedings of the Second International Conference on Informatics in Control, Automation and Robotics*, pages 229–232, Barcelona, September 2005. INSTICC.

[38] R. Kulhavý. *Recursive Nonlinear Estimation: A Geometric Approach*, volume 216 of *Lecture Notes in Control and Information Sciences*. Springer-Verlag, London, 1996.

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

[40] F. Liese and I. Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, 2006.

[41] Dennis Lindley. Reconciliation of probability distributions. *Operations Research*, 31(4):866–880, 1983.

[42] Jan R. Magnus and Heinz Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. J. Wiley & Sons, New York, London, Sydney, Toronto, 1999.

[43] Peter A. Morris. Decision analysis expert use. *Management Science*, 20(9):1233–1241, 1974.

[44] I. Nagy, M. Kárný, P. Nedoma, and Š. Voráčová. Bayesian estimation of traffic lane state. *International Journal of Adaptive Control and Signal Processing*, 17(1):51–65, 2003.

[45] Erik Quaeghebeur and Gert de Cooman. Imprecise probability models for inference in exponential families. In *Proceedings of the 4th Symposium on Imprecise Probabilities and Their Applications*, pages 1–10, Pittsburgh, Pennsylvania, USA, July 2005. Carnegie Mellon University.

[46] C.R. Rao. *Linear method of statistical inference and their applications*. Academia, Prague, 1987. in Czech.

[47] G. Schaffer. Savage revisited. *Statistical Science*, 1(4):463–501, 1986.

[48] J. Šindelář and M. Kárný. Adaptive control applied to financial market data. In *Advanced Mathematical Methods for Finance 2007*. European Science Foundation, 2007.

[49] Jan Šindelář, I. Vajda, and M. Kárný. Stochastic control optimal in the Kullback sense. *Kybernetika*, 2008. submitted.

[50] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Champan & Hall, London, New York, 1991. ISBN 0 412 286602 (HB).

[51] Robert L. Winkler. Combining probability distributions from dependent information sources. *Management Science*, 27(4):479–488, 1981.

Contents