CZECH TECHNICAL UNIVERSITY,
FACULTY OF ELECTRICAL ENGINEERING

METHODS OF PROBABILISTIC KNOWLEDGE INTEGRATION

PHD THESIS

PRAGUE, DECEMBER, 1999

JIŘÍ VOMLEL

CZECH TECHNICAL UNIVERSITY,
FACULTY OF ELECTRICAL ENGINEERING

METHODS OF PROBABILISTIC KNOWLEDGE INTEGRATION

PHD THESIS

AUTHOR JIŘÍ VOMLEL
PHD ADVISER RADIM JIROUŠEK
STUDY BRANCH ARTIFICIAL INTELLIGENCE AND BIOCYBERNETICS

PRAGUE, DECEMBER, 1999

Acknowledgments

I would like to thank my PhD adviser Radim Jiroušek. He put me on the track of interesting open problems and taught me how to solve them. He read carefully through my thesis and proposed many improvements to the text. I am grateful to all members of the research teams in Laboratory of Intelligent Systems at the University of Economics and in the Institute of Information Theory and Automation. I highly appreciate the consultations with František Matúš and Milan Studený. My research was supported by a grant of the Ministry of Education of the Czech Republic No. VS 96008 and a grant of the Grant Agency of the Academy of Sciences of the Czech Republic No. A 1075801 "Entropy functions and polymatroids".

Introduction

The main purpose of the thesis is a design and characterization of methods applicable to the process called *knowledge integration*. It is supposed that *knowledge* is represented by a set of probability distributions. Probabilistic models have the ability to handle uncertain information that systems used in artificial intelligence ought to posses. Within the probabilistic framework, *knowledge integration* is understood as a process of building a *joint probability distribution* from an *input set* of low-dimensional distributions. Finally, the process should yield the joint probability distribution embodying knowledge about a chosen domain.

Such a task can be solved by means of iterative procedures. The best-known example is the *Iterative Proportional Fitting Procedure* (IPFP). It was proved by Csizsár [10] (for the case of discrete probability distributions) that IPFP converges if the *input set* of low-dimensional probability distributions is *consistent*, i.e. if there exists a probability distribution whose marginals equal to the probability distributions from the *input set*. If the initial distribution is decomposable and all the input distributions corresponds to this decomposable model then the distributions computed during IPFP process are *decomposable* as well, so that they can be effectively stored and manipulated. *Decomposability* makes possible the fast computation of conditional probabilities possible. Csiszár showed that IPFP converges to the distribution that is I_1-projection of initial distribution to the set of distributions having their marginals equal to probability distributions from *input set*. Furthermore, if the initial distribution is uniform (or a product of positive one-dimensional probability distributions), then the resulting distribution maximizes *Shannon entropy* within the set of distributions having its marginals equal to the probability distributions from *input set*. Arguments in favor of *maximum entropy principle* are based on the fact that the distribution maximizing *Shannon entropy* contains least additional information[20, 23]. In Chapter 3, description of IPFP and its behavior on both the *consistent* and *inconsistent input set* is examined.

Generally, I-divergence geometry, introduced by Csiszár [10] (using I_1-*projection*) and Čencov [7] (using I_2-*projection*), extended by Matúš [32, 33] is an effective instrument enabling us to study *iterative procedures* since these procedures are based on the operation of I-projection to a set of probability distributions. In Section 1.4, we introduce the *additively constrained sets* and the *multiplicatively constrained sets*. In Chapter 2, I-projections to both the *additively* and *multiplicatively constrained sets* undergo an extensive exploration. Though, most of theoretical results in Chapter 2 are not original, it is advantageous to put them together since they were not published within a unifying framework.

A theoretical problem, we are dealing with in Chapter 4, is a behavior of IPFP in the case of *input set* generated by *decomposable generating class*. For this case, it was already proved by S. Haberman [17] that there exists an ordering of elements of the *input set* such that IPFP converges within one *cycle*. Furthermore, for certain decomposable generating classes IPFP converges within one cycle regardless the ordering used. S. Haberman [17] showed that

strongly (totally) decomposable generating classes meet this property. *Strong decomposability* is a sufficient but not a necessary condition. We have proved another sufficient but not necessary condition which extends the *family of generating classes* that posses this property (see Theorem 4.6). We propose a complete characterization of this family. We have rejected the conjecture given in [17] that for all *decomposable generating classes* IPFP converges within two *cycles* regardless from the order of *generating class*.

In the case of an *inconsistent input set* of probability distributions, the IPFP procedure tends to come in cycles. Nevertheless, we wish to get some representative of *input knowledge* even in the case when it is inconsistent. We endeavor to find *iterative procedures* that converge even for *inconsistent input sets*, the marginals of the resulting distribution should be somehow close to the input distributions and independent of the ordering of the input low-dimensional distributions. Application to *consistent input sets* should lead to the same distributions as classical IPFP. In Chapter 5 five iterative methods are described: *convex* and *log-convex conservative modifications* of IPFP, *methods of iterative geometric* and *arithmetic averages*, and a method that is an instance of *GEM-algorithm*. Discussion of their properties and experimental results are given together with several examples. Since *inconsistent input sets* of probability distributions can be derived from *incomplete (missing) data* we conclude the chapter with an overview of methods that are used for *missing data* in statistics. A comparison of these methods with the proposed iterative procedures has been carried out on a simple problem.

Contents

1 Preliminaries

1.1 Mathematical symbols

To make the text easy to read, mathematical notation obeys following rules:

- Low-case characters (e.g. $i, j, k, s, \ldots$) are used for numbers (usually integers), hypergraph (or graph) nodes, etc.
- Upper-case characters (e.g. $A, B, E_i, V, \ldots$) are used for sets of numbers, hypergraph edges, etc.
- Upper-case characters P, Q, R, S, T, and U (possibly with an accent) are reserved for probability distributions. If the superscript is symbol of a set then it means marginalization (e.g. P^A is marginal of P). If the superscript is a symbol of a real number then it means regular exponentiation. If indices are needed to distinguish distribution (especially in iterative processes) then subscripts are in parenthesis (e.g. $Q_{(i+1)}$), otherwise a subscript can be used to define index set of variables for which is a distribution defined (e.g. $P_V \in \mathcal{P}_V$).
- Upper-case character X denotes random variables (e.g. X_i) or more dimensional one ($X^A = (X_i)_{i \in A}$.
- Blackboard bold upper-case character $\mathbb{X}_i$ denote a set of values of a random variable X_i. $\mathbb{X}^A$ means Cartesian product of several variables values' set ($\mathbb{X}^A = \times_{i \in A} \mathbb{X}_i$).
- Low-case characters x_i, y_i denote arbitrary values of a random variable X_i. x^A is a vector of several variables' values ($x^A \in \mathbb{X}^A$).
- Caligraphic fonts are used for sets of sets (classes, families), (e.g.. sets of hypergraph nodes $\mathcal{E}_1, \mathcal{F}(F, \mathcal{E})$).
- Caligraphic fonts $\mathcal{P}$, $\mathcal{Q}, \mathcal{R}, \mathcal{S}$, and $\mathcal{T}$ are reserved for sets of distributions. Possible subscript somehow characterizes the set. For example, A in $\mathcal{S}_A$ means the set of variables for which the distributions from this set are defined, $\mathcal{E}_i$ in $\mathcal{R}_{\mathcal{E}_i}$ means the class of distributions decomposable with respect to $\mathcal{E}_1$. Their interpretation will be clear from the context.

- Low-case Greek character π with a subscript followed by a symbol of a distribution (e.g. $\pi_{\mathcal{R}}P$) denotes another distribution that is the I_1-projection of distribution P to the class of distributions $\mathcal{R}$ defined in subscript. Symbol π' is used for I_2-projection.

Details can be found in the sections where symbols are defined. In order to find easily a particular definition, index on page 121 may be helpful.

1.2 Graphs and Hypergraphs

In this section all definitions from graph and hypergraph theory, necessary in our text, will be given. Since *unoriented graphs without loops* can be viewed as *hypergraphs* having only edges that connect two nodes we will start with definitions of hypergraphs' properties. Both, graphs and hypergraphs are supposed to be finite.

1.2.1 Hypergraphs

In this subsection, we are going to define two types of acyclicity in hypergraphs, together with definition of decomposable hypergraphs. We will use the standard definition of hypergraphs (Berge [3]).

Definition 1.1 (Hypergraph) *Hypergraph* $H = (V, \mathcal{E})$ is a pair of two finite sets:

- set of vertices V and
- set of hypergraph edges. Edges $E \in \mathcal{E}$ are arbitrary nonempty subsets of V.

In the sequel, the notion of *reduced* and *nontrivial hypergraph* will be needed.

Definition 1.2 (Nontrivial hypergraph) Hypergraph $H = (V, \mathcal{E})$ is *nontrivial* if $\mathcal{E}$ consists of more than one element.

Definition 1.3 (Reduced set) Set $\mathcal{E}$ of subsets of V is *reduced* iff there are no two different subsets $(E, F \in \mathcal{E}, E \neq F)$ such that one is subset of another $(E \subseteq F)$.

The definition of reduced set (Definition 1.3) will be used to farther restrict the notion of hypergraph. Hypergraph where no edge is a proper subset of any other edge are called *reduced hypergraphs*.

Definition 1.4 (Reduced hypergraph) *Hypergraph* $H = (V, \mathcal{E})$ is *reduced* iff set of its edges $\mathcal{E}$ is *reduced*.

In the subsequent text all *hypergraphs* will be supposed to be *reduced*.

There are two basic ways how to construct hierarchy of hypergraphs. Both ways are going to be defined as an operation on the set of hypergraph edges. The first operation is based on *node removal* and it is called *hypergraph restriction* [17]. The second one is based on edge removal and it results in *subhypergraph*.

Definition 1.5 (Restriction of hypergraph) Let $M \subseteq V$ be a set of hypergraph nodes. *Restriction of hypergraph* $H = (V, \mathcal{E})$ is hypergraph $H' = (M, \mathcal{E}(M)$, where

$$\mathcal{E}(M) = \{E \cap M : E \in \mathcal{E}\}$$

is *the set of edges generated by set* M.

Definition 1.6 (Subhypergraph) The hypergraph $H' = (V', \mathcal{E}')$ is *subhypergraph* of hypergraph $H = (V, \mathcal{E})$ iff $\mathcal{E}' \subset \mathcal{E}$ and $V' \subseteq V$.

Definition 1.7 (Path in a hypergraph) Given a hypergraph $H = (V, \mathcal{E})$ a *path* from node n to node m is a sequence of distinct edges $(E_1, E_2, \ldots, E_q)$, where $1 \leq i \leq q : E_i \in \mathcal{E}$, such that:

- $n \in E_1$,
- $m \in E_q$, and
- for every $1 \leq i < q : E_i \bigcap E_{i+1} \neq \emptyset$.

Definition 1.8 (Connected hypergraph) Hypergraph $H = (V, \mathcal{E})$ is *connected* if for every pair of edges E_i, E_j there exist a path from node $n \in E_i$ to node $m \in E_j$.

Definition 1.9 (Articulation set) G is an *articulation* set of hypergraph $H = (V, \mathcal{E})$ iff hypergraph $(V \setminus G, \mathcal{E}(V \setminus G))$ is not *connected* and there exist two different edges $E, F \in \mathcal{E}$ such that $G = E \cap F$.

Ronald Fagin [15] defined four basic degrees of acyclicity for relational database schemes. They are *Berge-acyclicity, γ-acyclicity, β-acyclicity*, and *α-acyclicity*. They satisfy the following ordering:

$$\text{Berge-acyclicity} \Rightarrow \gamma\text{-acyclicity} \Rightarrow \beta\text{-acyclicity} \Rightarrow \alpha\text{-acyclicity}.$$

We will take advantage of two last mentioned types of acyclicity. Let us start with α-acyclicity which is hereditary with respect to *node removal*.

Definition 1.10 (α-acyclicity) Hypergraph H is *α-acyclic* iff all its nontrivial, connected restrictions have an articulation set.

Example 1.1 (α-acyclic hypergraph) Figure 1.1 displays an example of α-acyclic hypergraph $H = (\{1, 2, 3, 4, 5, 6\}, \{E_1, E_2, E_3, E_4\})$. For example the articulation set of $\{E_1, E_4\}$ is $\{2, 4\}$. Subset $\mathcal{E}' = \{E_1, E_2, E_3\}$ has no articulation set, but it is not a restriction of H. Observe that $\mathcal{E}' = \{E_1, E_2, E_3\} \subset \mathcal{E} = \{E_1, E_2, E_3, E_4\}$ and therefore $H' = (\{1, 2, 3, 4, 5, 6\}, \mathcal{E}')$ is a cyclic subhypergraph of H.

It is contra intuitive if an acyclic hypergraph contains cyclic subhypergraph. Thus R. Fagin [15] proposed another definition of acyclicity which has not this property. He call it β-acyclicity. It is hereditary not only with respect to *node removal* but also with respect to *edge removal*.

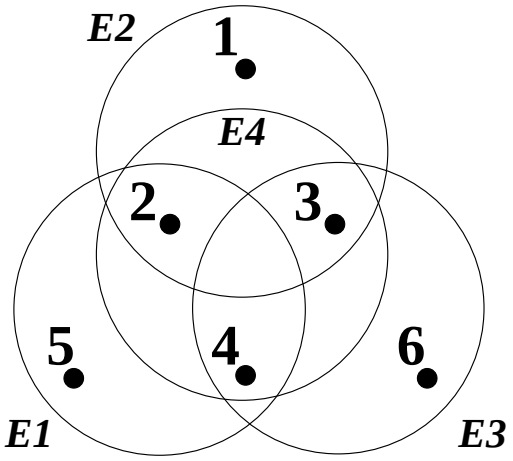


Figure 1.1: α-acyclic hypergraph

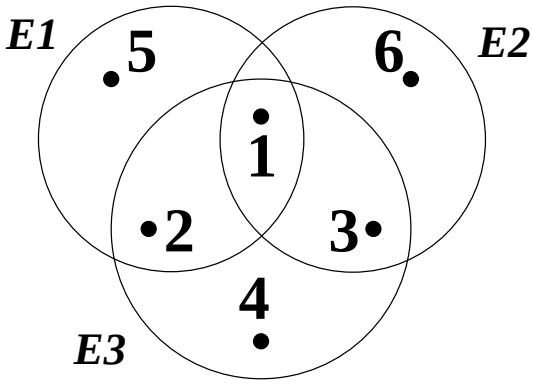


Figure 1.2: β-acyclic hypergraph

Definition 1.11 (β-acyclicity) Hypergraph H is *β-acyclic* iff all nontrivial, connected subsets of the set of edges have an articulation set.

Example 1.2 (β-acyclic hypergraph) An example of β-acyclic hypergraph is given in Figure 1.2.

S. Lauritzen et al. [30] defined *decomposability* of hypergraphs, equivalently to the definition of *decomposable generating class* given by Shelby Haberman [17].

Definition 1.12 (Decomposable hypergraph) Hypergraph $H = (V, \mathcal{E})$ is *decomposable* if either has one edge or is the union of two disjoint decomposable hypergraphs $H_1 = (V_1, \mathcal{E}_1)$ and $H_2 = (V_2, \mathcal{E}_2)$, i.e. $V_1 \subseteq V$,$V_2 \subseteq V$, $V = V_1 \cup V_2$, $\mathcal{E}_1 \cup \mathcal{E}_2 = \mathcal{E}$, and $\mathcal{E}_1 \cap \mathcal{E}_2 = \emptyset$, such that there exist edges $E_1 \in \mathcal{E}_1$ and $E_2 \in \mathcal{E}_2$ that:

$$((\cup_{E' \in \mathcal{E}_1} E') \cap (\cup_{E'' \in \mathcal{E}_2} E'')) = E_1 \cap E_2.$$

Figure 1.3 displays a decomposable graph. A sequence of intersections that can decompose the hypergraph is

$$E_1 \cap E_5, \quad E_2 \cap E_5, \quad E_3 \cap E_5, \quad E_4 \cap E_5$$

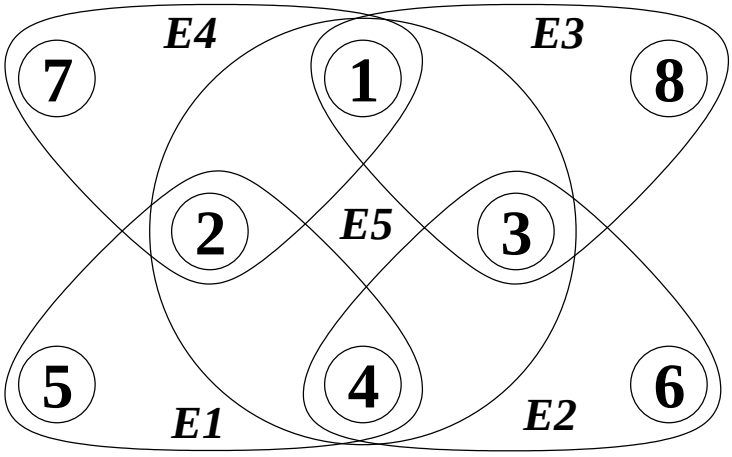


Figure 1.3: Decomposable hypergraph

In [30] it was shown that α-*acyclicity* is equivalent to hypergraph *decomposability*.

Theorem 1.1 *Hypergraph H is α-acyclic iff it is decomposable.*

Remark. Notice that notions of acyclicity and decomposability of hypergraphs are properties of hypergraph edges only. Therefore, in the sequel, we will often use these notions for sets of subsets of V. When we use, for example, $\mathcal{E}$ is decomposable, it means that hypergraph $H = (V', \mathcal{E})$ is *decomposable* for $V' = \cup_{E \in \mathcal{E}} E$.

Orderings of $\mathcal{E}$ that obey*Running Intersection Property*, which we are going to define, has several useful properties exploited particularly in Chapter 4.

Definition 1.13 (Running Intersection Property) Let E_k for $k = 1, \ldots, s$ be an ordering of sets from $\mathcal{E}$. This ordering meets the *Running Intersection Property (RIP)* if

$$\forall l = 2, \ldots, s \quad \exists k \quad (1 \leq k < l) \quad ((E_l \cap \bigcup_{m=1}^{l-1} E_m) \subset E_k).$$

Example 1.3 (Running Intersection Property) A possible ordering of edges of hypergraph from Figure 1.3 that meets RIP is $\{E_1, E_5, E_2, E_3, E_4\}$:

- $l = 2 : E_5 \cap E_1 \subset E_1$
- $l = 3 : E_2 \cap (E_1 \cup E_5) \subset E_5$
- $l = 4 : E_3 \cap (E_1 \cup E_5 \cup E_2) \subset E_5$
- $l = 5 : E_4 \cap (E_1 \cup E_5 \cup E_2 \cup E_3) \subset E_5$

For every decomposable hypergraph there exist orderings of its edges such that a property called Running Intersection Property (RIP) holds (Lemma 1.1). This property was proved by several authors, we refer at least to [17, 23].

Lemma 1.1 (Existence of orderings satisfying RIP) *Suppose $\mathcal{E}$ is decomposable. For any $E \in \mathcal{E}$ the sets in $\mathcal{E}$ can be ordered so that $E_1 = E$ and the ordering $\{E_1, E_2, \dots E_s\}$ satisfies RIP.*

Operation *join* and *meet* between two hypergraphs together with a relation of partial order for hypergraphs can be defined [30, 20].

Definition 1.14 (Join and meet operations)
Let $H_1 = (V_1, \mathcal{E}_1)$ and $H_2 = (V_2, \mathcal{E}_2)$ be two hypergraphs. The *join* operation $\vee$ is defined by

$$(V_3, \mathcal{E}_3) = (V_1, \mathcal{E}_1) \vee (V_2, \mathcal{E}_2), \quad \text{where}$$

$$V_3 = V_1 \cup V_2 \quad \text{and} \quad \mathcal{E}_3 \quad \text{is reduced set of} \quad \mathcal{E}_1 \cup \mathcal{E}_2.$$

The *meet* operation $\wedge$ is defined by

$$(V_4, \mathcal{E}_4) = (V_1, \mathcal{E}_1) \wedge (V_2, \mathcal{E}_2), \quad \text{where}$$

$$V_4 = V_1 \cap V_2 \quad \text{and} \quad \mathcal{E}_4 \quad \text{is reduced set of} \quad \mathcal{E}_1 \cap \mathcal{E}_2.$$

Definition 1.15 (Relation of partial order of hypergraphs)
If $V_1 \subseteq V_2$ and for every $E_i \in \mathcal{E}_1$ there exists $E_j \in \mathcal{E}_2$ such that $E_i \subseteq E_j$ then we write

$$(V_1, \mathcal{E}_1) \leq (V_2, \mathcal{E}_2)$$

.

If for $j = 1, 2, 3, 4 : V_j = \cup_{E \in \mathcal{E}_j} E$ we can omit conditions for V_j in the definitions of *join*, *meet*, and *partial order* of hypergraphs. Then these definitions can be understood as definitions of operations between sets of subsets of V and we will write $\mathcal{E}_3 = \mathcal{E}_1 \vee \mathcal{E}_2$, $\mathcal{E}_4 = \mathcal{E}_1 \wedge \mathcal{E}_2$, and $\mathcal{E}_1 \leq \mathcal{E}_2$, respectively.

1.2.2 Graphs

All graphs in this text are supposed to be unoriented graphs without loops and multiple edges (so called simple graphs).

Definition 1.16 (Graph) Graph $G = (V, \mathcal{E})$ consist of set of vertices V and set of unoriented pairs (called edges) $(v_i, v_j) \in \mathcal{E}, v_i, v_j \in V, v_i \neq v_j,$.

Definition 1.17 (Complete graph) *Complete graph* is graph, where all nodes are adjacent by an edge.

Definition 1.18 (Clique of a graph) *Clique* is a maximal complete subgraph of a graph.

Remark. The word maximal in the previous definition is to be understood that the set of nodes can not be extended by another node without violating the condition of completeness.

An example of graph is given in Figure 1.4. The graph has three cliques:

$$E_1 = \{1, 2, 3, 4\}, E_2 = \{1, 2, 5\}, \quad \text{and} \quad E_3 = \{1, 3, 6\}.$$

A *subgraph* is a just a special case of *subhypergraph*. *Path in a graph* is a sequence of graph edges that is a special case of sequence of *partial hypergraph edges*. *Cycle* in a graph is a *path* beginning and ending at the same node.

Definition 1.19 (Cycle in a graph)
Given a graph G, a *cycle* c in G is sequence $(v_1, v_2, \ldots, v_q)$, $q \geq 3$ of distinct graph nodes such that for every $1 \leq i < q$, (v_i, v_{i+1}) are graph edges and (v_1, v_q) is a graph edge too. A cycle is even if q is even.

If we distinguish between two types of chord we can define two special classes of graphs: *chordal graphs*, which are also called *triangulated graphs*, and *strongly chordal graphs*.

Definition 1.20 (Chord)
Chord in a cycle $c = (v_1, v_2, \ldots, v_q)$ is an edge (v_i, v_j), such that $1 < |i - j| < q - 1$.

Definition 1.21 (Chordal graph)
A *chordal graph* (triangulated graph) is a graph in which every cycle with at least four distinct nodes has a chord.

Definition 1.22 (Strong chord)
Strong chord in even cycle $c = (v_1, v_2, \ldots, v_q)$ is an edge (v_i, v_j), such that $1 < |i - j| < q - 1$ and $|i - j|$ is odd (*i.e.*$|i - j| = 3, 5, 7, \ldots$).

Definition 1.23 (Crossing chords)
Let $c = (v_1, v_2, \ldots, v_q), q \geq 4$ is a cycle of a graph G, (v_i, v_j) and (v_k, v_l) are graph edges such that $1 \leq i < j \leq q$, and $1 \leq k < l \leq q$. If $(i < k, j < l, j \neq k)$ or $(k < i, l < j, i \neq l)$ then edges (v_i, v_j) and (v_k, v_l) are called *crossing chords in cycle c*.

Definition 1.24 (Strongly chordal graph)
A *strongly chordal graph* is a graph that in every even cycle with at least six nodes contains a strong chord.

Example 1.4 (Chordal graphs) Figures 1.4 and 1.5 display chordal graphs. For example cycle $(1, 2, 4, 3)$ in graph on Figure 1.4 has two chords $(1, 4)$ and $(2, 3)$. This graph is also strongly chordal. In the only cycle with at least six nodes $(1, 5, 2, 4, 3, 6)$ it has strong chord $(1, 4)$. Graph from Figure 1.5 is not strongly chordal. All its even cycles with six nodes contains a strong chord. But the only cycle with eight nodes $(1, 7, 2, 5, 4, 6, 3, 8)$ does not contain any strong chord. All chords in this cycle $(1, 2)$, $(1, 3)$, $(1, 4)$, $(2, 3)$, $(2, 4)$, and $(3, 4)$ are not strong since their distance in the cycle is always even.

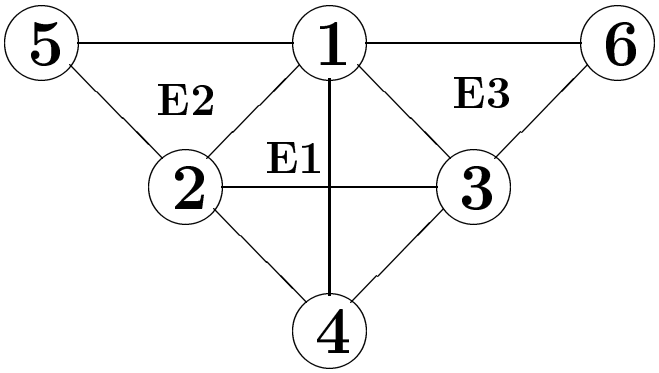


Figure 1.4: Example of a chordal graph

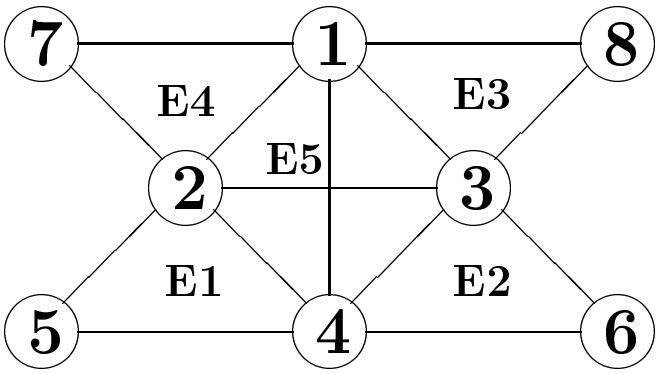


Figure 1.5: Another example of a chordal graph

1.2.3 Relation between acyclic hypergraphs and chordal graphs

In this subsection we will utilize the fact that every α-*acyclic* hypergraph can be uniquely transformed to a graph. Consequently, the study of different classes of acyclic hypergraphs can exploit results from Graph Theory.

Definition 1.25 (Graph of a hypergraph)
A graph $G(H)$ is *graph of hypergraph* $H = (V, \mathcal{E})$ iff it has the same vertex set V and two vertices are joined by an edge if they occur together in some edge $E \in \mathcal{E}$.

Definition 1.26 (Conformal hypergraph) Hypergraph H is *conformal* iff cliques of its graph $G(H)$ are exactly edges of hypergraph H.

The following theorem given independently by several authors e.g. [41, 2] can be used to establish bijection between acyclic hypergraphs and graphs.

Theorem 1.2 *All α-acyclic hypergraphs are conformal.*

Consider the mapping that associates to each graph G a hypergraph H of its maximal cliques. This mapping is bijection. Inverse mapping associates to each conformal hypergraph H its graph $G(H)$ (see [30, 2]).

The following two theorems gives relation between chordal graphs and acyclic hypergraphs. The proofs can be found in [12].

Theorem 1.3 *Hypergraph H is α-acyclic iff its graph $G(H)$ is chordal.*

Theorem 1.4 *Hypergraph H is β-acyclic iff its graph $G(H)$ is strongly chordal.*

1.3 Discrete Probability Distributions

We begin this section with description of notation used for probability distributions in this text. Standard Kolmogorov's definition of probability is used. Details can be found in any textbook of Mathematical Statistics or Probability Theory (e.g. in [42]).

Let $(\Omega, \mathcal{A}, P)$ be a *probability space*, where Ω is a set of elementary random events, A is a *Boolean algebra* of subsets of Ω, and P is a *probability measure*. If $\mathbb{X}$ is a finite set of real numbers then *finite random variable* $X(\omega)$ is a real function $X : \Omega \mapsto \mathbb{X}$ such that for each $x \in \mathbb{X}$ the event $E_x = \{\omega \in \Omega : X(\omega) = x\}$ belongs to $\mathcal{A}$. The set $\mathbb{X}$ is called *sample space* of $X(\omega)$. In this paper all random variables are supposed to be finite random variables (possibly multidimensional).

For $V = \{1, 2, \ldots, s\}$, $P(X_1, \ldots, X_s)$ (abbreviated P_V) will denote s-dimensional *probability distribution* of a s-dimensional random variable

$$\{X_i\}_{i \in V} : \Omega \mapsto \times_{i \in V} \mathbb{X}_i.$$

Thus for every $x = (x_1, \ldots x_s) \in \times_{i \in V} \mathbb{X}_i$

$$P(x) = P(X_1 = x_1, \ldots, X_s = x_s) = P(\{\omega : X_i(\omega) = x_i \quad \text{for all} \quad i = 1, 2, \ldots, s\}).$$

Let $A \subseteq V$. Symbol $\mathcal{P}_A$ denotes the set of all probability distributions defined for $\{X_i\}_{i \in A}$. Symbol $\mathcal{P}_V$ is going to be abbreviated to $\mathcal{P}$. Probability $P(X_1 = x_1, \ldots, X_k = x_k)$ will be often abbreviated to $P(x)$. For $x \in \mathbb{X} = \mathbb{X}^V$, $x^A = (x_i)_{i \in A}$ will denote coordinate projection of $x = (x_i)_{i \in V} \in \mathbb{X}$ on $\mathbb{X}^A = \times_{i \in A} \mathbb{X}_i$.

Definition 1.27 (Marginal probability distribution) Let $L \subseteq K \subseteq V$ and $P_K \in \mathcal{P}_K$ be probability distribution. By $P_K^L(x^L) \in \mathcal{P}_L$ we will denote a *marginal probability distribution* of $P_K(y)$, $y = (y_i)_{i \in K}$ defined as

$$P_K^L(x^L) = \sum_{y \in \mathbb{X}_K,\ y^L = x^L} P_K(y).$$

If $L = \emptyset$ then $P_K^L = 1$.

Definition 1.28 (Conditional probability distribution) Let $A, B \subseteq V$ be disjoint and $P \in \mathcal{P}$. Then a system of probability distributions $P^{A|B}(x^A|x^B)$ is called *conditional probability distribution* if it for all $x^A \in \mathbb{X}^A$, and $x^B \in \mathbb{X}^B$ satisfies

$$P^{A|B}(x^A|x^B) \cdot P^B(x^B) = P^{A \cup B}(x^{A \cup B}).$$

The previous definition leaves a conditional probability distribution $P^{A|B}(x^A|x^B)$ undefinite in the case of $P^B(x^B) = 0$.

Definition 1.29 (Probabilistic conditional independence)
Suppose that $P((X_i)_{i\in V}) \in \mathcal{P}$ is a probability distribution. Having a triplet $(A, B|C)$ of pairwise disjoint subsets of V, where A and B are nonempty, we say that $\{X_i\}_{i\in A}$ is *conditionally independent* of $\{X_i\}_{i\in B}$ given $\{X_i\}_{i\in C}$ iff $\forall (x_i)_{i\in V} \in \times_{i\in V}\mathcal{X}_i$:

$$P^{A\cup B\cup C}((x_i)_{i\in A\cup B\cup C}) \cdot P^C((x_i)_{i\in C}) = P^{A\cup C}((x_i)_{i\in A\cup C}) \cdot P^{B\cup C}((x_i)_{i\in B\cup C}).$$

Definition 1.30 (Dominance) Let $A \subseteq V$ and $P, Q \in \mathcal{P}_A$ be two probability distributions. The symbol $P \ll Q$ means that P *is dominated by* Q that is $\{x \in \mathbb{X}_A, P(x) > 0\} \subseteq \{y \in \mathbb{X}_A, Q(y) > 0\}$.

Remark. It is important to note that *probability distributions' dominance* will be used to guarantee that zero in the denominator stands always together with zero in the numerator. In such a case the values of the following expressions, containing such a fraction, will be defined to be

$$\frac{0}{0} = 0 \quad \text{and} \quad 0 \cdot \log \frac{0}{0} = 0.$$

We conclude this section with definition of the well known concepts - *Shannon entropy* $H(P)$, *information content* $I(P)$, *I-divergence*, and *total variance*.

Definition 1.31 (Shannon entropy) Let $P \in \mathcal{P}$ be a probability distribution. *Shannon entropy* is nonnegative real function

$$H(P) = -\sum_{x\in\mathbb{X},\, P(x)>0} P(x) \log P(x).$$

Definition 1.32 (Information content) Let $P \in \mathcal{P}$ be a probability distribution. *Information content* is nonnegative real function

$$I(P) = \sum_{x\in\mathbb{X},\, P(x)>0} P(x) \log \frac{P(x)}{\prod_{i\in V} P^{\{i\}}(x^{\{i\}})}.$$

Several functions to measure distance between two probability distributions can be used. In spite of it we use only two of them in this text. We take advantage of *I-divergence* and *total variance*. I-divergence also called *Kullback-Leibler divergence* or *Cross-entropy* is often used as a distance "measure" in Information Theory [39].

Definition 1.33 (I-divergence)

$$I(P \parallel Q) = \begin{cases} \sum\limits_{x\in\mathbb{X},\, P(x)>0} P(x) \log \dfrac{P(x)}{Q(x)} & \text{if } P \ll Q \\ +\infty & \text{if } P \not\ll Q \end{cases}$$

Let us note that *I-divergence* is not generally symmetric and that $I(P \parallel Q) \geq 0$ with equality only if P and Q are identical. If $\mathcal{T}$ is a *compact set* and $P \in \mathcal{T}$ then the function $I(P \parallel \cdot)$ is continuous and strictly convex.

In contrast to *I-divergence*, *total variance* satisfies *measure properties*. We shall need it to prove IPFP convergence in this chapter.

Definition 1.34 (Total variance)

$$|P - Q| = \sum_{x \in \mathbb{X}} |P(x) - Q(x)|$$

1.4 Sets of Probability Distributions

In this section properties of sets of probability distributions are reviewed. We are going to exploit two basic types: *additively constrained sets* and *multiplicatively constrained sets*. At first, some standard set properties are defined.

Definition 1.35 (Inner point) A probability distribution Q is called an *algebraic inner point* of a set of probability distributions $\mathcal{T} \subseteq \mathcal{P}$ iff for any $P_0 \in \mathcal{T}$ there exist $0 < \alpha < 1$ and $P_1 \in \mathcal{T}$ such that $Q = (1 - \alpha)P_0 + \alpha P_1 \in \mathcal{T}$.

Definition 1.36 (Compact set) A set of probability distributions $\mathcal{T} \subseteq \mathcal{P}$ is *compact* iff any sequence $\{P_n\}_{n=1}^{\infty}$ of elements from $\mathcal{T}$ contains a convergent subsequence with the limit belonging to the set $\mathcal{T}$.

Definition 1.37 (Neigbourhood) Let $\mathcal{T} \subseteq \mathcal{P}$ and $P \in \mathcal{T}$. $\mathcal{T}$ is *neigbourhood* of P iff it contains an open set which contains P.

Definition 1.38 (Closed set) A set of probability distributions $\mathcal{T} \subseteq \mathcal{P}$ is *closed* iff all distributions $P \in \mathcal{P}$, such that their every *neigbourhood* that contains at least one $Q \neq P$, $Q \in \mathcal{T}$, belongs to $\mathcal{T}$.

Every compact set $\mathcal{T} \subseteq \mathcal{P}$ is closed. A proof can be found for example in Yosida [45, p. 5].

In the following chapters, namely in Chapter 2, I-projections on *additively* and *multiplicatively constrained sets* of probability distributions are frequently used. The definitions of these sets follows.

Definition 1.39 (Additively constrained set)
Let $\mathcal{E} = \{E_1, E_2, \ldots, E_s\}$, $\bigcup_{i=1}^{s} E_i \subseteq A \subseteq V$, and $S_{E_i} \in \mathcal{P}_{E_i}, i = 1, 2, \ldots, s$. *Additively constrained set* $\mathcal{S}_{\mathcal{E},A}$ is a subset of the set $\mathcal{P}_A$ restricted by $S_{E_i}, i = 1, 2, \ldots, s$:

$$\mathcal{S}_{\mathcal{E},A} = \{P \in \mathcal{P}_A, P^{E_1} = S_{E_1}, \ldots, P^{E_s} = S_{E_s}\}.$$

It may happen that distributions $S_{E_i}, i = 1, 2, \ldots, s$ have got such values that set $\mathcal{S}_{\mathcal{E},A} = \emptyset$. Set $\mathcal{P}$ is defined to be *additively constrained* as well (with $s = 0$). In the sequel, if $A = V$ then $\mathcal{S}_{\mathcal{E},V}$ will be abbreviated to $\mathcal{S}_{\mathcal{E}}$. Lemma 1.2 states that every *additively constrained set* is *affine*.

Definition 1.40 (Affine set) A set of probability distributions $\mathcal{S} \subseteq \mathcal{P}_A, A \subseteq V$ is *affine* iff for any $P_0, P_1 \in \mathcal{S}$ and for any $\alpha \in \mathbb{R}$ such that $P_\alpha(x^A) = (1-\alpha)P_0(x^A) + \alpha P_1(x^A)$ is a probability distribution, $P_\alpha \in \mathcal{S}$.

Affine sets are also called *linear sets* (e.g. in Cziszar [10]).

Definition 1.41 (Convex set) A set of probability distributions $\mathcal{S} \subseteq \mathcal{P}_A, A \subseteq V$ is *convex* iff for any $P_0, P_1 \in \mathcal{S}$, and for every $0 < \alpha < 1 : P_\alpha(x^A) = (1-\alpha)P_0(x^A) + \alpha P_1(x^A) \in \mathcal{S}$.

It is obvious that every affine set is convex. It follows from the next two lemmata, that within the framework of our consideration (i.e. within sets of discrete probability distributions) *additively constrained sets* are equivalent to *affine sets*.

Lemma 1.2 (Affinity of a additively constrained set)
Every additively constrained set $\mathcal{S}_{\mathcal{E},A}$ is affine.

Proof. Take any $\alpha \in \mathbb{R}$, $0 < i \leq s$, and $P_0(x^A), P_1(x^A) \in \mathcal{S}_{\mathcal{E},A}$. Since $P_0^{E_i}(x^{E_i}) = S^{E_i}(x^{E_i}) = P_1^{E_i}(x^{E_i})$, we can get

$$P_\alpha^{E_i}(x^{E_i}) = (1-\alpha)P_0^{E_i}(x^{E_i}) + \alpha P_1^{E_i}(x^{E_i}) = (1-\alpha)S^{E_i}(x^{E_i}) + \alpha S^{E_i}(x^{E_i}) = S^{E_i}(x^{E_i}).$$

Therefore, any P_α such that it is a probability distribution belongs to $\mathcal{S}_{\mathcal{E},A}$. That is just the definition of an affine set. □

Any affine set $\mathcal{S} \subseteq \mathcal{P}^A$ can be looked as the intersection of simplex in $\times_{i\in A}\mathbb{R}$ representing $\mathcal{P}^A$ with an affine set from $\times_{i\in A}\mathbb{R}$. Therefore $\mathcal{S}$ is closed and can be represented by a finite number of linear constraints. This is an idea of the proof of the following lemma from [10].

Lemma 1.3 *Any affine set $\mathcal{S}$ can be represented by a finite number of linear constraints.*

In the following part of this section we will describe *multiplicatively constrained sets* similarly to the previous part, where *additively constrained sets* were thought about.

Definition 1.42 (Factorization of probability distribution)
Let $\psi_{E_i} : \mathbb{X}^{E_i} \mapsto \mathbb{R}, i = 1, 2, \ldots, s$. Then probability distribution $P \in \mathcal{P}_A, \cup_{i=1}^s E_i \subseteq A \subseteq V$ *factorizes with respect to* $\mathcal{E}$ iff

$$P(x^A) = \prod_{i=1}^{s} \psi_{E_i}(x^{E_i}).$$

Definition 1.43 (Multiplicatively constrained set)
Let $\mathcal{E} = \{E_1, E_2, \ldots, E_s\}$ and $\cup_{i=1}^s E_i \subseteq A \subseteq V$. Then *multiplicatively constrained set* $\mathcal{R}_{\mathcal{E},A}$ is a nonempty subset of the set $\mathcal{P}_A$ containing all distributions from $\mathcal{P}_A$ that *factorize with respect to* $\mathcal{E}$.

The set $\mathcal{P}_A, A \subseteq V$ is defined to be a *multiplicatively constrained* too (with $s = 1$ and $\mathcal{E} = \{A\}$). In the sequel, if $A = V$ then $\mathcal{R}_{\mathcal{E},V}$ is abbreviated to $\mathcal{R}_{\mathcal{E}}$. Every *multiplicatively constrained set* is *log-affine* (see Lemma 1.4. We note that the following definition of *log-affine* set corresponds to the definition of *weakly log-convex* set given in [33]. For usual definition of *log-affine* see [29]. In the case of strictly positive probability distributions both definitions are equivalent. In the following definition the superscripts α and $(1-\alpha)$ are real numbers and therefore they mean regular exponentiation, which, in the case of probability distributions, means that every value of a probability distribution is exponentiated.

Definition 1.44 (Log-affine set) A set of probability distributions $\mathcal{R} \subseteq \mathcal{P}_A, A \subseteq V$ is *log-affine* iff for any $P_0, P_1 \in \mathcal{R}$ such that $\exists x^A \in \mathbb{X}^A, P_0(x^A) > 0, P_1(x^A) > 0$ and for any $\alpha \in \mathbb{R}$

$$P_\alpha(x^A) = \frac{1}{c(P_0, P_1, \alpha)} \cdot P_0^{(1-\alpha)}(x^A) \cdot P_1^{\alpha}(x^A),$$

where $c(P_0, P_1, \alpha) = \sum_{x^A \in \mathbb{X}^A} P_0^{(1-\alpha)}(x^A) \cdot P_1^{\alpha}(x^A)$, belongs to $\mathcal{R}$.

Next, it will be shown that within the framework of sets of discrete probability distributions *multiplicatively constrained sets* are *log-affine*.

Lemma 1.4 (Log-affinity of a multiplicatively constrained set)
Every multiplicatively constrained set $\mathcal{R}_{\mathcal{E},A}$ is log-affine.

Proof. Take any $\alpha \in \mathbb{R}$ and any $P_0, P_1 \in \mathcal{R}_{\mathcal{E},A}$ such that $\exists x^A \in \mathbb{X}^A, P_0(x^A) > 0, P_1(x^A) > 0$. We can write

$$P_0(x^A) = \prod_{i=1}^{s} \psi_{E_i}(x^{E_i}) \quad \text{and} \quad P_1(x^A) = \prod_{i=1}^{s} \phi_{E_i}(x^{E_i}),$$

where $i = 1, 2, \ldots, s : \psi_{E_i}, \phi_{E_i} : \mathbb{X}^{E_i} \mapsto \mathbb{R}$. If we define functions

$$\begin{aligned} \chi_{E_i}(x^{E_i}) &= \psi_{E_i}^{(1-\alpha)}(x^{E_i}) \cdot \phi_{E_i}^{\alpha}(x^{E_i}) \quad \text{for} \quad i = 1, 2, \ldots, s-1 \quad \text{and} \\ \chi_{E_s}(x^{E_i}) &= \psi_{E_s}^{(1-\alpha)}(x^{E_s}) \cdot \phi_{E_s}^{\alpha}(x^{E_s}) \cdot c(P_0, P_1, \alpha) \end{aligned}$$

then we can write:

$$P_\alpha = \frac{1}{c(P_0, P_1, \alpha)} \cdot P_0^{(1-\alpha)} \cdot P_1^{\alpha} = \frac{1}{c(P_0, P_1, \alpha)} \cdot \prod_{i=1}^{s} \psi_{E_i}^{(1-\alpha)}(x^{E_i}) \cdot \prod_{i=1}^{s} \phi_{E_i}^{\alpha}(x^{E_i}) = \prod_{i=1}^{s} \chi_{E_i}(x^{E_i})$$

Therefore, P_α belongs to $\mathcal{R}_{\mathcal{E},A}$ and consequently $\mathcal{R}_{\mathcal{E},A}$ is log-affine. □

If we do not want to exclude probability distributions containing zeroes then for the description of some limit properties of iterative procedures we need extend the definition of $\mathcal{R}_{\mathcal{E},A}$ to its closure.

Definition 1.45 (Closure of $\mathcal{R}_{\mathcal{E},A}$) $\bar{\mathcal{R}}_{\mathcal{E},A}$ is the closure of $\mathcal{R}_{\mathcal{E},A}$ i.e.

$$\bar{\mathcal{R}}_{\mathcal{E},A} = \{P_A \in \mathcal{P}_A, \exists \{Q_{(i)}\}_{i=1}^{\infty} : \ Q_{(i)} \in \mathcal{R}_{\mathcal{E},A} \text{ for } i = 1, 2, \ldots \text{ and } \lim_{i \to \infty} Q_{(i)} = P_A\}$$

It will be used in the sequel that both $\bar{\mathcal{R}}_{\mathcal{E},A}$ and $\mathcal{S}_{\mathcal{E},A}$ are *compact* sets.

Relations between $\mathcal{S}_{\mathcal{E},A}$ and $\bar{\mathcal{R}}_{\mathcal{E},A}$ will be studied deeply in next chapters. However, the following theorem does not need any further assumption, therefore it will be given here. The theorem together with its proof, given by Lauritzen, is cited from [29, Lemma 3.14]. Since we are going to use similar technique to that used in the proof, the theorem is given together with its proof.

Theorem 1.5 (Intersection of $\mathcal{S}_{\mathcal{E},A}$ and $\bar{\mathcal{R}}_{\mathcal{E},A}$)
If there exists a probability distribution $P \in \mathcal{S}_{\mathcal{E},A} \cap \bar{\mathcal{R}}_{\mathcal{E},A}$ then it is unique.

Proof. Consider two distributions $P, Q \in \bar{\mathcal{R}}_{\mathcal{E},A}$. We need prove that if for all $E_i \in \mathcal{E}$: $P^{E_i} = Q^{E_i}$ then $P = Q$.

$$\begin{aligned} P &= \lim_{n\to\infty} \prod_{i=1}^{s} \psi_{E_i,(n)} \\ Q &= \lim_{n\to\infty} \prod_{i=1}^{s} \phi_{E_i,(n)} \end{aligned}$$

If we define $0 \cdot \log 0 = 0$ then we can write

$$\begin{aligned} \sum_{x^A \in \mathbb{X}^A} P(x^A) \log Q(x^A) &= \sum_{x^A \in \mathbb{X}^A} P(x^A) \log \left(\lim_{n\to\infty} \prod_{i=1}^{s} \phi_{E_i,(n)}(x^{E_i}) \right) \\ &= \lim_{n\to\infty} \sum_{x^A \in \mathbb{X}^A} P(x^A) \sum_{i=1}^{s} \log \phi_{E_i,(n)}(x^{E_i}) \\ &= \lim_{n\to\infty} \sum_{i=1}^{s} \sum_{x^{E_i} \in \mathbb{X}^{E_i}} P^{E_i}(x^{E_i}) \log \phi_{E_i,(n)}(x^{E_i}) \\ &= \lim_{n\to\infty} \sum_{i=1}^{s} \sum_{x^{E_i} \in \mathbb{X}^{E_i}} Q^{E_i}(x^{E_i}) \log \phi_{E_i,(n)}(x^{E_i}) \\ &= \lim_{n\to\infty} \sum_{x^A \in \mathbb{X}^A} Q(x^A) \sum_{i=1}^{s} \log \phi_{E_i,(n)}(x^{E_i}) \\ &= \sum_{x^A \in \mathbb{X}^A} Q(x^A) \log \left(\lim_{n\to\infty} \prod_{i=1}^{s} \phi_{E_i,(n)}(x^{E_i}) \right) \\ &= \sum_{x^A \in \mathbb{X}^A} Q(x^A) \log Q(x^A) \end{aligned}$$

If $P(x^A)$ and $Q(x^A)$ are exchanged then similarly to the previous it can be shown that

$$\sum_{x^A \in \mathbb{X}^A} Q(x^A) \log P(x^A) = \sum_{x^A \in \mathbb{X}^A} P(x^A) \log P(x^A)$$

In the next we will use a version of the so-called *information inequality* :

$$\sum_{x^A \in \mathbb{X}^A} Q(x^A) \log P(x^A) \leq \sum_{x^A \in \mathbb{X}^A} Q(x^A) \log Q(x^A),$$

which holds with equality only if $P = Q$.
Thus we can write:

$$\begin{aligned} \sum_{x^A \in \mathbb{X}^A} P(x^A) \log P(x^A) &= \sum_{x^A \in \mathbb{X}^A} Q(x^A) \log P(x^A) \leq \sum_{x^A \in \mathbb{X}^A} Q(x^A) \log Q(x^A), \\ \sum_{x^A \in \mathbb{X}^A} Q(x^A) \log Q(x^A) &= \sum_{x^A \in \mathbb{X}^A} P(x^A) \log Q(x^A) \leq \sum_{x^A \in \mathbb{X}^A} P(x^A) \log P(x^A). \end{aligned}$$

The inequalities holds with equality if and only if $P = Q$. Thus the proof is completed. □

2 I-projections

The objective of this chapter is to create a unifying framework for operations with probability distributions, that will be used for description of *iterative procedures* in the subsequent chapters. Most of theoretical results of this Chapter are not original, since deep study of I-projections was already done by I. Cziszár [10, 11] and recently extended by F. Matúš [32, 33]. However, it seems to be useful to rewrite some theorems and proofs so that they fit well to the situation we study. Moreover, if these theorems are given together one can easily see their mutual relations. The proof of Theorem 2.9 is new.

Probability distributions from a given set minimizing *I-divergence* with respect to a given distribution will be called *I-projections*. Depending on the minimized argument of *I-divergence* projections are called *I_1-projections* or *I_2-projections*. They can be generally defined for any set of probability distributions. However, we are going to be interested in *additively* and *multiplicatively constrained sets* only. In order to simplify proofs it is supposed that $A = \cup_{E_i \in \mathcal{E}} E_i = V$, in this section. It means $\mathcal{S}_{\mathcal{E},A} = \mathcal{S}_\mathcal{E}$ and $\mathcal{R}_{\mathcal{E},A} = \mathcal{R}_\mathcal{E}$. For the most part of this section the generalization is straightforward. At the end of this section the way of the generalization of results for $\mathcal{R}_{\mathcal{E},A}$ will be shown (see Lemmata 2.5 and 2.6).

2.1 I_1-projections on $\mathcal{S}_\mathcal{E}$

In this section we shall deal with I_1-projection on *additively constrained sets*.

Definition 2.1 (I_1-projection on $\mathcal{S}_\mathcal{E}$) Let probability distribution $Q \in \mathcal{P}$ and *additively constrained set* $\mathcal{S}_\mathcal{E}$ contain at least one probability distribution P, $P \ll Q$. Then for probability distribution

$$\pi_{\mathcal{S}_\mathcal{E}} Q = \arg \min_{\{P \in \mathcal{S}_\mathcal{E}\}} I(P \parallel Q)$$

it holds that $I(\pi_{\mathcal{S}_\mathcal{E}} Q \parallel Q)$ is finite and $\pi_{\mathcal{S}_\mathcal{E}} Q$ is called *I_1-projection* of Q on $\mathcal{S}_\mathcal{E}$.

Every *additively constrained set* is *convex* and *compact* (see Section 1.4). Therefore for a fixed Q the function $I(P \parallel Q) : \mathcal{S}_\mathcal{E} \mapsto \mathbb{R}^+$ (nonnegative real numbers including infinity) is strictly convex and continuous. It is supposed that $\mathcal{S}_\mathcal{E}$ contains at least one probability distribution P, $P \ll Q$. Consequently, probability distribution $\pi_{\mathcal{S}_\mathcal{E}} Q$ is unique.

Next, we present important theorem that is a special case of the well known *Minimum discrimination information theorem* given by Kullback and Khairat [27]. Theorem 2.1 gives a proposal, how to find I_1-projection on *additively constrained set*

$$\mathcal{S}_{\{E\}} = \{P \in \mathcal{P}, P^E = S_E \}, E \subseteq V.$$

Theorem 2.1 (I_1-projection on $\mathcal{S}_{\{E\}}$)
Let E be a subset of index set V and $S_E \in \mathcal{P}_E$ be a given probability distribution. Further, let $\mathcal{S}_{\{E\}}$ be nonempty subset of $\mathcal{P}$ defined by $\mathcal{S}_{\{E\}} = \{P \in \mathcal{P}, P^E = S_E \}$ and $Q \in \mathcal{P}$ be a given probability distribution such that $S_E \ll Q^E$. Finally, let probability distribution $R \in \mathcal{S}_{\{E\}}$ be defined by

$$R(x) \quad = \quad \begin{cases} 0, & \text{for } Q^E(x^E) = 0, \\ Q(x)\dfrac{S_E(x^E)}{Q^E(x^E)}, & \text{for } Q^E(x^E) > 0. \end{cases} \tag{2.1}$$

Then

(i) *$R = \pi_{\mathcal{S}_{\{E\}}} Q$ i.e. R is I_1-projection of Q on set $\mathcal{S}_{\{E\}}$ and*

(ii) *for every $P \in \mathcal{S}_{\{E\}}$ the following Three Points Property holds:*

$$I(P \parallel Q) \quad = \quad I(P \parallel R) + I(R \parallel Q). \tag{2.2}$$

Proof.
As $I(R \parallel Q)$ is always nonnegative, the assertion **(i)** follows immediately from **(ii)**. Therefore we shall proof only **(ii)**. The proof splits into two parts according to whether $P \ll Q$ or not.

(A) $[P \in \mathcal{S}_{\{E\}}, P \not\ll Q]$
In this case $I(P \parallel Q) = \infty$. It follows from equation 2.1 that $R \ll Q$, therefore $I(R \parallel Q) < \infty$. $P \not\ll Q$ and $R \ll Q$ implies that $P \not\ll R$ and, consequently, $I(P \parallel R) = \infty$, which proves **(ii)** for this particular case.

(B) $[P \in \mathcal{S}_{\{E\}}, P \ll Q]$
Since R is defined so that $R \ll Q$, $I(R \parallel Q)$ is always finite and defined by the expression

$$I(R \parallel Q) \quad = \quad \sum_{x \in \mathbb{X},\, R(x)>0} R(x) \log \frac{R(x)}{Q(x)}.$$

In the case of $x \in \mathbb{X} : R(x) > 0$, $R(x) = Q(x)\dfrac{S_E(x^E)}{Q^E(x^E)}$ and the I-divergence $I(R \parallel Q)$ can be rewritten:

$$\begin{aligned} I(R \parallel Q) \quad &= \quad \sum_{x \in \mathbb{X},\, Q(x)>0,\, S_E(x^E)>0} Q(x)\frac{S_E(x^E)}{Q^E(x^E)} \log \frac{S_E(x^E)}{Q^E(x^E)} \\ &= \quad \sum_{x^E \in \mathbb{X}^E,\, S_E(x^E)>0} \quad \sum_{x^{V\setminus E} \in \mathbb{X}^{V\setminus E},\, Q(x)>0} Q(x^E, x^{V\setminus E})\frac{S_E(x^E)}{Q^E(x^E)} \log \frac{S_E(x^E)}{Q^E(x^E)} \end{aligned}$$

$$
\begin{aligned}
&= \sum_{x^E \in \mathbb{X}^E,\ S_E(x^E)>0} \frac{S_E(x^E)}{Q^E(x^E)} \log \frac{S_E(x^E)}{Q^E(x^E)} \sum_{x^{V\setminus E} \in \mathbb{X}^{V\setminus E},\ Q(x)>0} Q(x^E, x^{V\setminus E}) \\
&= \sum_{x^E \in \mathbb{X}^E,\ S_E(x^E)>0} S_E(x^E) \frac{Q^E(x^E)}{Q^E(x^E)} \log \frac{S_E(x^E)}{Q^E(x^E)} \\
&= \sum_{x^E \in \mathbb{X}^E,\ S_E(x^E)>0} S_E(x^E) \log \frac{S_E(x^E)}{Q^E(x^E)}.
\end{aligned}
$$

$P \in \mathcal{S}_{\{E\}}$ therefore $P^E = S_E$ and

$$
\begin{aligned}
I(R \parallel Q) &= \sum_{x^E \in \mathbb{X}^E,\ P^E(x^E)>0} P^E(x^E) \log \frac{S_E(x^E)}{Q^E(x^E)} \\
&= \sum_{x \in \mathbb{X},\ P(x)>0} P(x) \log \frac{S_E(x^E)}{Q^E(x^E)}
\end{aligned}
$$

$P \ll Q \Rightarrow P^E \ll Q^E \Rightarrow S_E \ll Q^E$ therefore the formula for $I(R \parallel Q)$ can be further rewritten

$$
\begin{aligned}
I(R \parallel Q) &= \sum_{x \in \mathbb{X},\ P(x)>0} P(x) \log \frac{Q(x)\frac{S_E(x^E)}{Q^E(x^E)}}{Q(x)} \\
&= \sum_{x \in \mathbb{X},\ P(x)>0} P(x) \log \frac{R(x)}{Q(x)} \qquad (2.3)
\end{aligned}
$$

For every $x \in \mathbb{X}$ it holds that $R(x) = 0$ if and only if $Q(x) = 0$ or $S_E(x^E) = P^E(x^E) = 0$. Since $P \ll Q$, $Q(x) = 0$ implies $P(x) = 0$. Obviously, $P^E(x^E) = 0$ implies $P(x) = 0$. Thus for every $P \in \mathcal{S}_{\{E\}}$ it holds that $P \ll R$ and we can write

$$
I(P \parallel Q) = \sum_{x \in \mathbb{X},\ P(x)>0} P(x) \log \frac{P(x) \cdot R(x)}{R(x) \cdot Q(x)} = I(P \parallel R) + \sum_{x \in \mathbb{X},\ P(x)>0} P(x) \log \frac{R(x)}{Q(x)}
$$

Using equation 2.3 we immediately get equation 2.2

$$
I(P \parallel Q) = I(P \parallel R) + I(R \parallel Q).
$$

That finishes the proof. □

Example 2.1 (I_1-projection on an additively constrained set $\mathcal{S}_{\{E\}}$)
Let $V = \{1, 2, 3\}$, $E = \{2, 3\}$, $Q(x)$ be a three dimensional distribution, and $S_E(x^E)$ be a two dimensional distribution, both defined for dichotomic random variables. The distributions are given in Table 2.1. Let $\mathcal{S}_{\{E\}} = \{P \in \mathcal{P}, P^E(x^E) = S_E(x^E)\}$. Since $S_E \ll Q^E$ resulting I_1-projection $\pi_{\mathcal{S}_{\{E\}}} Q(x)$ can be calculated using Theorem 2.1. I_1-projection of Q on $\mathcal{S}_{\{E\}}$ is given in Table 2.1 as well.

Lemma 2.1 *Let $\mathcal{S}_{\{E\}}$ be a nonempty subset of $\mathcal{P}$ defined by $\mathcal{S}_{\{E\}} = \{P \in \mathcal{P}, P^E = S_E\}$, E be a subset of index set V, and $S_E \in \mathcal{P}_E$ be a given probability distribution. I_1-projection of $Q \in \mathcal{P}$ on $\mathcal{S}_{\{E\}}$ exists if and only if $S_E \ll Q^E$.*

Distribution $Q(x)$

	$X_1=0$ $X_2=0$	$X_1=0$ $X_2=1$	$X_1=1$ $X_2=0$	$X_1=1$ $X_2=1$
$X_3=0$	0	1/6	1/6	0
$X_3=1$	1/6	1/6	2/6	0

Distribution $Q^E(x^E)$

	$X_1=0$ $X_2=0$	$X_1=0$ $X_2=1$	$X_1=1$ $X_2=0$	$X_1=1$ $X_2=1$
	1/6	2/6	3/6	0

Distribution $S_E(x^E)$

	$X_1=0$ $X_2=0$	$X_1=0$ $X_2=1$	$X_1=1$ $X_2=0$	$X_1=1$ $X_2=1$
	1/3	0	2/3	0

Distribution $R = \pi_{\mathcal{S}_{\{E\}}} Q(x)$

	$X_1=0$ $X_2=0$	$X_1=0$ $X_2=1$	$X_1=1$ $X_2=0$	$X_1=1$ $X_2=1$
$X_3=0$	0	0	2/9	0
$X_3=1$	1/3	0	4/9	0

Table 2.1: Distributions from Example 2.1

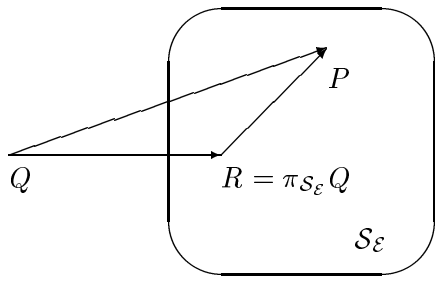
Proof. I_1-projection of $Q \in \mathcal{P}$ on $\mathcal{S}_{\{E\}} = \{P \in \mathcal{P}, P^E = S_E\}$ exists iff $\mathcal{S}_{\{E\}}$ contains at least one probability distribution P, $P \ll Q$. Then $P \ll Q \Rightarrow P^E \ll Q^E \Rightarrow S_E \ll Q^E$.
The proof of the converse is a consequence of Theorem 2.1. If there exist a probability distribution $S \in \mathcal{S}_{\{E\}}$, $S^E \ll Q^E$ then formula 2.1 of this theorem defines I_1-projection of $Q \in \mathcal{P}$ on $\mathcal{S}_{\{E\}}$. □

The following theorem is cited from [10, Theorem 2.2]. The Three Points Property holds for any additively constrained set $\mathcal{S}_\mathcal{E}$ containing at least one probability distribution $P \in \mathcal{S}_\mathcal{E}, P \ll Q$.

Theorem 2.2 (Three Points Property)
$R = \pi_{\mathcal{S}_\mathcal{E}} Q$ iff for every $P \in \mathcal{S}_\mathcal{E}$ the Three Points Property holds:

$$I(P \parallel Q) \quad = \quad I(P \parallel R) + I(R \parallel Q).$$

Figure 2.1 displays the situation described by Theorem 2.2. I-divergence can play a role of squared Euclidean distance. Then the Three Points Property is an analogy of Pythagoras' theorem. It should be noted that not all conceptions from Euclidean geometry work in I-divergence geometry. Figure 2.2 displays the situation described by Theorem 2.3.



$$I(P \parallel Q) = I(P \parallel \pi_{\mathcal{S}_\mathcal{E}} Q) + I(\pi_{\mathcal{S}_\mathcal{E}} Q \parallel Q)$$

Figure 2.1: Three Points Property

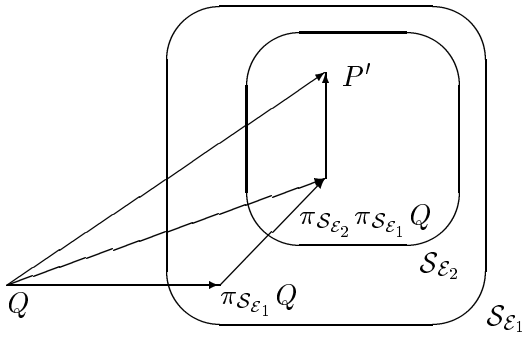


Figure 2.2: I_1-projection on a subset

Theorem 2.3 (I_1-projection on a subset) Csiszár [10, Theorem 2.3]
Let $\mathcal{S}_{\mathcal{E}_1} \subseteq \mathcal{P}$ and $\mathcal{S}_{\mathcal{E}_2} \subseteq \mathcal{S}_{\mathcal{E}_1}$ be two additively constrained sets, further let $Q \in \mathcal{P}$ be a given probability distribution such that there exist a probability distribution $P \in \mathcal{S}_{\mathcal{E}_2}, P \ll Q$ then there exist probability distributions $\pi_{\mathcal{S}_{\mathcal{E}_1}} Q$, $\pi_{\mathcal{S}_{\mathcal{E}_2}} Q$ and the following equation holds:

$$\pi_{\mathcal{S}_{\mathcal{E}_2}} \pi_{\mathcal{S}_{\mathcal{E}_1}} Q \quad = \quad \pi_{\mathcal{S}_{\mathcal{E}_2}} Q \tag{2.4}$$

Proof.
According to Theorem 2.2 taking $\mathcal{S}_{\mathcal{E}_1}$ for $\mathcal{S}_\mathcal{E}$:

$$\forall P' \in \mathcal{S}_{\mathcal{E}_1} : \quad I(P' \parallel Q) \quad = \quad I(P' \parallel \pi_{\mathcal{S}_{\mathcal{E}_1}} Q) + I(\pi_{\mathcal{S}_{\mathcal{E}_1}} Q \parallel Q),$$

and thus

$$\forall P' \in \mathcal{S}_{\mathcal{E}_1} : I(\pi_{\mathcal{S}_{\mathcal{E}_1}} Q \parallel Q) \quad = \quad I(P' \parallel Q) - I(P' \parallel \pi_{\mathcal{S}_{\mathcal{E}_1}} Q). \tag{2.5}$$

Since $\pi_{\mathcal{S}_{\mathcal{E}_2}} \pi_{\mathcal{S}_{\mathcal{E}_1}} Q \in \mathcal{S}_{\mathcal{E}_2} \subseteq \mathcal{S}_{\mathcal{E}_1}$ equation 2.5 holds for this particular $P' = \pi_{\mathcal{S}_{\mathcal{E}_2}} \pi_{\mathcal{S}_{\mathcal{E}_1}} Q$

$$I(\pi_{\mathcal{S}_{\mathcal{E}_2}} \pi_{\mathcal{S}_{\mathcal{E}_1}} Q \parallel \pi_{\mathcal{S}_{\mathcal{E}_1}} Q) \quad = \quad I(\pi_{\mathcal{S}_{\mathcal{E}_2}} \pi_{\mathcal{S}_{\mathcal{E}_1}} Q \parallel Q) - I(\pi_{\mathcal{S}_{\mathcal{E}_1}} Q \parallel Q) \tag{2.6}$$

Using Theorem 2.2 for $\mathcal{S}_{\mathcal{E}_1}$ and for $\mathcal{S}_{\mathcal{E}_2}$ it holds $\forall P' \in \mathcal{S}_{\mathcal{E}_2}$:

$$\begin{aligned} I(P' \parallel \pi_{\mathcal{S}_{\mathcal{E}_1}} Q) \quad &= \quad I(P' \parallel \pi_{\mathcal{S}_{\mathcal{E}_2}} \pi_{\mathcal{S}_{\mathcal{E}_1}} Q) + I(\pi_{\mathcal{S}_{\mathcal{E}_2}} \pi_{\mathcal{S}_{\mathcal{E}_1}} Q \parallel \pi_{\mathcal{S}_{\mathcal{E}_1}} Q) \\ I(\pi_{\mathcal{S}_{\mathcal{E}_2}} \pi_{\mathcal{S}_{\mathcal{E}_1}} Q \parallel \pi_{\mathcal{S}_{\mathcal{E}_1}} Q) \quad &= \quad I(P' \parallel \pi_{\mathcal{S}_{\mathcal{E}_1}} Q) - I(P' \parallel \pi_{\mathcal{S}_{\mathcal{E}_2}} \pi_{\mathcal{S}_{\mathcal{E}_1}} Q) \end{aligned} \tag{2.7}$$

Substituting $I(\pi_{\mathcal{S}_{\mathcal{E}_2}}\pi_{\mathcal{S}_{\mathcal{E}_1}}Q \parallel \pi_{\mathcal{S}_{\mathcal{E}_1}}Q)$ from equation 2.6 to equation 2.7 we get $\forall P' \in \mathcal{S}_{\mathcal{E}_2}$:

$$I(\pi_{\mathcal{S}_{\mathcal{E}_2}}\pi_{\mathcal{S}_{\mathcal{E}_1}}Q \parallel Q) - I(\pi_{\mathcal{S}_{\mathcal{E}_1}}Q \parallel Q) = I(P' \parallel \pi_{\mathcal{S}_{\mathcal{E}_1}}Q) - I(P' \parallel \pi_{\mathcal{S}_{\mathcal{E}_2}}\pi_{\mathcal{S}_{\mathcal{E}_1}}Q) \quad (2.8)$$

After the substitution of $I(\pi_{\mathcal{S}_{\mathcal{E}_1}}Q \parallel Q)$ from equation 2.5 to equation 2.8 we get for all $P' \in \mathcal{S}_{\mathcal{E}_2}$:

$$\begin{aligned} I(\pi_{\mathcal{S}_{\mathcal{E}_2}}\pi_{\mathcal{S}_{\mathcal{E}_1}}Q \parallel Q) - I(P' \parallel Q) + I(P' \parallel \pi_{\mathcal{S}_{\mathcal{E}_1}}Q) &= I(P' \parallel \pi_{\mathcal{S}_{\mathcal{E}_1}}Q) - I(P' \parallel \pi_{\mathcal{S}_{\mathcal{E}_2}}\pi_{\mathcal{S}_{\mathcal{E}_1}}Q) \\ I(P' \parallel Q) &= I(P' \parallel \pi_{\mathcal{S}_{\mathcal{E}_2}}\pi_{\mathcal{S}_{\mathcal{E}_1}}Q) + I(\pi_{\mathcal{S}_{\mathcal{E}_2}}\pi_{\mathcal{S}_{\mathcal{E}_1}}Q \parallel Q) \end{aligned}$$

which, according to Theorem 2.2, induces 2.4. □

Following theorem provides a lower bound of I-divergence stated using total variance. The proof of the theorem was given by Kullback [26]. Since lower bound of I-divergence plays important role in proofs of convergence of iterative procedures we will paraphrase his proof in more details.

Theorem 2.4 (Lower bound on I-divergence by total variance) Kullback [26]
If $P, Q \in \mathcal{P}$ are two probability distributions then the following inequality holds:

$$|P - Q| \leq 2\sqrt{I(P \parallel Q)} \quad (2.9)$$

Proof.
$|P - Q|$ is always finite therefore if $P \not\ll Q$ then $I(P \parallel Q) = +\infty$ and inequality 2.9 holds. The proof for $P \ll Q$ uses the Jensen and Cauchy-Schwarz inequalities.

$$\begin{aligned} I(P \parallel Q) &= \sum_{x\in\mathbb{X},\, P(x)>0} P(x)\log\frac{P(x)}{Q(x)} \\ &= \sum_{x\in\mathbb{X},\, P(x)>0} 2\sqrt{P(x)}\sqrt{P(x)}\log\frac{\sqrt{P(x)}}{\sqrt{Q(x)}} \\ &= 2(\sum_{y\in\mathbb{X}}\sqrt{P(y)}) \sum_{x\in\mathbb{X},\, P(x)>0} \frac{\sqrt{P(x)}}{\sum\limits_{y\in\mathbb{X}}\sqrt{P(y)}}\sqrt{P(x)}\log\frac{\sqrt{P(x)}}{\sqrt{Q(x)}} \end{aligned}$$

Jensen inequality will be used in the following form

$$\sum_x \alpha_x f(a(x), b(x)) \geq f(\sum_x \alpha_x a(x), \sum_x \alpha_x b(x)),$$

where $\sum\limits_x \alpha_x = 1$ and $f(a(x), b(x))$ is a convex function.

In this case $f(a(x), b(x)) = a_x(\log a(x) - \log b(x))$, $\alpha_x = \dfrac{\sqrt{P(x)}}{\sum\limits_{y\in\mathbb{X}}\sqrt{P(y)}}$, $a(x) = \sqrt{P(x)}$, and $b(x) = \sqrt{Q(x)}$. Then the following holds:

$$\begin{aligned} I(P \parallel Q) &\geq 2(\sum_{y\in\mathbb{X}}\sqrt{P(y)})(\sum_{x\in\mathbb{X},\, P(x)>0} \frac{\sqrt{P(x)}}{\sum\limits_{y\in\mathbb{X}}\sqrt{P(y)}}\sqrt{P(x)}) \cdot \log\frac{\dfrac{\sqrt{P(x)}}{\sum\limits_{y\in\mathbb{X}}\sqrt{P(y)}}\sqrt{P(x)}}{\dfrac{\sqrt{P(x)}}{\sum\limits_{y\in\mathbb{X}}\sqrt{P(y)}}\sqrt{Q(x)}} \\ &\geq -2\log \sum_{x\in\mathbb{X},\, P(x)>0} \sqrt{P(x)}\sqrt{Q(x)} \end{aligned}$$

Since $-\log x \geq 1 - x$ the inequality can be extended to:

$$\begin{aligned} I(P \parallel Q) &\geq 2(1 - \sum_{x\in\mathbb{X},\, P(x)>0} \sqrt{P(x)}\sqrt{Q(x)}) \\ &\geq \sum_{x\in\mathbb{X},\, P(x)>0} P(x) + \sum_{x\in\mathbb{X},\, P(x)>0} Q(x) - 2 \cdot \sum_{x\in\mathbb{X},\, P(x)>0} \sqrt{P(x)}\sqrt{Q(x)} \\ &\geq \sum_{x\in\mathbb{X},\, P(x)>0} (\sqrt{P(x)} - \sqrt{Q(x)})^2 \end{aligned} \tag{2.10}$$

The total variance can be rewritten as follows:

$$\begin{aligned} |P - Q|^2 &= \left(\sum_{x\in\mathbb{X}} |P(x) - Q(x)|\right)^2 \\ &= \left(\sum_{x\in\mathbb{X}} \left|\sqrt{P(x)} - \sqrt{Q(x)}\right| \cdot \left(\sqrt{P(x)} + \sqrt{Q(x)}\right)\right)^2 \end{aligned}$$

The Cauchy-Schwarz inequality will be used in the form:

$$(\sum_x u(x)v(x))^2 \leq \sum_x u(x)^2 \cdot \sum_x v(x)^2$$

Substituting $u(x) = |\sqrt{P(x)} - \sqrt{Q(x)}|$ and $v(x) = (\sqrt{P(x)} + \sqrt{Q(x)})$ we can write:

$$\begin{aligned} |P - Q|^2 &\leq \sum_{x\in\mathbb{X}} \left|\sqrt{P(x)} - \sqrt{Q(x)}\right|^2 \cdot \sum_{x\in\mathbb{X}} \left(\sqrt{P(x)} + \sqrt{Q(x)}\right)^2 \\ &\leq \sum_{x\in\mathbb{X}} \left|\sqrt{P(x)} - \sqrt{Q(x)}\right|^2 \cdot \sum_{x\in\mathbb{X}} \left(P(x) + 2\sqrt{P(x)}\sqrt{Q(x)} + Q(x)\right) \\ &\leq \sum_{x\in\mathbb{X}} \left|\sqrt{P(x)} - \sqrt{Q(x)}\right|^2 \cdot \left(1 + 2\sum_{x\in\mathbb{X}} \sqrt{P(x)}\sqrt{Q(x)} + 1\right) \end{aligned}$$

The Cauchy-Schwarz inequality for the second part of the previous expression is used. This time it is in the form:

$$\sum_x u(x)v(x)) \leq \sqrt{\sum_x u(x)^2} \cdot \sqrt{\sum_x v(x)^2}$$

Substituting $u(x) = \sqrt{P(x)}$ and $v(x) = \sqrt{Q(x)}$ we can write:

$$\begin{aligned} |P - Q|^2 &\leq \sum_{x\in\mathbb{X}} \left|\sqrt{P(x)} - \sqrt{Q(x)}\right|^2 \cdot \left(2 + 2\sqrt{\sum_{x\in\mathbb{X}} P(x)} \cdot \sqrt{\sum_{x\in\mathbb{X}} Q(x)}\right) \\ &\leq 4\sum_{x\in\mathbb{X}} \left(\sqrt{P(x)} - \sqrt{Q(x)}\right)^2 \end{aligned} \tag{2.11}$$

It follows from inequalities 2.10 and 2.11 that

$$I(P \parallel Q) \geq \sum_{x\in\mathbb{X},\, P(x)>0} \left(\sqrt{P(x)} - \sqrt{Q(x)}\right)^2 \geq \frac{1}{4}|P - Q|^2$$

which immediately gives inequality 2.9 we wanted to prove. □

The following theorem given by Csiszár [10, Theorem 3.2] has a fundamental significance. Its proof is based on *the Three Points Property* (Theorem 2.2), *I_1-projection on a subset* (Theorem 2.3), and on the connection between I-divergence and total variance (Theorem 2.4).

Theorem 2.5 (Convergence of recursive I_1-projections)
Let $\mathcal{S}_{\mathcal{E}_1}, \ldots, \mathcal{S}_{\mathcal{E}_s} \subseteq \mathcal{P}$ be arbitrary additively constrained sets, such that $\mathcal{S}_{\mathcal{E}} = \cap_{i=1}^{s} \mathcal{S}_{\mathcal{E}_s} \neq \emptyset$ and $Q_{(0)} \in \mathcal{P}$ be a probability distribution to which there exists $P \in \mathcal{S}_{\mathcal{E}}$ with $P \ll Q_{(0)}$. Define $Q_{(1)}, Q_{(2)}, \ldots$ recursively by letting

$$Q_{(n)} = \pi_{\mathcal{S}_{\mathcal{E}_j}} Q_{(n-1)}, \quad n = 1, 2, \ldots, \quad j = (n-1) \bmod s + 1.$$

Then the sequence of probability distributions $Q_{(n)}$ converges (point-wise or, equivalently, in variation) to the probability distribution

$$Q^* = \pi_{\mathcal{S}_{\mathcal{E}}} Q_{(0)}.$$

Proof. Take an arbitrary $P \in \mathcal{S}_{\mathcal{E}_1} \cap \ldots \cup \mathcal{S}_{\mathcal{E}_s}$. For $n = 1, 2, \ldots$, $Q_{(n)}$ is an I_1-projection therefore we can apply the Three points property (Theorem 2.2) according to which we can write:

$$\begin{aligned} I(P \parallel Q_{(0)}) &= I(P \parallel Q_{(1)}) + I(Q_{(1)} \parallel Q_{(0)}), \\ I(P \parallel Q_{(1)}) &= I(P \parallel Q_{(2)}) + I(Q_{(2)} \parallel Q_{(1)}), \\ &\ldots \end{aligned}$$

Therefore using limit transition we can write

$$\begin{aligned} I(P \parallel Q_{(0)}) &= I(P \parallel Q_{(n)}) + I(Q_{(1)} \parallel Q_{(0)}) + \ldots + I(Q_{(n)} \parallel Q_{(n-1)}). \\ I(P \parallel Q_{(0)}) &= \lim_{n\to\infty} I(P \parallel Q_{(n)}) + \sum_{n=1}^{\infty} I(Q_{(n)} \parallel Q_{(n-1)}) \qquad (2.12) \\ I(P \parallel Q_{(0)}) &\geq \sum_{n=1}^{\infty} I(Q_{(n)} \parallel Q_{(n-1)}). \end{aligned}$$

All $I(Q_{(n)} \parallel Q_{(n-1)})$ are nonnegative and since $P \ll Q_{(0)}$, $I(P \parallel Q_{(0)})$ is finite. Therefore

$$I(Q_{(n)} \parallel Q_{(n-1)}) \quad \to \quad 0. \qquad (2.13)$$

Since all probability distributions are defined for finite random variables, using Theorem 2.4, we get the convergence of the sequence $Q_{(0)}, Q_{(1)}, \ldots$. Let the limit probability distribution be denoted Q^*.

If we consider an arbitrary index m, $1 \leq m \leq s$ then all distributions in sequence

$$Q_{(m)}, Q_{(m+s)}, Q_{(m+2s)}, Q_{(m+3s)}, \ldots$$

are *I_1-projections* of some distribution on $\mathcal{S}_{\mathcal{E}_m}$. $\mathcal{S}_{\mathcal{E}_m}$ is *compact set* therefore the limit of the sequence belongs to $\mathcal{S}_{\mathcal{E}_m}$ as well. All subsequences of convergent sequence has the same limit, hence the sequence converges to Q^* and since m was arbitrary, we get:

$$Q^* \in \mathcal{S}_{\mathcal{E}} = \cap_{i=1}^{s} \mathcal{S}_{\mathcal{E}_i}.$$

Since P in equation 2.12 was arbitrary this equation must holds also for Q^*. Using equation 2.13 we can write:

$$\begin{aligned} I(Q^* \parallel Q_{(0)}) &= \lim_{n\to\infty} I(Q^* \parallel Q_{(n)}) + \sum_{n=1}^{\infty} I(Q_{(n)} \parallel Q_{(n-1)}) \\ &= \sum_{n=1}^{\infty} I(Q_{(n)} \parallel Q_{(n-1)}) \end{aligned} \tag{2.14}$$

Using equations 2.12 and 2.14 we get for all $P \in \mathcal{S}_\mathcal{E}$:

$$\begin{aligned} I(P \parallel Q_{(0)}) &= \lim_{n\to\infty} I(P \parallel Q_{(n)}) + \sum_{n=1}^{\infty} I(Q_{(n)} \parallel Q_{(n-1)}) \\ &= I(P \parallel Q^*) + \sum_{n=1}^{\infty} I(Q_{(n)} \parallel Q_{(n-1)}) \\ &= I(P \parallel Q^*) + I(Q_{(0)} \parallel Q^*). \end{aligned}$$

which according to the Three points property (Theorem 2.2) means that

$$Q^* = \pi_{\mathcal{S}_\mathcal{E}} Q_{(0)}.$$

□

2.2 I_2-projections

Csiszár [10] developed geometry of I-divergence using I_1-projection. Many results on I-divergence for the reverse order of arguments were achieved by Čencov [7]. The properties of I_1- and I_2-*projections* are in many aspects different. The differences begin with the fact that minimizer of $\arg\min_{\{Q\in\mathcal{T}\}} I(P \parallel Q)$ is not generally unique even if $\mathcal{T}$ is convex and compact set (see Section 2.2.1).

2.2.1 I_2-projections on $\mathcal{S}_\mathcal{E}$

In spite of the fact that results given in this subsection will not be used in next chapters, we are going to state them here, since they can help to build up better understanding of *I-divergence geometry.*

In the case of I_2-projections on $\mathcal{S}_\mathcal{E}$ the uniqueness could be guaranteed if we had restricted ourselves to strictly positive probability distributions, but this is not the way we want to follow. Instead of it we will use selections scheme from set of minimizers proposed by Matúš [32]. It is based on *maximum entropy principle*, so that it can be shown that I_2-*projection* on $\mathcal{S}_\mathcal{E}$ defined such a way can be computed using formula similar to one of I_1-projection.

Definition 2.2 (Set of I_2-minimizers on $\mathcal{S}_\mathcal{E}$) Set of I_2-*minimizers* with respect to a probability distribution $Q \in \mathcal{P}$ and to *additively constrained set* $\mathcal{S}_\mathcal{E}$ of probability distributions containing at least one probability distribution P, $Q \ll P$, is the set $\mathcal{T}_{\mathcal{S}_\mathcal{E}}(Q) = \arg\min_{\{P\in\mathcal{S}_\mathcal{E}\}} I(Q \parallel P)$.

The following lemma, which is cited from Matúš [32], approve that in the case of strictly positive probability distributions I_2-projections on $\mathcal{S}_\mathcal{E}$ is unique.

Lemma 2.2 *All probability distributions $T \in \mathcal{T}_{\mathcal{S}_\mathcal{E}}(Q)$ coincide on the set $\{x \in \mathbb{X}, Q(x) > 0\}$.*

In the subsequent text attention is given to properties of I_2-projection on *additively constrained sets* $\mathcal{S}_\mathcal{E}$ defined in the following way.

Definition 2.3 (I_2-projection on $\mathcal{S}_\mathcal{E}$) *I_2-projection* of a probability distribution $Q \in \mathcal{P}$ on an *additively constrained set* $\mathcal{S}_\mathcal{E}$ containing at least one probability distribution P, $Q \ll P$ is the probability distribution

$$\pi'_{\mathcal{S}_\mathcal{E}} Q = \pi_{\mathcal{T}_{\mathcal{S}_\mathcal{E}}(Q)} U = \arg \min_{\{T \in \mathcal{T}_{\mathcal{S}_\mathcal{E}}(Q)\}} I(T \parallel U),$$

where U denotes the uniform probability distribution.

The following lemma states an explicit formula for computing *set of I_2-minimizers* on *additively constrained set* $\mathcal{S}_{\{E\}}$. The formula will be used in Theorem 2.6, where it is restricted so that it leads to unique I_2-projection on $\mathcal{S}_{\{E\}}$.

Lemma 2.3 (Set of I_2-minimizers on $\mathcal{S}_{\{E\}}$)
Let $E \subseteq V$, and $Q \in \mathcal{P}$, $S_E \in \mathcal{P}_E$ be given probability distributions such that $Q^E \ll S_E$. Further let $\mathcal{S}_{\{E\}} = \{P \in \mathcal{P}, P^E(x^E) = S_E(x^E)\}$. Then

$$\mathcal{T}_{\mathcal{S}_{\{E\}}}(Q) \quad = \quad \{P^* \in \mathcal{S}_{\{E\}} : \forall x, Q^E(x^E) > 0 : P^*(x) = Q(x)\frac{S_E(x^E)}{Q^E(x^E)}\} \tag{2.15}$$

Proof. It suffices to prove that

$$\forall P \in \mathcal{S}_{\{E\}}, \forall P^* \in \mathcal{T}_{\mathcal{S}_{\{E\}}} : I(Q \parallel P^*) \quad \leq \quad I(Q \parallel P).$$

By the definition

$$I(Q \parallel P) \quad = \quad \sum_{x \in \mathbb{X},\, Q(x)>0} Q(x) \log \frac{Q(x)}{P(x)}$$

Since $Q^E(x^E) = 0$ implies $Q(x) = 0$ it can be derived from the right side of expression 2.15 that for all x such that $Q(x) > 0$, $P^*(x) > 0$ as well, and

$$\begin{aligned} I(Q \parallel P) \quad &= \quad \sum_{x \in \mathbb{X},\, Q(x)>0} Q(x) \log \frac{Q(x)P^*(x)}{P(x)P^*(x)} \\ &= \quad I(Q \parallel P^*) + \sum_{x \in \mathbb{X},\, Q(x)>0} Q(x) \log \frac{P^*(x)}{P(x)} \end{aligned}$$

It remains to prove that

$$\sum_{x \in \mathbb{X},\, Q(x)>0} Q(x) \log \frac{P^*(x)}{P(x)} \quad \geq \quad 0$$

In order to simplify notation define the set $B = V \setminus E$.

$$\begin{aligned}
&\sum_{x\in\mathbb{X},\ Q(x)>0} Q(x)\log\frac{P^*(x)}{P(x)} = \\
&= \sum_{x\in\mathbb{X},\ Q(x)>0} Q(x)\log\frac{Q(x)S_E(x^E)}{P(x)Q^E(x^E)} \\
&= \sum_{x\in\mathbb{X},\ Q(x)>0} Q^E(x^E)Q^{B|E}(x^B|x^E)\log\frac{Q^E(x^E)Q^{B|E}(x^B|x^E)S_E(x^E)}{P^E(x^E)P^{B|E}(x^B|x^E)Q^E(x^E)}
\end{aligned}$$

Since $P \in \mathcal{S}_{\{E\}}$ therefore $P^E = S_E$, and

$$\begin{aligned}
&\sum_{x\in\mathbb{X},\ Q(x)>0} Q(x)\log\frac{P^*(x)}{P(x)} = \\
&= \sum_{x^E\in\mathbb{X}^E,\ Q(x)>0} Q^E(x^E) \sum_{x^B\in\mathbb{X}^B,\ Q(x)>0} Q^{B|E}(x^B|x^E)\log\frac{Q^{B|E}(x^B|x^E)}{P^{B|E}(x^B|x^E)} \\
&= \sum_{x^E\in\mathbb{X}^E,\ Q(x)>0} Q^E(x^E) I(Q^{B|E} \parallel P^{B|E}) \\
&\geq 0
\end{aligned}$$

That finishes the proof. □

Remark. It follows from Lemma 2.3 that the minimization of $I(Q \parallel P)$ over P does not give any restriction for values of $P(x)$ for such x that $Q^E(x^E) = 0$.

Theorem 2.6 (I_2-projection on $\mathcal{S}_{\{E\}}$)
Let $E \subseteq V$, and $Q \in \mathcal{P}$, $S_E \in \mathcal{P}_E$ be given probability distributions such that $Q^E \ll S_E$. Further let $\mathcal{S}_{\{E\}} = \{P \in \mathcal{P}, P^E(x^E) = S_E(x^E)\}$, and $U(x)$ is the uniform probability distribution. Then

$$R(x) = \begin{cases} U(x)\dfrac{S_E(x^E)}{U^E(x^E)}, & \text{if } Q^E(x^E) = 0 \\ Q(x)\dfrac{S_E(x^E)}{Q^E(x^E)}, & \text{if } Q^E(x^E) > 0, \end{cases}$$

is I_2-projection of Q on $\mathcal{S}_{\{E\}}$, i.e. $R(x) = \pi'_{\mathcal{S}_{\{E\}}}Q(x)$.

Proof. It suffices to prove that

$$\forall P^* \in \mathcal{T}_{\mathcal{S}_{\{E\}}}(Q) : I(R \parallel U) \leq I(P^* \parallel U) \tag{2.16}$$

with equality only if $P^*(x) = R(x)$. According to Lemma 2.3

$$\begin{aligned}
\mathcal{T}_{\mathcal{S}_{\{E\}}}(Q) &= \arg\min_{\{P\in\mathcal{S}_{\{E\}},\ Q\ll P\}} I(Q \parallel P) \\
&= \{P^* \in \mathcal{S}_{\{E\}} : \forall x,\ Q^E(x^E) > 0 : P^*(x) = Q(x)\frac{S_E(x^E)}{Q^E(x^E)}\}
\end{aligned}$$

$$
\begin{aligned}
I(R \parallel U) &= \sum_{x\in\mathbb{X},\ R(x)>0} R(x)\log\frac{R(x)}{U(x)} \\
&= \sum_{x\in\mathbb{X},\ Q(x)>0,\ Q^E(x^E)>0} Q(x)\frac{S_E(x^E)}{Q^E(x^E)}\log\frac{R(x)}{U(x)} \\
&\quad + \sum_{x\in\mathbb{X},\ S_E(x^E)>0,\ Q^E(x^E)=0} U(x)\frac{S_E(x^E)}{U^E(x^E)}\log\frac{S_E(x^E)}{U^E(x^E)} \\
&= \sum_{x\in\mathbb{X},\ P^*(x)>0,\ Q^E(x^E)>0} P^*(x)\log\frac{R(x)}{U(x)} \\
&\quad + \sum_{x^E\in\mathbb{X}^E,\ S_E(x^E)>0,\ Q^E(x^E)=0} S_E(x^E)\log\frac{S_E(x^E)}{U^E(x^E)}
\end{aligned}
$$

Since $P^* \in \mathcal{T}_{\mathcal{S}_{\{E\}}}(Q)$, $P^* \in \mathcal{S}_{\{E\}}$, and consequently $P^{*E} = S_E$ we can write

$$
\begin{aligned}
I(R \parallel U) &= \sum_{x\in\mathbb{X},\ P^*(x)>0,\ Q^E(x^E)>0} P^*(x)\log\frac{R(x)}{U(x)} + \\
&\quad \sum_{x^E\in\mathbb{X}^E,\ P^{*E}(x^E)>0,\ Q^E(x^E)=0} P^{*E}(x^E)\log\frac{S_E(x^E)}{U_E(x^E)} \\
&= \sum_{x\in\mathbb{X},\ P^*(x)>0,\ Q^E(x^E)>0} P^*(x)\log\frac{R(x)}{U(x)} + \\
&\quad \sum_{x\in\mathbb{X},\ P^*(x)>0,\ Q^E(x^E)=0} P^*(x)\log\frac{U(x)\frac{S_E(x^E)}{U^E(x^E)}}{U(x)} \\
&= \sum_{x\in\mathbb{X},\ P^*(x)>0,\ Q^E(x^E)>0} P^*(x)\log\frac{R(x)}{U(x)} + \\
&\quad \sum_{x\in\mathbb{X},\ P^*(x)>0,\ Q^E(x^E)=0} P^*(x)\log\frac{R(x)}{U(x)} \\
&= \sum_{x\in\mathbb{X},\ P^*(x)>0} P^*(x)\log\frac{R(x)}{U(x)} \qquad (2.17)
\end{aligned}
$$

Since $P^* \ll U$, and $R \ll U$ the substitutions $f(x) = \frac{P^*(x)}{U(x)}$, and $f^*(x) = \frac{R(x)}{U(x)}$ to the definition of I-divergence $I(P^* \parallel S)$, and to equation 2.17 can be used.

$$
I(P^* \parallel U) = \sum_{x\in\mathbb{X},\ P^*(x)>0} U(x)f(x)\log f(x) \qquad (2.18)
$$

$$
I(R \parallel U) = \sum_{x\in\mathbb{X}, P^*(x)>0} U(x)f(x)\log f^*(x) \qquad (2.19)
$$

As a direct consequence of equations 2.18, and 2.19 the proof of inequality 2.16 is equivalent to the proof of the statement that for all $f(x)$ such that $P \in \mathcal{S}_{\{E\}}$ the next inequality is valid

with equality only if $f(x) = f^*(x)$.

$$\sum_{x\in\mathbb{X},\, f(x)>0} f(x)Q(x)\log\frac{f^*(x)}{f(x)} \quad \leq \quad 0,$$

which has already been proved in the proof of Theorem 2.1. □

Following example displays calculation of I_2-projection on a linearly constrained set $\mathcal{S}_{\{E\}}$.

Example 2.2 (I_2-projection) Let $V = \{1,2,3\}$, $E = \{2,3\}$, and $Q(x)$, $S(x)$ be three dimensional and two dimensional distributions, respectively, defined for dichotomic random variables. They are given in Table 2.2. Let $\mathcal{S}_{\{E\}} = \{P \in \mathcal{P}, P^E(x^E) = S_E(x^E)\}$. It is obvious

Distribution $Q(x)$

	$X_1=0$ $X_2=0$	$X_1=0$ $X_2=1$	$X_1=1$ $X_2=0$	$X_1=1$ $X_2=1$
$X_3=0$	0	1/6	2/6	0
$X_3=1$	0	0	3/6	0

Distribution $Q^E(x^E)$

	$X_1=0$ $X_2=0$	$X_1=0$ $X_2=1$	$X_1=1$ $X_2=0$	$X_1=1$ $X_2=1$
	0	1/6	5/6	0

Distribution $S_E(x^E)$

	$X_1=0$ $X_2=0$	$X_1=0$ $X_2=1$	$X_1=1$ $X_2=0$	$X_1=1$ $X_2=1$
	2/6	2/6	2/6	0

Distribution $R = \pi'_{\mathcal{S}_{\{E\}}}Q(x)$

	$X_1=0$ $X_2=0$	$X_1=0$ $X_2=1$	$X_1=1$ $X_2=0$	$X_1=1$ $X_2=1$
$X_3=0$	1/6	2/6	2/15	0
$X_3=1$	1/6	0	3/15	0

Table 2.2: Distributions from Example 2.2

that $Q^E(x^E) \ll S_E(x^E)$. Therefore, Theorem 2.6 can be used. The resulting I_2-projection $\pi'_{\mathcal{S}_{\{E\}}}Q(x)$ is given in the Table 2.2.

There is no analogy of the *Three Points Property* for *I_2-projections* on $\mathcal{S}_\mathcal{E}$. Instead of it an inequality called *Four Points Property* holds. This property was introduced by Csiszár and Tusnády in [11]. First, we prove its particular case for *additively constrained set* $\mathcal{S}_{\{E\}}$.

Theorem 2.7 (Four Points Property on $\mathcal{S}_{\{E\}}$)
Let Q_1, Q_2 be probability distributions from $\mathcal{P}$ such that $Q_2 \ll Q_1$, $E \subseteq V$. Further, let $S_E \in \mathcal{P}_E$ be given probability distribution such that $Q_1^E(x^E) \ll S_E(x^E)$. Finally, let $\mathcal{S}_{\{E\}} = \{P \in \mathcal{P}, P^E(x^E) = S_E(x^E)\}$ then

$$I(Q_2 \parallel \pi'_{\mathcal{S}_{\{E\}}} Q_1) \leq I(Q_2 \parallel Q_1) + I(Q_2 \parallel \pi'_{\mathcal{S}_{\{E\}}} Q_2).$$

Proof.

$$\begin{aligned} I(Q_2 \parallel Q_1) + I(Q_2 \parallel \pi'_{\mathcal{S}_{\{E\}}} Q_2) &= \sum_{x \in \mathbb{X},\, Q_2(x)>0} Q_2(x) \log \frac{Q_2(x) Q_2(x) \pi'_{\mathcal{S}_{\{E\}}} Q_1(x)}{Q_1(x) \pi'_{\mathcal{S}_{\{E\}}} Q_2(x) \pi'_{\mathcal{S}_{\{E\}}} Q_1(x)} \\ &= \sum_{x \in \mathbb{X},\, Q_2(x)>0} Q_2(x) \log \frac{Q_2(x)}{\pi'_{\mathcal{S}_{\{E\}}} Q_1(x)} \\ &\quad + \sum_{x \in \mathbb{X},\, Q_2(x)>0} Q_2(x) \log \frac{Q_2(x) \pi'_{\mathcal{S}_{\{E\}}} Q_1(x)}{Q_1(x) \pi'_{\mathcal{S}_{\{E\}}} Q_2(x)} \end{aligned} \tag{2.20}$$

In order to substitute for I_2-projections $\pi'_{\mathcal{S}_{\{E\}}} Q_1(x)$ and $\pi'_{\mathcal{S}_{\{E\}}} Q_2(x)$ Theorem 2.6 will be used. Fortunately, not all value combinations of $Q_1^E(x^E)$, and $Q_2^E(x^E)$ are possible. Since the sum in equation 2.20 is only over positive values of $Q_2(x)$ it follows from the assumptions that $Q_2(x) > 0$ implies $Q_1(x) > 0$, and therefore $Q_1^E(x^E) > 0$. That is why it suffices to sum just over all $x \in \mathbb{X}$, $Q_1^E(x^E) > 0$, $Q_2^E(x^E) > 0$. In this case

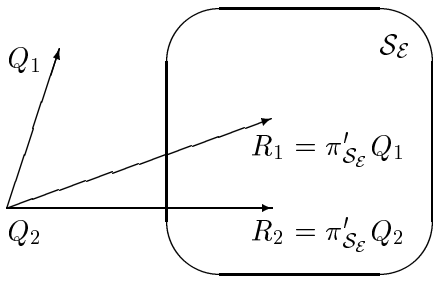
$$\begin{aligned} \pi'_{\mathcal{S}_{\{E\}}} Q_1(x) &= Q_1(x) \log \frac{S_E(x^E)}{Q_1^E(x^E)}, \\ \pi'_{\mathcal{S}_{\{E\}}} Q_2(x) &= Q_2(x) \log \frac{S_E(x^E)}{Q_2^E(x^E)}. \end{aligned}$$

Substituting the previous into equation 2.20

$$\begin{aligned} &I(Q_2 \parallel Q_1) + I(Q_2 \parallel \pi'_{\mathcal{S}_{\{E\}}} Q_2) = \\ &= I(Q_2 \parallel \pi'_{\mathcal{S}_{\{E\}}} Q_1) + \sum_{x \in \mathbb{X},\, Q_2(x)>0} Q_2(x) \log \frac{Q_2(x) Q_1(x) \frac{S_E(x^E)}{Q_1^E(x^E)}}{Q_1(x) Q_2(x) \frac{S_E(x^E)}{Q_2^E(x^E)}} \\ &= I(Q_2 \parallel \pi'_{\mathcal{S}_{\{E\}}} Q_1) + \sum_{x \in \mathbb{X},\, Q_2(x)>0} Q_2(x) \log \frac{Q_2^E(x^E)}{Q_1^E(x^E)} \\ &= I(Q_2 \parallel \pi'_{\mathcal{S}_{\{E\}}} Q_1) + I(Q_2^E \parallel Q_1^E) \\ &\geq I(Q_2 \parallel \pi'_{\mathcal{S}_{\{E\}}} Q_1), \end{aligned}$$

which concludes the proof. □

Theorem 2.8 cited from Csiszár and Tusnády [11, Lemma 3] states that the Four Points Property holds for any additively constrained set $\mathcal{S}_{\mathcal{E}}$. Figure 2.3 displays the situation described by this theorem.



$$I(Q_2 \parallel \pi'_{\mathcal{S}_\mathcal{E}} Q_1) \leq I(Q_2 \parallel Q_1) + I(Q_2 \parallel \pi'_{\mathcal{S}_\mathcal{E}} Q_2)$$

Figure 2.3: Four Points Property

Theorem 2.8 (Four Points Property)
Let $Q_1 \in \mathcal{P}$ be a probability distribution such that there exists at least one $P \in \mathcal{S}_\mathcal{E}, P \ll Q_1$. Then for every $Q_2 \in \mathcal{P}$ the Four Points Property *holds, i.e.*

$$I(Q_2 \parallel \pi'_{\mathcal{S}_\mathcal{E}} Q_1) \leq I(Q_2 \parallel Q_1) + I(Q_2 \parallel \pi'_{\mathcal{S}_\mathcal{E}} Q_2).$$

2.2.2 I_2-projections on $\bar{\mathcal{R}}_\mathcal{E}$

In this subsection we will exploit the uniqueness of intersection

$$\bar{\mathcal{R}}_\mathcal{E} \cap \mathcal{S}_\mathcal{E} = \bar{\mathcal{R}}_\mathcal{E} \cap \{P \in \mathcal{P}, P^{E_1} = S^{E_1}, \ldots, P^{E_s} = S^{E_s}\},$$

which is the statement of Theorem 1.5. Let us denote this unique distribution from the intersection as $\hat{S}$. The fact that distribution $\hat{S} \in \bar{\mathcal{R}}_\mathcal{E} \cap \mathcal{S}_\mathcal{E}$ means that $\hat{S}$ has its marginals equal to given $S^E, E \in \mathcal{E}$ and at the same time it factorizes with respect to $\mathcal{E}$ or is a limit of factorizing distributions. In the sequel it will be shown that such a distribution minimizes the *I-divergence* $I(\cdot \parallel Q)$ within the class $\bar{\mathcal{R}}_\mathcal{E}$ i.e. for every $Q \in \bar{\mathcal{R}}_\mathcal{E}$ it holds that $I(S \parallel Q) \geq I(S \parallel \hat{S})$, which is the assertion of Theorem 2.9. Therefore I_2-projections of S on *closure of multiplicatively constrained set* can be defined (Definition 2.4) to be a distribution from intersection

$$\bar{\mathcal{R}}_\mathcal{E} \cap \mathcal{S}_\mathcal{E} = \bar{\mathcal{R}}_\mathcal{E} \cap \{P \in \mathcal{P}, P^{E_1} = S^{E_1}, \ldots, P^{E_s} = S^{E_s}\}.$$

In order to give a simple proof of the following lemma we need Theorem 3.4 and Theorem 3.2. Since both exploits the definition of *iterative proportional fitting procedure* (IPFP) they are not given until Chapter 3, which is devoted to IPFP.

Lemma 2.4
Let $\mathcal{S}_\mathcal{E}$ be an additively constrained set. For $\hat{S} \in \bar{\mathcal{R}}_\mathcal{E} \cap \mathcal{S}_\mathcal{E}$ and for any $S \in \mathcal{S}_\mathcal{E}$ it holds that $S \ll \hat{S}$.

Proof. It follows from Theorem 1.5 that $\hat{S} \in \bar{\mathcal{R}}_\mathcal{E} \cap \mathcal{S}_\mathcal{E}$ is unique. According to Theorem 3.4 $\hat{S}$ is the limit distribution of IPFP, as well. Using Theorem 3.2 applied for IPFP starting with

the uniform distribution U we can write for limit distribution of IPFP $\hat{S} = \pi_{\mathcal{S}_\mathcal{E}} U$. Therefore the Three Points Property (Theorem 2.2) can be applied, which gives

$$I(S \parallel U) \quad = \quad I(S \parallel \hat{S}) + I(\hat{S} \parallel U).$$

Both $I(S \parallel U)$ and $I(\hat{S} \parallel U)$ are finite therefore it follows that $I(S \parallel \hat{S})$ is finite as well. That assertion is equivalent to $S \ll \hat{S}$. □

Theorem 2.9 (Intersection $\bar{\mathcal{R}}_\mathcal{E} \cap \mathcal{S}_\mathcal{E}$) *Let $\mathcal{S}_\mathcal{E}$ be an additively constrained set. For $\hat{S} \in \bar{\mathcal{R}}_\mathcal{E} \cap \mathcal{S}_\mathcal{E}$, $Q \in \bar{\mathcal{R}}_\mathcal{E}$, and $S \in \mathcal{S}_\mathcal{E}$ it holds that*

$$I(S \parallel Q) \quad \geq \quad I(S \parallel \hat{S}).$$

Proof. It is obvious that for Q such that $S \not\ll Q$ the assertion of Theorem 2.9 holds. We will show for $S \ll Q$ that

$$I(S \parallel Q) - I(S \parallel \hat{S}) \geq 0.$$

Using the fact that both $Q, \hat{S} \in \bar{\mathcal{R}}_\mathcal{E}$ we can write

$$\begin{aligned} \hat{S} &= \lim_{n\to\infty} \prod_{i=1}^{s} \psi_{E_i,(n)} \\ Q &= \lim_{n\to\infty} \prod_{i=1}^{s} \phi_{E_i,(n)} \end{aligned}$$

It follows from Lemma 2.4 that $S \ll \hat{S}$. Therefore $I(S \parallel \hat{S})$ is finite and it holds that

$$\begin{aligned} I(S \parallel Q) - I(S \parallel \hat{S}) &= \sum_{x\in\mathbb{X},\, S(x)>0} S(x) \log \lim_{n\to\infty} \prod_{i=1}^{s} \psi_{E_i,(n)}(x^{E_i}) \\ &\quad - \sum_{x\in\mathbb{X},\, S(x)>0} S(x) \log \lim_{n\to\infty} \prod_{i=1}^{s} \phi_{E_i,(n)}(x^{E_i}) \\ &= \sum_{x\in\mathbb{X},\, S(x)>0} S(x) \lim_{n\to\infty} \sum_{i=1}^{s} \log \frac{\psi_{E_i,(n)}(x^{E_i})}{\phi_{E_i,(n)}(x^{E_i})} \end{aligned}$$

Since $x = (x^{E_i}, x^{V\setminus E_i})$ we can write

$$I(S \parallel Q) - I(S \parallel \hat{S}) \quad = \quad \sum_{i=1}^{s} \lim_{n\to\infty} \sum_{x^{E_i}\in\mathbb{X}^{E_i},\, S^{E_i}(x)>0} \log \frac{\psi_{E_i,(n)}(x^{E_i})}{\phi_{E_i,(n)}(x^{E_i})} \sum_{x^{V\setminus E_i}\in\mathbb{X}^{V\setminus E_i},\, S(x)>0} S(x)$$

Knowing $\hat{S} \in \mathcal{S}_\mathcal{E}$ it is obvious that $\hat{S}^{E_i} = S^{E_i} = \sum_{x^{V\setminus E_i}\in\mathbb{X}^{V\setminus E_i},\, S(x)>0} S(x)$

$$I(S \parallel Q) - I(S \parallel \hat{S}) \quad = \quad \sum_{i=1}^{s} \lim_{n\to\infty} \sum_{x^{E_i}\in\mathbb{X}^{E_i},\, \hat{S}^{E_i}(x)>0} \hat{S}^{E_i} \cdot \log \frac{\psi_{E_i,(n)}(x^{E_i})}{\phi_{E_i,(n)}(x^{E_i})}$$

$$
\begin{aligned}
&= \sum_{x\in\mathbb{X},\ \hat{S}(x)>0} \hat{S}(x)\cdot \lim_{n\to\infty} \log \frac{\prod_{i=1}^{s} \psi_{E_i,(n)}(x^{E_i})}{\prod_{i=1}^{s} \phi_{E_i,(n)}(x^{E_i})} \\
&= \sum_{x\in\mathbb{X},\ \hat{S}(x)>0} \hat{S}(x)\cdot \log \frac{\hat{S}(x)}{Q(x)} \\
&= I(\hat{S} \parallel Q) \geq 0,
\end{aligned}
$$

with equality iff $x \in \mathbb{X},\ \hat{S}(x) > 0 : Q(x) = \hat{S}(x)$. □

Due to Theorems 1.5 and 2.9 the following definition of I_2-projection on $\bar{\mathcal{R}}_{\mathcal{E}}$ define a unique distribution.

Definition 2.4 (I_2-projection on $\bar{\mathcal{R}}_{\mathcal{E}}$)
Let $S \in \mathcal{P}$ be a probability distribution and $\{P \in \mathcal{P}, P^{E_1} = S^{E_1}, \ldots, P^{E_s} = S^{E_s}\}$ be *additively constrained set* defined by marginals of S. *I_2-projection* of a probability distribution S on a *closure of multiplicatively constrained set* $\bar{\mathcal{R}}_{\mathcal{E}}$ is probability distribution

$$\pi'_{\bar{\mathcal{R}}_{\mathcal{E}}} S \quad = \quad \bar{\mathcal{R}}_{\mathcal{E}} \cap \mathcal{S}_{\mathcal{E}}.$$

The fact that I_2-projections of S on *closure of multiplicatively constrained set* belongs to $\mathcal{S}_{\mathcal{E}}$, which immediately follows from the definition, enable us to prove certain convergence properties of IPFP on *decomposable generating classes*. These proofs are part and parcel of Chapter 4.

At the beginning of this section we have simplified situation by assumption of $A = \cup_{E_i\in\mathcal{E}} E_i = V$. The next two lemmata present a way *I_2-projection* on *multiplicatively constrained sets* can be extended to the case $V \setminus A = D \neq \emptyset$. These lemmata are analogy of similar lemmata for *maximum likelihood estimation in contingency tables* given by S. Haberman in [17, Lemma 5.6 and 5.7].

Remark. In the subsequent text, notation will be simplified so that for any $A \subseteq V$ we will write just $\sum_{x^A} \ldots$ instead of $\sum_{x^A\in\mathbb{X}^A} \ldots$.

Lemma 2.5 *If $A = \bigcup_{E_i\in\mathcal{E}} E_i$, $D = V \setminus A$, $U_D \in \mathcal{P}_D$ is the uniform probability distribution, $P \in \mathcal{R}_{\mathcal{E}}$, and P^A is marginal of P then for all $x \in \mathbb{X}^V$*

$$P(x) = P^A(x^A)\cdot U_D(x^D) \quad \text{and} \quad P^A \in \mathcal{R}_{\mathcal{E}}.$$

Conversely, if $P^A \in \mathcal{R}_{\mathcal{E}}$ and for all $x \in \mathbb{X}^V$ $P(x) = P^A(x^A)\cdot U_D(x^D)$ then

$$P \in \mathcal{R}_{\mathcal{E}}.$$

Proof. Since $V \setminus A = D$ and $P(x)$ can be written as $\prod_{E_i\in\mathcal{E}} \psi_i(x^{E_i})$ it follows that

$$P^A(x^A) \quad = \quad \sum_{x^{V\setminus A}} \prod_{E_i\in\mathcal{E}} \psi_i(x^{E_i}) = \prod_{E_i\in\mathcal{E}} \psi_i(x^{E_i}) \sum_{x^D} 1 = \frac{P(x)}{U_D(x^D)}.$$

The proof of the second part is elementary as well. Since values of the uniform probability distribution are constant we can write

$$P(x) = P^A(x^A) \cdot U_D(x^D) = \prod_{E_i \in \mathcal{E}} \psi_i(x^{E_i}) \cdot U_D(x^D).$$

$$\begin{array}{lll} \text{If we define} & i = 1: & \phi_i(x^{E_i}) = \psi_i(x^{E_i}) \cdot U_D(x^D) \\ & i \geq 1: & \phi_i(x^{E_i}) = \psi_i(x^{E_i}) \\ \text{then} & P(x) = \prod_{E_i \in \mathcal{E}} \phi_i(x^{E_i}) & \Rightarrow \quad P \in \mathcal{R}_{\mathcal{E}}. \end{array}$$

□

Lemma 2.6 *If $A = \bigcup_{E_i \in \mathcal{E}} E_i$, $D = V \setminus A$, $U_D \in \mathcal{P}_D$ is the uniform probability distribution, $P \in \mathcal{P}$ is a given probability distribution, and $\pi'_{\mathcal{R}_{\mathcal{E},V}} P(x^V)$ exists then*

$$\pi'_{\mathcal{R}_{\mathcal{E},V}} P(x^V) = \pi'_{\mathcal{R}_{\mathcal{E},A}} P(x^A) \cdot U_D(x^D).$$

$\pi'_{\mathcal{R}_{\mathcal{E},V}} P(x^V)$ exists iff $\pi'_{\mathcal{R}_{\mathcal{E},A}} P(x^A)$ exists.

Proof. Class $\mathcal{S}_{\mathcal{E}}$ is defined as $\{S \in \mathcal{P} : S^E = P^E, E \in \mathcal{E}\}$. It is direct consequence of Definition 2.4 that

$$\pi'_{\mathcal{R}_{\mathcal{E},V}} P \in \mathcal{R}_{\mathcal{E}} \quad \text{and} \quad (\pi'_{\mathcal{R}_{\mathcal{E},V}} P)^{E_i} = P^{E_i}, \ E_i \in \mathcal{E}$$

Since also

$$\pi'_{\mathcal{R}_{\mathcal{E},A}} P \in \mathcal{R}_{\mathcal{E}} \quad \text{and} \quad (\pi'_{\mathcal{R}_{\mathcal{E},A}} P)^{E_i} = P^{E_i}, \ E_i \in \mathcal{E}$$

it is obvious that $(\pi'_{\mathcal{R}_{\mathcal{E},V}} P)^A = \pi'_{\mathcal{R}_{\mathcal{E},A}} P$. Therefore application of Lemma 2.5 gives:

$$(\pi'_{\mathcal{R}_{\mathcal{E},V}} P)^A = \frac{\pi'_{\mathcal{R}_{\mathcal{E},V}} P}{U_D} \ \Rightarrow \ \pi'_{\mathcal{R}_{\mathcal{E},V}} P = \pi'_{\mathcal{R}_{\mathcal{E},A}} P \cdot U_D.$$

To prove converse implication we show that if $\pi'_{\mathcal{R}_{\mathcal{E},A}} P$ exists then by Lemma 2.5 a distribution

$$\bar{P} = \pi'_{\mathcal{R}_{\mathcal{E},A}} P \cdot U_D \in \mathcal{R}_{\mathcal{E}} \text{ and } \bar{P}^A = \pi'_{\mathcal{R}_{\mathcal{E},A}} P.$$

It follows from Definition 2.4 that $(\pi'_{\mathcal{R}_{\mathcal{E},A}} P)^{E_i} = \bar{P}^{E_i}$, therefore again using Definition 2.4 for $\pi'_{\mathcal{R}_{\mathcal{E},V}} P$ we get

$$\bar{P} = \pi'_{\mathcal{R}_{\mathcal{E},V}} P.$$

□

If distributions from *multiplicatively constrained sets* satisfy a decomposability condition, then the following Lemma 2.7, which performs a way of their decomposition, can be applied. Lemma 2.8 shows that *I_2-projections* to *multiplicatively constrained sets* $\mathcal{R}_{\mathcal{E}}$ can be decomposed using *I_2-projections* to subsets of $\mathcal{E}$ in a similar manner. We have exploited parallelism between *maximum likelihood estimation of log-linear models* and *I_2-projecting to multiplicatively constrained sets* again. It enables us to rewrite lemmata of S. Haberman [17, Lemma 5.8, Lemma 5.9] so that they fit well to our problem.

Lemma 2.7 (Composition and decomposition of distributions from $\mathcal{R}_\mathcal{E}$)
Let $\mathcal{E}_1$ and $\mathcal{E}_2$ be a partition of $\mathcal{E}$, such that for some $E_1 \in \mathcal{E}_1$ and $E_2 \in \mathcal{E}_2$ it holds

$$A \cap B = C,$$

where $A = \cup_{E_i \in \mathcal{E}_1} E_i$, $B = \cup_{E_j \in \mathcal{E}_2} E_j$, and $C = E_1 \cap E_2$. Further let $D = V \setminus (A \cap B)$. If $P \in \mathcal{R}_\mathcal{E}$ then

$$P^A \in \mathcal{R}_{\mathcal{E}_1}, P^B \in \mathcal{R}_{\mathcal{E}_2} \quad \text{and} \quad \forall x \in \mathbb{X}^V : \; P(x) = \frac{P^A(x^A) \cdot P^B(x^B)}{P^C(x^C)} \cdot U_D(x^D).$$

Conversely, if $P_A(x^A) \in \mathcal{R}_{\mathcal{E}_1}, P_B(x^B) \in \mathcal{R}_{\mathcal{E}_2}$, such that

$$\begin{aligned} \forall x^C \in \mathbb{X}^C : \quad & P_A^C(x^C) = P_B^C(x^C) = P_C(x^C), \quad \textit{and} \\ \forall x \in \mathbb{X}^V : \quad & Q(x) = \frac{P_A(x^A) \cdot P_B(x^B)}{P_C(x^C)} \cdot U_D(x^D) \end{aligned}$$

then

$$Q \in \mathcal{R}_\mathcal{E}, Q^A = P_A, \quad \textit{and} \quad Q^B = P_B.$$

Proof. $P \in \mathcal{R}_\mathcal{E}$ therefore $P(x) = \prod_{E_i \in \mathcal{E}_1} \psi_i(x^{E_i}) \cdot \prod_{E_i \in \mathcal{E}_2} \psi_i(x^{E_i})$ and we can write

$$\begin{aligned} P^A(x^A) &= \sum_{x^{V \setminus A}} \prod_{E_i \in \mathcal{E}_1} \psi_i(x^{E_i}) \cdot \prod_{E_i \in \mathcal{E}_2} \psi_i(x^{E_i}) \\ &= \prod_{E_i \in \mathcal{E}_1} \psi_i(x^{E_i}) \cdot \sum_{x^{V \setminus A}} \prod_{E_i \in \mathcal{E}_2} \psi_i(x^{E_i}) \end{aligned}$$

Since $V \setminus A = V \setminus (\cup_{E_i \in \mathcal{E}_1} E_i) = D \cup (\cup_{E_j \in \mathcal{E}_2} E_j \setminus C) = D \cup (B \setminus C)$, thus the sum is only over all $x^{D \cup (B \setminus C)} \in \mathbb{X}^{D \cup (B \setminus C)}$:

$$\begin{aligned} P^A(x^A) &= \prod_{E_i \in \mathcal{E}_1} \psi_i(x^{E_i}) \cdot \sum_{x^{B \setminus C}} \prod_{E_i \in \mathcal{E}_2} \psi_i(x^{E_i}) \sum_{x^D} 1 \\ P^A(x^A) &= \frac{1}{U_D(x^D)} \cdot \prod_{E_i \in \mathcal{E}_1} \psi_i(x^{E_i}) \cdot \sum_{x^{B \setminus C}} \prod_{E_i \in \mathcal{E}_2} \psi_i(x^{E_i}) \end{aligned} \tag{2.21}$$

Similarly, we can show for P^B that

$$P^B(x^B) = \frac{1}{U_D(x^D)} \cdot \prod_{E_i \in \mathcal{E}_2} \psi_i(x^{E_i}) . \sum_{x^{A \setminus C}} \prod_{E_i \in \mathcal{E}_1} \psi_i(x^{E_i}) \tag{2.22}$$

$$(B \setminus C) \subset E_1 \quad \Rightarrow \quad P^A(x^A) \in \mathcal{R}_{\mathcal{E}_1}, \text{ and} \tag{2.23}$$

$$(A \setminus C) \subset E_2 \quad \Rightarrow \quad P^B(x^B) \in \mathcal{R}_{\mathcal{E}_2}. \tag{2.24}$$

$$\tag{2.25}$$

For P^C we can write:

$$\begin{aligned} P^C(x^C) &= \sum_{y^{A\setminus C}=x^{A\setminus C}} P^A(y^A) \\ &= \frac{1}{U_D(x^D)} \cdot \sum_{x^{A\setminus C}} [\prod_{E_i\in\mathcal{E}_1} \psi_i(x^{E_i}) \cdot \sum_{x^{B\setminus C}} \prod_{E_i\in\mathcal{E}_2} \psi_i(x^{E_i})] \\ &= \frac{1}{U_D(x^D)} \cdot [\sum_{x^{A\setminus C}} \prod_{E_i\in\mathcal{E}_1} \psi_i(x^{E_i})] \cdot [\sum_{x^{B\setminus C}} \prod_{E_i\in\mathcal{E}_2} \psi_i(x^{E_i})] \end{aligned} \tag{2.26}$$

Finally, using equations 2.21, 2.22, and 2.26 we can conclude first part of the proof:

$$\begin{aligned} &\frac{P^A(x^A)\cdot P^B(x^B)}{P^C(x^C)} \cdot \frac{1}{U_D(x^D)} = \\ &= \frac{[\prod_{E_i\in\mathcal{E}_1} \psi_i(x^{E_i}) \cdot \sum_{x^{B\setminus C}} \prod_{E_i\in\mathcal{E}_2} \psi_i(x^{E_i})] \cdot [\prod_{E_i\in\mathcal{E}_2} \psi_i(x^{E_i}) \cdot \sum_{x^{A\setminus C}} \prod_{E_i\in\mathcal{E}_1} \psi_i(x^{E_i})]}{[\sum_{x^{A\setminus C}} \prod_{E_i\in\mathcal{E}_1} \psi_i(x^{E_i})] \cdot [\sum_{x^{B\setminus C}} \prod_{E_i\in\mathcal{E}_2} \psi_i(x^{E_i})]} \\ &= \prod_{E_i\in\mathcal{E}_1} \psi_i(x^{E_i}) \cdot \prod_{E_i\in\mathcal{E}_2} \psi_i(x^{E_i}) \\ &= P(x). \end{aligned}$$

The proof of the converse:
Since it is supposed that $P_A \in \mathcal{R}_{\mathcal{E}_1}, P_B \in \mathcal{R}_{\mathcal{E}_2}$, and $P_C = P_B^C$ it applies

$$\begin{aligned} Q^A(x^A) &= \sum_{x^{V\setminus A}} Q(x) = \sum_{x^{V\setminus A}} [\frac{P_A(x^A)\cdot P_B(x^B)}{P_C(x^C)} \cdot U_D(x^D)] \\ &= \sum_{x^{V\setminus A}} [\prod_{E_i\in\mathcal{E}_1} \psi_i(x^{E_i}) \cdot \prod_{E_i\in\mathcal{E}_2} \psi_i(x^{E_i}) \cdot \frac{1}{\sum_{x^{B\setminus C}} \prod_{E_i\in\mathcal{E}_2} \psi_i(x^{E_i})} \cdot U_D(x^D) \\ &= \prod_{E_i\in\mathcal{E}_1} \psi_i(x^{E_i}) \cdot \frac{U_D(x^D)}{\sum_{x^{B\setminus C}} \prod_{E_i\in\mathcal{E}_2}(x^{E_i})} \cdot \sum_{x^{B\setminus C}} [\prod_{E_i\in\mathcal{E}_2} \psi_i(x^{E_i}) \cdot \sum_{x^D} 1] \\ &= P_A(x^A) \cdot \frac{U_D(x^D)}{\sum_{x^{B\setminus C}} \prod_{E_i\in\mathcal{E}_2}(x^{E_i})} \cdot \frac{\sum_{x^{B\setminus C}} \prod_{E_i\in\mathcal{E}_2} \psi_i(x^{E_i})}{U_D(x^D)} \\ &= P_A(x^A). \end{aligned}$$

Similarly $Q^B(x^B) = P_B(x^B)$.
It remains to prove that $Q \in \mathcal{R}_{\mathcal{E}}$. Since $P_A \in \mathcal{R}_{\mathcal{E}_1}$ and $P_B(x^B) \in \mathcal{R}_{\mathcal{E}_2}$ Lemma 2.5 can be applied so that

$$P_A \cdot U_D = Q^A \cdot U_D \in \mathcal{R}_{\mathcal{E}_1} \quad \text{and} \quad P_B \cdot U_D = Q^B \cdot U_D \in \mathcal{R}_{\mathcal{E}_2}.$$

Since

$$P_C(x^C) = \sum_{x^{A\setminus C}} P_A(x^A) = \sum_{x^{A\setminus C}} \prod_{E_i\in\mathcal{E}_1} \psi_i(x^{E_i}) = \phi(x^c)$$

it holds that

$$P_C \cdot U_D = Q^C \cdot U_D \in \mathcal{R}_{\{C\}}.$$

Finally, $\mathcal{R}_{\mathcal{E}_1} \subset \mathcal{R}_{\mathcal{E}}$, $\mathcal{R}_{\mathcal{E}_2} \subset \mathcal{R}_{\mathcal{E}}$, and $\mathcal{R}_{\{C\}} \subset \mathcal{R}_{\mathcal{E}} \Rightarrow Q(x) \in \mathcal{R}_{\mathcal{E}}$. □

Lemma 2.8 (Composition and decomposition of $\pi'_{\mathcal{R}_\mathcal{E}}P$) *Let $\mathcal{E}_1$ and $\mathcal{E}_2$ be a partition of $\mathcal{E}$, such that for some $E_1 \in \mathcal{E}_1$ and $E_2 \in \mathcal{E}_2$ it holds*

$$A \cap B = C,$$

where $A = \cup_{E_i \in \mathcal{E}_1} E_i$, $B = \cup_{E_j \in \mathcal{E}_2} E_j$, and $C = E_1 \cap E_2$. Further let $D = V \setminus (A \cap B)$. Then $\pi'_{\mathcal{R}_{\mathcal{E},V}}P$ exists iff both $\pi'_{\mathcal{R}_{\mathcal{E}_1,A}}P$ (abbreviated $\pi'_{\mathcal{R}_A}P$) and $\pi'_{\mathcal{R}_{\mathcal{E}_2,B}}P(x^B)$ (abbreviated $\pi'_{\mathcal{R}_B}P$) exist. If $\pi'_{\mathcal{R}_V}P$ exists, then for any $x^V \in \mathbb{X}^V$:

$$\pi'_{\mathcal{R}_V}P(x^V) = \frac{\pi'_{\mathcal{R}_A}P(x^A) \cdot \pi'_{\mathcal{R}_B}P(x^B)}{P^C_{E_1}(x^C)} \cdot U_D(x^D).$$

Proof. Class $\mathcal{S}_\mathcal{E}$ is defined as $\{S \in \mathcal{P} : S^E = P^E, E \in \mathcal{E}\}$. Therefore all distributions that belongs to $\mathcal{S}_\mathcal{E}$ have its marginals equal to $P^E, E \in \mathcal{E}$.
If $\pi'_{\mathcal{R}_V}P$ exists then it is a direct consequence of expressions (2.23) and (2.24) that

$$(\pi'_{\mathcal{R}_V}P)^A \in \mathcal{R}_{\mathcal{E}_1} \tag{2.27}$$
$$(\pi'_{\mathcal{R}_V}P)^B \in \mathcal{R}_{\mathcal{E}_2}. \tag{2.28}$$

Since $C \subset E_1$

$$(\pi'_{\mathcal{R}_V}P)^C = P^C_{E_1}.$$

For every $E \in \mathcal{E}_1$ according to Definition 2.4 $(\pi'_{\mathcal{R}_V}P)^E = P^E$. Thus together with (2.27) it is exactly the definition of $\pi'_{\mathcal{R}_A}P$ (Definition 2.4). Since previous consideration can be repeated also for $\pi'_{\mathcal{R}_B}P$ we can write

$$(\pi'_{\mathcal{R}_V}P)^A = \pi'_{\mathcal{R}_A}P$$
$$(\pi'_{\mathcal{R}_V}P)^B = \pi'_{\mathcal{R}_B}P$$

Since $\pi'_{\mathcal{R}_V}P \in \mathcal{R}_\mathcal{E}$ Lemma 2.7 can be applied:

$$\pi'_{\mathcal{R}_V}P = \frac{(\pi'_{\mathcal{R}_V}P)^A \cdot (\pi'_{\mathcal{R}_V}P)^B}{(\pi'_{\mathcal{R}_V}P)^C} \cdot U_D = \frac{\pi'_{\mathcal{R}_A}P \cdot \pi'_{\mathcal{R}_B}P}{(\pi'_{\mathcal{R}_{E_1}}P)^C} \cdot U_D.$$

On the other hand, suppose that both $\pi'_{\mathcal{R}_A}P$ and $\pi'_{\mathcal{R}_B}P$ exists, then

$$(\pi'_{\mathcal{R}_A}P)^C = (\pi'_{\mathcal{R}_B}P)^C = (\pi'_{\mathcal{R}_{E_1}}P)^C.$$

Let us define a distribution $\bar{P}$ so that

$$\bar{P}(x) = \frac{\pi'_{\mathcal{R}_A}P(x^A) \cdot \pi'_{\mathcal{R}_B}P(x^B)}{P^C_{E_1}(x^C)}.$$

Then, according to Lemma 2.7 (converse part)

$$\bar{P}^A = \pi'_{\mathcal{R}_A}P, \tag{2.29}$$
$$\bar{P}^B = \pi'_{\mathcal{R}_B}P, \tag{2.30}$$
$$\bar{P} \in \mathcal{R}_\mathcal{E}. \tag{2.31}$$

As a consequence of expressions (2.29), (2.30) and Definition 2.4 we can write

$$E \in \mathcal{E}_1 : \quad \bar{P}^E = (\pi'_{\mathcal{R}_A} P)^E = P^E, \tag{2.32}$$

$$E \in \mathcal{E}_2 : \quad \bar{P}^E = (\pi'_{\mathcal{R}_B} P)^E = P^E. \tag{2.33}$$

Therefore, it follows from (2.31), (2.32), (2.33), and from Definition 2.4 that $\bar{P} = \pi'_{\mathcal{R}_V} P$. $\square$

3 Iterative Proportional Fitting Procedure

In this chapter we are going to analyze the well-known *iterative proportional fitting procedure* (abbreviated IPFP). The first section outlines the history of its convergence proofs and problems it was applied to. IPFP can be used to find I-projection of initial distribution to an *additively constrained set*. As a consequence of Theorem 2.5 that claims that in the case of *consistent input set* it converges to I_1-projection of initial distribution to given *additively constrained set*. In the last section of this chapter results of our experiments with IPFP on *inconsistent input set* will be summarized by means of conjectures. Discussion of IPFP behavior on *inconsistent input set* is extended in Section 5.6 where the *limit procedures* and their *arithmetic and geometric* averages are compared using I_1 and I_2 aggregates (defined in Section 5.1). The study of IPFP behavior on *consistent input set* continues in Chapter 4 which is devoted to the procedure's convergence properties when it is applied to an consistent input generated by a *decomposable hypergraph*.

3.1 What is IPFP?

Iterative proportional fitting procedure dates back to 1940. It was proposed by W. E. Deming and F. F. Stephan [13] as a procedure for adjustment of frequencies in contingency tables. This approach is later referred to as *classical use* of IPFP [4]. We will use a simple example of Deming and Stephan to describe the type of problems for which they proposed to use IPFP.

Example 3.1 (Frequencies adjustment)
Suppose a sample from U.S. population answered two questions in a opinion poll. Total number n of U.S. citizens replied to the questions: "How old are you?" and "What is your education?". Respondents were classified to $r \times s$ groups according to their answers. It resulted to a contingency table $(n_{ij})_{i=1,j=1}^{r,s}$ where n_{ij} is number of citizens in a age group i

and education group j. It is well known how U.S. citizens are distributed according to the age. It is similar with their education. Let us denote $N_{i\cdot}$ and $N_{\cdot j}$ the total number of U.S. citizens in age group i and education group j respectively. Now the problem is: how to adjust the table $(n_{ij})_{1,1}^{r,s}$ so that it corresponds to the well known marginals $N_{i\cdot}$ and $N_{\cdot j}$, but stay close to the proportions of original one. W. E. Deming and F. F. Stephan [13] proposed IPFP which they conjecture to minimize weighted square distance

$$\sum_{i,j=1,1}^{r,s} \frac{(n_{ij} - n * P_{ij})^2}{n_{ij}},$$

where P_{ij} are adjusted proportions they searched for.

It was shown that IPFP does not minimize this function. However, several authors paid attention to IPFP, which is a natural method of iterative adjustment.

Several authors wrote proofs of IPFP convergence. In 1959, D. T. Brown [5] gave a proof based on the fact, that the likelihood of the iterated solution is a monotonic decreasing function. C. T. Ireland and S. Kullback's [21] proof contained an error (1968). In 1970, S. Fienberg [16] showed that IPFP does not minimize square distance and proved IPFP convergence for strictly positive distribution. S. J. Haberman [17] showed that IPFP is a special case of the cyclic ascent method of function maximization [46] and proved several properties for decomposable generating classes (1974). In 1975, I. Csiszár [10] proved IPFP convergence using I-divergence geometry without assumption of strict positivity. The convergence of IPFP for continuous case still remains an open problem. L. Rüschendorf [37] made an attempt and gave the convergence proof for continuous case valid under some regularity conditions.

IPFP was applied to several problems from different domains. IPFP can used to construct a probabilistic expert systems, where the knowledge is represented by a joint probability distribution [23, 20]. S. J. Haberman [17] used IPFP for maximum likelihood estimation in contingency tables. Several other applications, listed in L. Rüschendorf [37], are tomography, ridge-type regression models, and Hoeffding's decomposition (see [37] for references).

Before the precise definition of IPFP we will describe what we mean by a *generating class*. This notion will enable us to assign hypergraph properties to *sets of low-dimensional probability distributions* that come into an iterative process as an *input set*.

Definition 3.1 (Generating class and input set) Let $H = (V, \mathcal{E}')$ be a *reduced hypergraph* (see Definition 1.4) and $P_{\mathcal{E}} = \{P_{E_1} \in \mathcal{P}_{E_1}, \ldots, P_{E_s} \in \mathcal{P}_{E_s}\}$ be a set of low-dimensional probability distributions (for $i = 1, 2, \ldots s : E_i \subseteq V$). The set of hypergraph edges $\mathcal{E}'$ is called *generating class* of $P_{\mathcal{E}}$ if every edge from $\mathcal{E}'$ defines domain of one low-dimensional probability distribution from $P_{\mathcal{E}}$, i.e. $\mathcal{E}' = \mathcal{E} = \{E_1, \ldots, E_s\}$. The set of distributions $\mathcal{P}_{\mathcal{E}}$ that has a set of edges of a reduced hypergraph as its generating set will be referred to as *input set*.

It should be noted that in the notation of S. Haberman [17] (which we will use as well) *α-acyclic generating classes* is denoted as *decomposable* and *β-acyclic generating classes* is denoted as *totally decomposable* .

Definition 3.2 (IPFP) Let $\mathcal{E} = \{E_1, \ldots E_s\}$ be a *generating class* of *input set* $\mathcal{P}_{\mathcal{E}}$, and $Q_{(0)} \in \mathcal{P}$ be an initial probability distribution. Then *iterative proportional fitting procedure*

is computational process defined for $x \in \mathbb{X}_V$ and for i and $j = ((i-1) \bmod s+1)$ such that $P_{E_j} \ll Q_{(i)}^{E_j}$ by the following formula:

$$Q_{(i)}(x) = \begin{cases} 0, & \text{for } x \text{ such that } Q_{(i-1)}^{E_j}(x^{E_j}) = 0, \\ Q_{(i-1)}(x)\dfrac{P_{E_j}(x^{E_j})}{Q_{(i-1)}^{E_j}(x^{E_j})}, & \text{for } x \text{ such that } Q_{(i-1)}^{E_j}(x^{E_j}) > 0. \end{cases} \tag{3.1}$$

The following example illustrate how the computations of IPFP are performed.

Example 3.2 (IPFP)
Suppose that *input set* consists of two probability distributions

$$P_{\{1,2\}}(X_1, X_2) \text{ and } P_{\{2,3\}}(X_2, X_3).$$

Values of these distributions are denoted by small characters and can be seen on Figure 3.1. Figures 3.1, 3.2, and 3.3) correspond to computational steps 0,1, and 2, respectively. The joint distributions computed during the iterative process is denoted by $Q_{(i)} \in \mathcal{P}_{\{1,2,3\}}, i = 0, 1, 2, \ldots$.

- The initial distribution $Q_{(0)} \in \mathcal{P}_{\{1,2,3\}}$ is uniform (see Figure 3.1).
- The operation of *fitting* input distribution $P_{\{1,2\}}(X_1, X_2)$ brings joint probability distribution $Q_{(1)}$ to have its corresponding marginal $Q_{(1)}^{\{1,2\}}$ equal to the distribution being fitted (Figure 3.2). For example

$$\begin{aligned} &Q_{(1)}(X_1 = 1, X_2 = 1, X_3 = 1) = \\ &= \frac{Q_{(0)}(X_1 = 1, X_2 = 1, X_3 = 1) \cdot P_{\{1,2\}}(X_1 = 1, X_2 = 1)}{Q_{(0)}(X_1 = 1, X_2 = 1, X_3 = 1) + Q_{(0)}(X_1 = 1, X_2 = 1, X_3 = 2)} \\ &= \frac{\frac{1}{8} \cdot a}{\frac{1}{8} + \frac{1}{8}} = \frac{a}{2}. \end{aligned}$$

- *Fitting* the second input distribution $P_{\{2,3\}}(X_2, X_3)$ has similar effect on marginal $Q_{(2)}^{\{2,3\}}$ (Figure 3.3). For example

$$\begin{aligned} &Q_{(2)}(X_1 = 1, X_2 = 1, X_3 = 1) = \\ &= \frac{Q_{(1)}(X_1 = 1, X_2 = 1, X_3 = 1) \cdot P_{\{2,3\}}(X_2 = 1, X_3 = 1)}{Q_{(1)}(X_1 = 1, X_2 = 1, X_3 = 1) + Q_{(1)}(X_1 = 2, X_2 = 1, X_3 = 1)} \\ &= \frac{\frac{a}{2} \cdot e}{\frac{a}{2} + \frac{b}{2}} = \frac{a \cdot e}{a + b}. \end{aligned}$$

From this figure it can be seen that if

$$e + f = a + b \text{ and } g + h = c + d, \tag{3.2}$$

then $Q_{(2)}$ preserves its marginal from previous step, i.e.

$$Q_{(2)}^{\{1,2\}} = Q_{(1)}^{\{1,2\}} = P_{\{1,2\}}.$$

The condition 3.2 equivalently written as $P^{\{2\}}_{\{1,2\}} = P^{\{2\}}_{\{2,3\}}$ is a property of *input set* which we will later define to be a condition of *weak consistency.*

- If we would continue by fitting $P_{\{1,2\}}(X_1, X_2)$ again it would not make any change to $Q_{(2)}$. Let us check one value of $Q_{(3)}$ again:

$$\begin{aligned} & Q_{(3)}(X_1 = 1, X_2 = 1, X_3 = 1) = \\ = \; & \frac{Q_{(2)}(X_1 = 1, X_2 = 1, X_3 = 1) \cdot P_{\{1,2\}}(X_2 = 1, X_3 = 1)}{Q_{(2)}(X_1 = 1, X_2 = 1, X_3 = 1) + Q_{(2)}(X_1 = 2, X_2 = 1, X_3 = 1)} \\ = \; & \frac{\frac{a \cdot e}{a+b} \cdot a}{\frac{a \cdot e}{a+b} + \frac{a \cdot f}{a+b}} = \frac{a \cdot e}{e + f}, \end{aligned}$$

  which, having condition 3.2 fulfilled, equals to $\frac{a \cdot e}{a+b}$.

- If we would continue by fitting $P_{\{2,3\}}(X_2, X_3)$, again, we could derive, in the same way, that it would not bring any change to $Q_{(3)}$.

For this example, we may conclude that if *input set* is *consistent* IPFP stops after two iterations. Latter, we will continue with this example for the case when condition 3.2 does not hold (see Example 3.4).

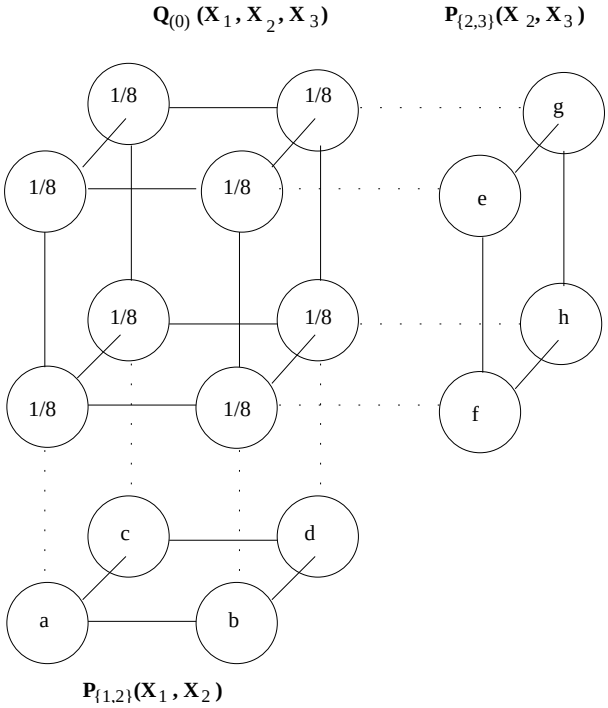


Figure 3.1: Initial and input distributions of IPFP

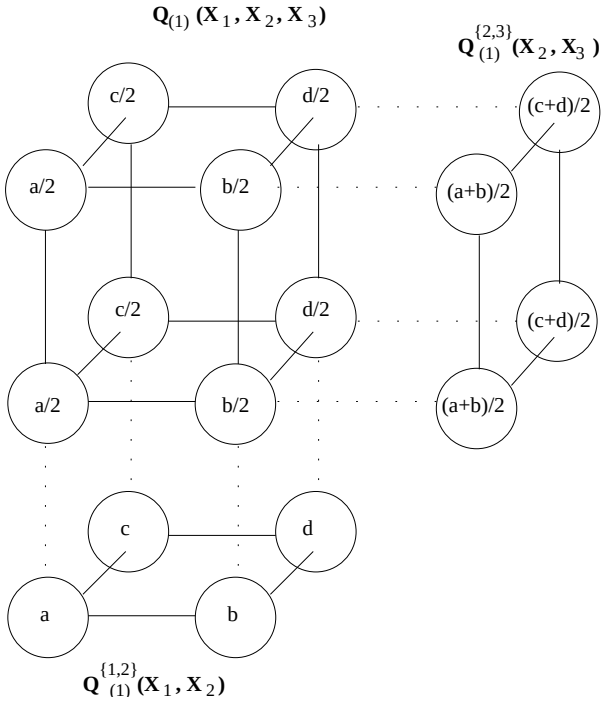


Figure 3.2: Joint distribution and its marginals after fitting P_1

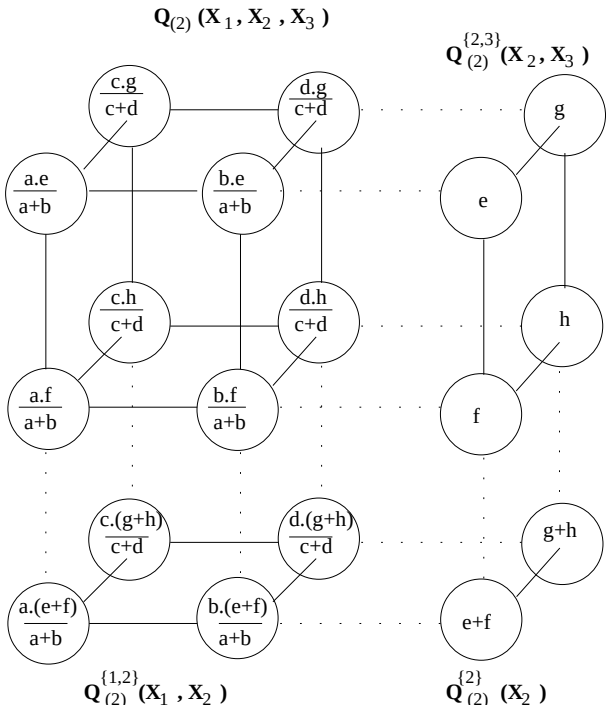


Figure 3.3: Joint distribution and its marginals after fitting P_1 and P_2

The following theorem is essential for *Space-saving implementation of IPFP*, which was proposed by R. Jiroušek. The theorem, proved in [22], is instructive and explains the computational process of *Space-saving implementation of IPFP*. We will use it latter to prove a sufficient condition for IPFP convergence within *one cycle* (s iterative steps).

Theorem 3.1 (Space-saving implementation of IPFP) *Let $\mathcal{E}$ be a generating class of input set $\mathcal{P}_\mathcal{E}$, such that there exists $P \in \mathcal{S}_\mathcal{E}, P \ll Q_{(0)}$, and distributions $Q_{(i)}, i = 1, 2, \ldots$ be computed according to formula 3.1 of Definition 3.2. Further let a class $\mathcal{F} \geq \mathcal{E}$ be decomposable and $Q_{(0)} \in \mathcal{R}_\mathcal{F}$. If for every $j = ((i-1) \bmod s + 1)$ class $\mathcal{F}$ is reordered so that $E_j \subseteq F_{\delta(1)}$ and $\{F_{\delta(1)}, F_{\delta(2)}, \ldots, F_{\delta(t)}\}$ satisfies RIP, and $A_k = \cup_{l=1}^{k-1} F_{\delta(l)}$ then for every finite $i = 1, 2, \ldots, Q_{(i)} \in \mathcal{R}_\mathcal{F}$ and*

$$Q_{(i)}^{F_{\delta(k)}} = \begin{cases} P_{E_j} \dfrac{Q_{(i-1)}^{F_{\delta(k)}}}{Q_{(i-1)}^{E_j}} & \text{for } k = 1 \\ Q_{(i)}^{F_{\delta(k)} \cap A_k} \dfrac{Q_{(i-1)}^{F_{\delta(k)}}}{Q_{(i-1)}^{F_{\delta(k)} \cap A_k}} & \text{for } k = 2, \ldots, t. \end{cases}$$

The behavior of IPFP varies according to *consistency* of *input set*. The definitions of *strong* and *weak consistency* follows.

Definition 3.3 (Strong consistency)
Input set $\{P_{E_1}, \ldots, P_{E_s}\}$ is *strongly consistent* iff

$$\mathcal{S}_\mathcal{E} = \{Q(x) \in \mathcal{P} : Q^{E_i} = P_{E_i}, i = 1, 2, \ldots, s\} \neq \emptyset.$$

The following is the definition of *weak consistency* that is sometimes called *pair-wise consistency* or *projectivity*.

Definition 3.4 (Weak consistency)
Input set $\{P_{E_1}, \ldots, P_{E_s}\}$ is *weakly consistent* iff

$$\forall k, l \in \{1, \ldots, s\} : P_{E_k}^{E_k \cap E_l} = P_{E_l}^{E_k \cap E_l}.$$

3.2 IPFP on consistent input set

If an iterative step of IPFP is compared with Theorem 2.1 it can be immediately seen that in every step IPFP computes I_1-projection of the distribution from previous iteration to an *additively constrained set* $\mathcal{S}_{\{E_j\}}, j = 1, 2, \ldots s$. Sets $\mathcal{S}_{\{E_j\}}$ are step by step alternating.

Theorem 2.1 suppose that distribution S_E that defines *additively constrained set* $\mathcal{S}_{\{E\}}$ is dominated by marginal of the projected distribution Q^E. Therefore, in order to allow its application, the following lemma is necessary.

Lemma 3.1 (Inheritance of dominance) *If there exists a probability distribution $P \in \mathcal{S}_\mathcal{E}$ such that $P \ll Q_{(0)}$ and $Q_{(i)}$ for $i = 1, 2, \ldots$ is computed according to formula 3.1 of Definition 3.2 then for $i = 1, 2, \ldots$ and $j = ((i-1) \bmod s + 1) : P^{E_j} \ll Q_{(i-1)}^{E_j}$.*

Proof. We will prove this lemma by induction over i. Since it holds that for $P \in \mathcal{S}_\mathcal{E} : P \ll Q_{(0)}$ it suffices to show that:

$$P \ll Q_{(i-1)} \Rightarrow P \ll Q_{(i)}, \quad \text{for} \quad i = 1, 2, \ldots .$$

Assume $P \ll Q_{(i-1)}$ It follows from formula 3.1 of Definition 3.2 that for all $x \in \mathbb{X}$ it holds

$$Q_{(i)}(x) = 0 \quad \text{if and only if} \quad Q_{(i-1)}(x) = 0 \text{ or } P^{E_j}(x^{E_j}) = 0.$$

$P \ll Q_{(i-1)}$, which means that for all $x \in \mathbb{X}$:

$$Q_{(i-1)}(x) = 0 \;\; \Rightarrow \;\; P(x) = 0.$$

For $y \in \mathbb{X}, y^{E_j} = x^{E_j}$:

$$P^{E_j}(x^{E_j}) = 0 \;\; \Rightarrow \;\; P(y) = 0.$$

This proves that

$$Q_{(i)}(x) = 0 \;\; \Rightarrow \;\; P(x) = 0,$$

which means that $P \ll Q_{(i)}$. □

By now, we are ready to present the Csiszár's proof [10] of main theorem of **IPFP convergence** in a narrow way.

Theorem 3.2 (Convergence of IPFP) *Let $\{P_1, \ldots, P_s\}$ be a strongly consistent input set (i.e. $\mathcal{S}_\mathcal{E} \neq \emptyset$). If there exists a probability distribution $Q \in \mathcal{S}_\mathcal{E}$ such that $Q \ll Q_{(0)}$ then sequence of the joint distributions computed by formula 3.1 of Definition 3.2 is convergent. In the case of convergence*

$$Q^* = \lim_{n\to\infty} Q_{(n)} = \pi_{\mathcal{S}_\mathcal{E}} Q_{(0)}.$$

Proof. The assertion of the theorem is a direct consequence of Theorem 2.1, Lemma 3.1, and Theorem 2.5. It follows from Theorem 2.1 that distributions $Q_{(n)}$ computed by formula 3.1 are I_1-projections to $\mathcal{S}_{\{E_j\}}, E_j \in \mathcal{E}$, where $j = ((n-1) \bmod s + 1)$. Lemma 3.1 is used to assure that $Q^{E_j} \ll Q_{(i-1)}^{E_j}$, which allows the application of Theorem 2.5, which proves the IPFP convergence. □

Remark. IPFP convergence will be latter referred to as **C1.IPFP** property. The fact that the procedure converges to $\pi_{\mathcal{S}_\mathcal{E}} Q_{(0)}$ will be referred as **C3.IPFP** property.. The IPFP convergence theorem implies that the ordering of the *input set* has no influence on the resulting distribution of the procedure (**C2.IPFP** property).

It is easy to prove that if the initial distribution belongs to $\mathcal{R}_\mathcal{E}$ then all distributions computed within finite number of IPFP iterations belongs to $\mathcal{R}_\mathcal{E}$ as well. On the other side, if the initial distribution does not belong to $\mathcal{R}_\mathcal{E}$ then all distributions computed by IPFP within finite number of iterations does not belong to $\mathcal{R}_\mathcal{E}$ either. This assertions are formalized in the following theorem.

Theorem 3.3 (Inheritance of factorizability (G1.IPFP)) *Let $Q_{(0)} \in \mathcal{P}$ and $Q_{(i)} \in \mathcal{P}, i = 1, 2, \ldots$ are computed by IPFP process (i.e. formula 3.1 of Definition 3.2) applied to the* input set *$\mathcal{P}_\mathcal{E}$. Then for all (finite) $i = 1, 2, \ldots$ it holds that*

$$Q_{(i)} \in \mathcal{R}_\mathcal{E} \;\; \text{iff} \;\; Q_{(i-1)} \in \mathcal{R}_\mathcal{E}.$$

Proof.

(A) $[Q_{(i-1)} \in \mathcal{R}_{\mathcal{E}} \Longrightarrow Q_{(i)} \in \mathcal{R}_{\mathcal{E}}]$

Since $Q_{(i-1)} \in \mathcal{R}_{\mathcal{E}}$ we can write $Q_{(i-1)}(x) = \prod_{i=1}^{s} \psi_{E_i}(x^{E_i})$. Hence for $x : Q_{(i-1)}^{E_j}(x^{E_j}) > 0$

$$Q_{(i)}(x) \quad = \quad Q_{(i-1)}(x)\frac{P_{E_j}(x^{E_j})}{Q_{(i-1)}^{E_j}(x^{E_j})} = \prod_{i=1}^{s} \psi_{E_i}(x^{E_i})\frac{P_{E_j}(x^{E_j})}{Q_{(i-1)}^{E_j}(x^{E_j})}$$

Let $\phi_{E_j}(x^{E_j}) = \psi_{E_j}(x^{E_j})\dfrac{Q_{(i-1)}^{E_j}(x^{E_j})}{P_{E_j}(x^{E_j})}$ and if $i \neq j : \phi_{E_i}(x^{E_i}) = \psi_{E_i}(x^{E_i})$. For x such that $Q_{(i-1)}^{E_j}(x^{E_j}) = 0$ also $P_{E_j}(x^{E_j}) = 0$. In this case $Q_{(i)}^{E_j}(x^{E_j})$ is defined to be equal zero too. It is obvious that for such x the function $\phi_{E_i}(x^{E_i})$ can be defined to be equal zero as well. Then

$$Q_{(i)}(x) \quad = \quad \prod_{i=1}^{s} \phi_{E_i}(x^{E_i}), \quad \text{i.e.} \quad Q_{(i)} \in \mathcal{R}_{\mathcal{E}}.$$

(B) $[Q_{(i)} \in \mathcal{R}_{\mathcal{E}} \Longrightarrow Q_{(i-1)} \in \mathcal{R}_{\mathcal{E}}]$

Similarly to **(A)** we can write $Q_{(i)}(x) = \prod_{i=1}^{s} \phi_{E_i}(x^{E_i})$. Hence for $x : P_{E_j}(x^{E_j}) > 0$

$$Q_{(i-1)}(x) \quad = \quad Q_{(i)}(x)\frac{Q_{(i-1)}^{E_j}(x^{E_j})}{P_{E_j}(x^{E_j})} = \prod_{i=1}^{s} \phi_{E_i}(x^{E_i})\frac{Q_{(i-1)}^{E_j}(x^{E_j})}{P_{E_j}(x^{E_j})}$$

Let $\psi_{E_j}(x^{E_j}) = \phi_{E_j}(x^{E_j})\dfrac{Q_{(i-1)}^{E_j}(x^{E_j})}{P_{E_j}(x^{E_j})}$ and if $i \neq j : \psi_{E_i}(x^{E_i}) = \phi_{E_i}(x^{E_i})$. For x such that $P_{E_j}(x^{E_j}) = 0$ also $Q_{(i)}^{E_j}(x^{E_j}) = 0$ and consequently $\phi_{E_j}(x^{E_j}) = 0$. In this case we define $\psi_{E_j}(x^{E_j}) = 0$ as well. Then

$$Q_{(i-1)}(x) \quad = \quad \prod_{i=1}^{s} \psi_{E_i}(x^{E_i}), \quad \text{i.e.} \quad Q_{(i-1)} \in \mathcal{R}_{\mathcal{E}}.$$

□

Theorem 3.4 *If the input set $\{P_1, \ldots, P_s\}$ is strongly consistent, i.e. $\mathcal{S}_{\mathcal{E}} \neq \emptyset$, there exists a probability distribution $Q \in \mathcal{S}_{\mathcal{E}}$ such that $Q \ll Q_{(0)}$, and $Q_{(0)} \in \mathcal{R}_{\mathcal{E}}$ then the sequence of probability distributions computed by IPFP (formula* 3.1 *of Definition* 3.2*) converges to*

$$Q^* \in \bar{\mathcal{R}}_{\mathcal{E}} \cap \mathcal{S}_{\mathcal{E}}.$$

Proof. If there exists a probability distribution $Q \in \mathcal{S}_{\mathcal{E}}$ such that $Q \ll Q_{(0)}$ then sequence of the joint distributions computed by formula 3.1 of Definition 3.2 is convergent to $Q^* \in \mathcal{S}_{\mathcal{E}}$ (Theorem 3.2). If $Q_{(0)} \in \mathcal{R}_{\mathcal{E}}$ then according to Theorem 3.3 $Q_{(i)} \in \mathcal{R}_{\mathcal{E}}, i = 1, 2, \ldots$, thus $Q^* \in \bar{\mathcal{R}}_{\mathcal{E}}$. □

For the uniform distribution U it holds that $U \in \mathcal{R}_{\mathcal{E}}$ for any $\mathcal{E}$. Thus, it turns out that if the initial distribution is uniform then according to Theorem 3.3 the limit distribution $Q^* \in \bar{\mathcal{R}}_{\mathcal{E}}$.

Theorem 3.5 (I_2-projection to $\bar{\mathcal{R}}_\mathcal{E}$) *If an input set $\{P_1, \ldots, P_s\}$ is strongly consistent, i.e. $\mathcal{S}_\mathcal{E} \neq \emptyset$, there exists a probability distribution $Q \in \mathcal{S}_\mathcal{E}$ such that $Q \ll Q_{(0)}$, and $Q_{(0)} \in \mathcal{R}_\mathcal{E}$ then the sequence of probability distributions computed by IPFP (formula 3.1 of Definition 3.2) converges to*

$$Q^* = \pi'_{\bar{\mathcal{R}}_\mathcal{E}} Q.$$

Proof. We have shown that for the limit distribution Q^* of the iterative process defined by formula 3.1 of Definition 3.2 it holds $Q^* \in \mathcal{S}_\mathcal{E}$ (Theorem 3.2) and $Q^* \in \bar{\mathcal{R}}_\mathcal{E}$ (Theorem 3.4). Therefore it follows from Theorem 2.9 that Q^* is also I_2-projection of any $Q \in \mathcal{S}_\mathcal{E}$ to $\bar{\mathcal{R}}_\mathcal{E}$. □

Suppose we want to solve another problem. A probability distribution Q is given (e.g. estimated from data). We search for another probability distribution that factorizes with respect to given $\mathcal{E}$, i.e it belongs to $\bar{\mathcal{R}}_\mathcal{E}$, such that it is "a best" approximation of known data Q. It is a consequence of the previous theorem, that IPFP is applicable to such a problem whereas the criteria is I-divergence. The best approximation equals

$$\arg \min_{\{R \in \bar{\mathcal{R}}_\mathcal{E}\}} I(Q \parallel R) = \pi'_{\bar{\mathcal{R}}_\mathcal{E}} Q.$$

Next, an example of an *input set* consisting of four distributions is going to be introduced. It will be used to demonstrate the behavior of IPFP.

Example 3.3 (Consistent input set) In Table 3.1 a *consistent input set of distributions* is defined. Figure 3.4 displays that in this example IPFP convergence rate is high since after 2 cycles (i.e. 8 iterations) the values of $Q_{(i)}$ differ less than 0.01%. Since it might give false impression that joint distribution does not change after 4th iteration Figure 3.5 displays detail of IPFP behavior in between steps 4 and 8.

Iteration i	$Q_{(i)}(X_1 = 1, X_2 = 1, X_3 = 2, X_4 = 1)$
4	0.15685
8	0.15691
12	0.15691

The plot in Figure 3.4 and all plots in next chapters describe dependence between probability values (vertical axis) of joint distribution $Q_{(i)}$ for one combination of random variables' values $X_1 = 1, X_2 = 1, X_3 = 2, X_4 = 1$ and number of iterations i (horizontal axis).

P_{E_1}	$X_2 = 1$	$X_2 = 2$
$X_1 = 1$	89/288	43/144
$X_1 = 2$	7/36	19/96

P_{E_2}	$X_3 = 1$	$X_3 = 2$
$X_2 = 1$	17/96	47/144
$X_2 = 2$	43/288	25/72

P_{E_3}	$X_4 = 1$	$X_4 = 2$
$X_3 = 1$	31/144	1/9
$X_3 = 2$	67/144	5/24

P_{E_4}	$X_4 = 1$	$X_4 = 2$
$X_1 = 1$	15/32	5/36
$X_1 = 2$	61/288	13/72

Table 3.1: Consistent input set of distributions

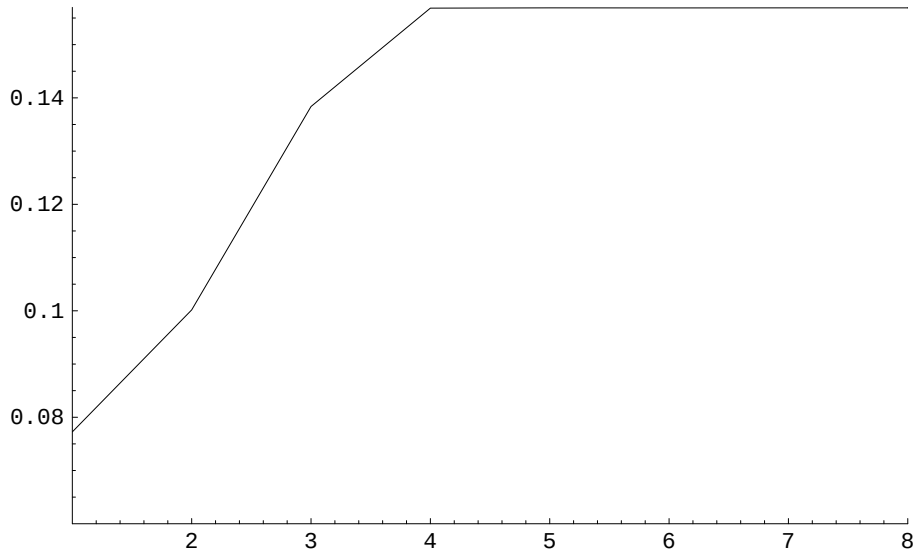


Figure 3.4: IPFP on consistent input $\{P_{E_1}, P_{E_2}, P_{E_3}, P_{E_4}\}$

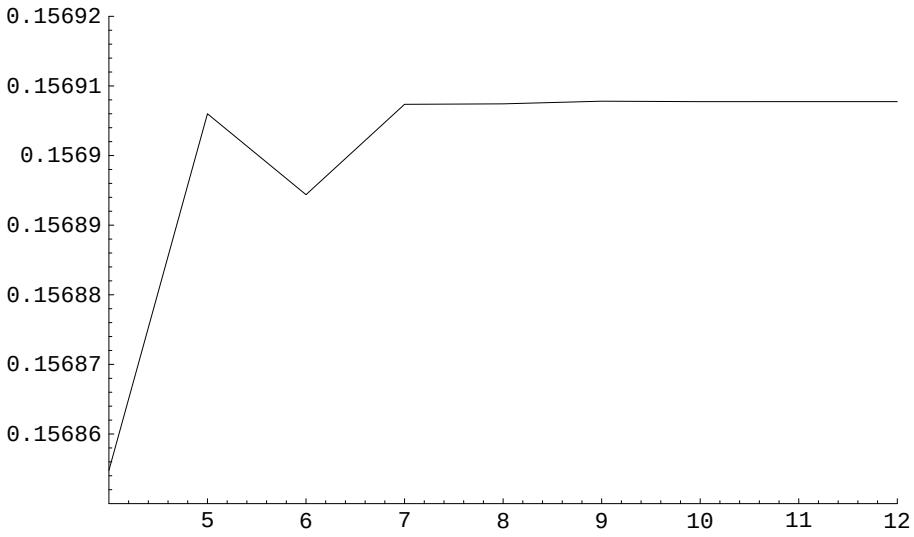


Figure 3.5: IPFP on consistent input $\{P_{E_1}, P_{E_2}, P_{E_3}, P_{E_4}\}$ (detail)

3.3 IPFP on inconsistent input set

What can be said about the IPFP process when the *input set* $\mathcal{P}_{\mathcal{E}}$ is not consistent? First of all, we will continue with Example 3.2 to show IPFP behavior when probability distributions from *input set* are *inconsistent*.

Example 3.4 (IPFP on inconsistent input set (continued Example 3.2)) Suppose that the *input set* consists of two distributions $P_{\{1,2\}}(X_1, X_2)$ and $P_{\{2,3\}}(X_2, X_3)$. Their values were defined by Figure 3.1 in Example 3.2. Recall, that after two iterations of IPFP we have got probability distribution $Q_{(2)}$. It is given by the following Figure 3.6.

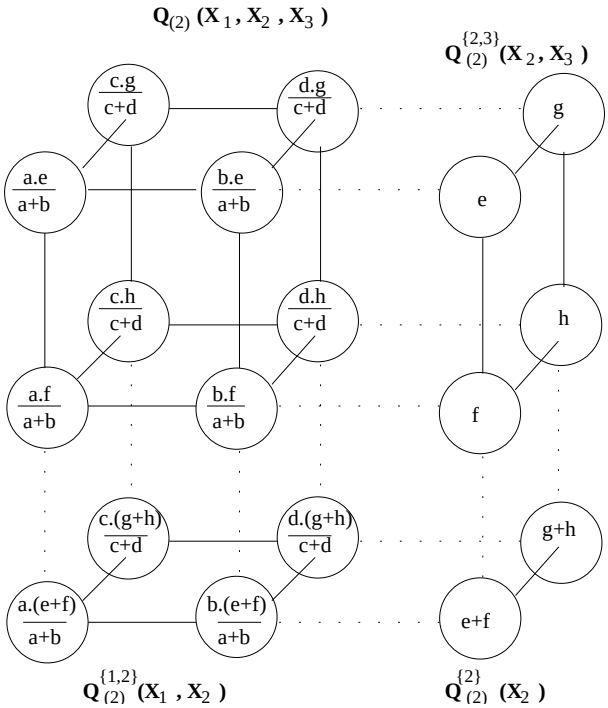


Figure 3.6: Joint distribution and its marginals after fitting P_1 and P_2

- We have supposed that the *input set* is inconsistent therefore

$$e + f \neq a + b \quad \text{and} \quad g + h \neq c + d.$$

Thus, by fitting $P_{\{1,2\}}(X_1, X_2)$ again, we get distribution $Q_{(3)}$, displayed on Figure 3.7. Let us check one value of $Q_{(3)}$ again:

$$\begin{aligned} & Q_{(3)}(X_1 = 1, X_2 = 1, X_3 = 1) = \\ &= \frac{Q_{(2)}(X_1 = 1, X_2 = 1, X_3 = 1) \cdot P_{\{1,2\}}(X_1 = 1, 2_3 = 1)}{Q_{(2)}(X_1 = 1, X_2 = 1, X_3 = 1) + Q_{(2)}(X_1 = 1, X_2 = 1, X_3 = 2)} \\ &= \frac{\frac{a \cdot e}{a+b} \cdot a}{\frac{a \cdot e}{a+b} + \frac{a \cdot f}{a+b}} = \frac{a \cdot e}{e + f}. \end{aligned}$$

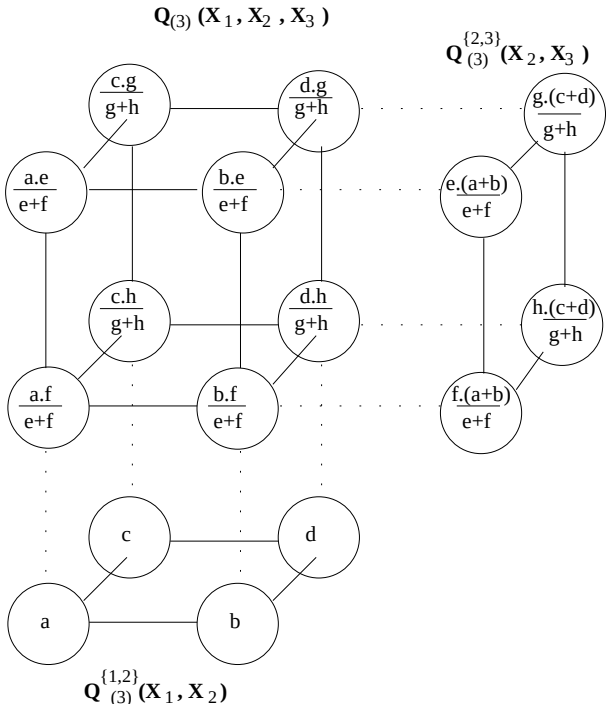


Figure 3.7: Joint distribution and its marginals after fitting P_1, P_2, and P_1 again

- By fitting $P_{\{2,3\}}(X_2, X_3)$ again we get distribution $Q_{(4)}$ which turns out to be equal $Q_{(2)}$ already displayed on Figure 3.6. Let us check one value of $Q_{(4)}$ again:

$$\begin{aligned} & Q_{(4)}(X_1 = 1, X_2 = 1, X_3 = 1) = \\ = \; & \frac{Q_{(3)}(X_1 = 1, X_2 = 1, X_3 = 1) \cdot P_{\{2,3\}}(X_2 = 1, X_3 = 1)}{Q_{(3)}(X_1 = 1, X_2 = 1, X_3 = 1) + Q_{(3)}(X_1 = 2, X_2 = 1, X_3 = 1)} \\ = \; & \frac{\frac{a \cdot e}{e+f} \cdot e}{\frac{a \cdot e}{e+f} + \frac{b \cdot e}{e+f}} = \frac{a \cdot e}{a+b}. \end{aligned}$$

We have shown that $Q_{(4)} = Q_{(2)}$, consequently $Q_{(5)} = Q_{(3)}$, which proves that IPFP stabilizes in the cycle of two joint probability distributions $Q_{(2i)} = Q_{(2)}$ and $Q_{(2i+1)} = Q_{(3)}$ for $i = 2, 3, \ldots$ (**I1.IPFP** property). The fact that IPFP stabilized in the limit cycle might be exploited to get a *unique joint probability distribution* either as an arithmetic Q_a^* or geometric averages Q_g^* of distributions from the limit cycle. A question is whether these distributions obey some reasonable properties we would like the resulting *joint probability distribution* to have. We will further continue with this example in Chapter 5 (Example 5.1). In this chapter several requirements on *joint probability distribution* to be regarded as a reasonable result of *inconsistent knowledge integration* are discussed in detail.

After a number of computational experiments with the regular IPFP applied to both the weakly consistent and inconsistent *input sets* we believe that only two possibilities may take place.

Conjecture 3.1 *Let $Q_0 \in \mathcal{P}$ be the uniform probability distribution defined for all variables $(X_i)_{i\in V}$. For any strongly inconsistent input set the application of the IPFP (Definition 3.2) leads to one of the following two situations:*

- *The computation fails because there exists $1 < j \leq s$ such that the distribution P_{E_j} is not dominated by $Q_{(j-1)}^{E_j}$ (i.e. $\exists x : P_j(x) > 0, Q_{(j-1)}^{C_j}(x) = 0$), and thus the distribution $Q_{(j)}$ is not defined.*
- *The sequence $Q_{(1)}, Q_{(2)}, \ldots$ tends to come in cycles. More precisely: for all $1 \leq j \leq s$ the subsequences $Q_{(j)}, Q_{(j+s)}, Q_{(j+2s)}, \ldots, Q_{(j+ns)}, \ldots$ are convergent. At least two limit distributions of these subsequences differ each from other.*

Let us illustrate these two possibilities using another example.

Example 3.5 (Three two-dimensional distributions)
Consider the *input set* which consists of three two-dimensional distributions given in Table 3.2

$$\{P_1(X_1, X_2), P_2(X_2, X_3), P_3(X_1, X_3)\}.$$

Table 3.2: Input distributions $\{P_1,\ P_2,\ P_3\}$

P_1	$X_2 = 1$	$X_2 = 2$
$X_1 = 1$	$0.5 - \varepsilon$	$0 + \varepsilon$
$X_1 = 2$	$0 + \varepsilon$	$0.5 - \varepsilon$

P_2	$X_3 = 1$	$X_3 = 2$
$X_2 = 1$	$0.5 - \varepsilon$	$0 + \varepsilon$
$X_2 = 2$	$0 + \varepsilon$	$0.5 - \varepsilon$

P_3	$X_3 = 1$	$X_3 = 2$
$X_1 = 1$	$0 + \varepsilon$	$0.5 - \varepsilon$
$X_1 = 2$	$0.5 - \varepsilon$	$0 + \varepsilon$

Taking $\varepsilon = 0$, the computation fails because the distribution P_3 is not dominated by $Q_2^{E_3}$. Namely, for $x_{12} = (X_1 = 1, X_3 = 2)$ $P_3(x_{12}) > 0$ and $Q_2^{E_3}(x_{12}) = 0$ as well as for $x_{21} = (X_1 = 2, X_3 = 1)$ $P_3(x_{21}) > 0$ and $Q_2^{E_3}(x_{21}) = 0$, and thus the distribution Q_3 is not defined. The respective distributions Q_1 and Q_2 are in Table 3.3.

Table 3.3: Application of IPFP to the distributions from Table 3.2 with $\varepsilon = 0$.

X_1	X_2	X_3	Q_0	Q_1	Q_2	Q_3
1	1	1	0.1250	0.2500	0.5000	0.0000
1	1	2	0.1250	0.2500	0.0000	-
1	2	1	0.1250	0.0000	0.0000	0.0000
1	2	2	0.1250	0.0000	0.0000	-
2	1	1	0.1250	0.0000	0.0000	0.0000
2	1	2	0.1250	0.0000	0.0000	-
2	2	1	0.1250	0.2500	0.0000	0.0000
2	2	2	0.1250	0.2500	0.0000	-

Table 3.4: Application of IPFP to the distributions from Table 3.2 with $\varepsilon = 0.1$. and ordering $\{P_1, P_2, P_3\}$

X_1	X_2	X_3	Q_0	Q_1	Q_2	Q_3	...	Q_{19+3n}	Q_{20+3n}	Q_{21+3n}
1	1	1	0.1250	0.2000	0.3200	0.0941		0.1562	0.2439	0.1000
1	1	2	0.1250	0.2000	0.0800	0.2000		0.2438	0.1000	0.1562
1	2	1	0.1250	0.0500	0.0200	0.0059		0.0000	0.0000	0.0000
1	2	2	0.1250	0.0500	0.0800	0.2000		0.1000	0.1562	0.2438
2	1	1	0.1250	0.0500	0.0800	0.2000		0.1000	0.1562	0.2438
2	1	2	0.1250	0.0500	0.0200	0.0059		0.0000	0.0000	0.0000
2	2	1	0.1250	0.2000	0.0800	0.2000		0.2438	0.1000	0.1562
2	2	2	0.1250	0.2000	0.3200	0.0941		0.1562	0.2439	0.1000

Table 3.5: Application of IPFP to the distributions from Table 3.2 with $\varepsilon = 0.1$ and ordering $\{P_3, P_2, P_1\}$

X_1	X_2	X_3	Q_0	Q_1	Q_2	Q_3	...	Q_{19+3n}	Q_{20+3n}	Q_{21+3n}
1	1	1	0.1250	0.0500	0.0800	0.2000		0.1000	0.1561	0.2438
1	1	2	0.1250	0.2000	0.0800	0.2000		0.2438	0.1000	0.1562
1	2	1	0.1250	0.0500	0.0200	0.0059		0.0000	0.0000	0.0000
1	2	2	0.1250	0.0200	0.3200	0.0941		0.1562	0.2439	0.1000
2	1	1	0.1250	0.0200	0.3200	0.0941		0.1562	0.2439	0.1000
2	1	2	0.1250	0.0500	0.0200	0.0059		0.0000	0.0000	0.0000
2	2	1	0.1250	0.2000	0.0800	0.2000		0.2438	0.1000	0.1562
2	2	2	0.1250	0.0500	0.0800	0.0200		0.1000	0.1561	0.2438

The other situation occurs for $\varepsilon = 0.1$. In this case the sequence of distribution $Q_1, Q_2, \ldots$ "converges" to the cycle of the three distributions shown in Table 3.4. When the order of input set $\mathcal{P}_\mathcal{E}$ is $\{P_3, P_2, P_1\}$ then the distributions in the cycle are different (see Table 3.5).

Let us stress that even for a weakly consistent *input set* (but strongly inconsistent) the systems of limit distributions may differ when the distributions in $\mathcal{P}_\mathcal{E}$ are reordered. From the next example it can be seen that both *arithmetic and geometric averages* of distributions from limit cycles are dependent on the order of distributions in *input set*.

Example 3.6 (Four two-dimensional distributions) Suppose that *input set* consist of four *inconsistent* probability distributions given by the following table. Distributions of *input set* are given in Table 3.6.

P_{E_1}	$X_2 = 1$	$X_2 = 2$
$X_1 = 1$	1/10	3/10
$X_1 = 2$	2/10	4/10

P_{E_2}	$X_3 = 1$	$X_3 = 2$
$X_2 = 1$	7/10	1/10
$X_2 = 2$	1/10	1/10

P_{E_3}	$X_4 = 1$	$X_4 = 2$
$X_3 = 1$	2/10	1/10
$X_3 = 2$	2/10	5/10

P_{E_4}	$X_4 = 1$	$X_4 = 2$
$X_1 = 1$	3/10	3/10
$X_1 = 2$	3/10	1/10

Table 3.6: Inconsistent input set of distributions

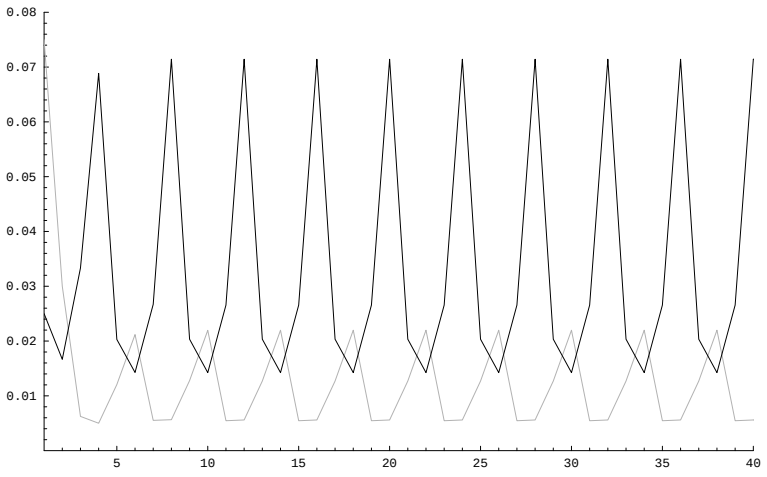


Figure 3.8: IPFP behavior on inconsistent input for the order $\{E_1, E_2, E_3, E_4\}$ (upper dark line) and $\{E_4, E_2, E_1, E_3\}$ (lower gray line)

4 Decomposable generating classes

The theory we are going to present at the beginning of this chapter exploits results of S. Haberman [17] for *log-linear models* on a large scale. However, the theory of *log-linear models* assumes that all probability distributions are strictly positive. Our definition of I_2-*projection* to *multiplicatively constrained sets* $\mathcal{R}_{\mathcal{E}}$ enable us to put aside assumption of strict positivity. Furthermore, some proofs in [17] contain several misprints, several "obvious" facts were not so evident. Therefore we believe it is meaningful to rewrite the fundamental theorems together with their proofs.

In the second part of the chapter new results concerning *decomposable generating classes* are presented. They concern characterization of types and orderings of *generating classes* that guarantee convergence of IPFP within *one* or *two cycles*. The main result is a counterexample to the conjecture that for all *decomposable generating classes* IPFP converges within *two cycles* no matter how *generating class* is ordered.

At first, we are going to introduce some class functions defined in order to enable better characterization of *generating classes*. We follow S. Haberman [17].

Definition 4.1 (Generating class functions) Suppose $\mathcal{E}$ is a *generating class*.
Intersection class $\mathcal{F}(\mathcal{E})$ is defined to be the set of all intersection

$$\mathcal{F}(\mathcal{E}) = \{E \cap E' : E \in \mathcal{E}, E' \in \mathcal{E}, E \neq E'\}.$$

Let F be an element of *intersection class* $\mathcal{F}(\mathcal{E})$. *Raw replication number* $c(F, \mathcal{E})$ is the number of sets in $\mathcal{E}$ that contain F.
Class $\mathcal{F}(F, \mathcal{E})$ contains sets from $\mathcal{F}(\mathcal{E})$ that has set F as its proper subset:

$$\mathcal{F}(F, \mathcal{E}) = \{F' \in \mathcal{F}(\mathcal{E}) : F \subset F', F \neq F'\}.$$

The *adjusted replication number* $d(F, \mathcal{E})$ is defined recursively by equation:

$$d(F, \mathcal{E}) = c(F, \mathcal{E}) - 1 - \sum_{F' \in \mathcal{F}(F, \mathcal{E})} d(F', \mathcal{E}),$$

where the recursion is given by the hierarchy of supersets of F in the intersection class $\mathcal{F}(\mathcal{E})$ until there is no superset. Then $d(F', \mathcal{E}) = c(F', \mathcal{E}) - 1$.
Hierarchical class $\mathcal{H}(\mathcal{E})$ generated by $\mathcal{E}$ is the set

$$\{A \subseteq E : E \in \mathcal{E}\}.$$

In order to shed light upon previous definition we will give a simple example to show how these function acts.

Example 4.1 (Generating class functions) Let $\mathcal{E}$ is a *generating class* given by Figure 4.1 i.e. $\mathcal{E} = \{\{1,2,4\},\{2,3,5\},\{2,4,5\},\{4,5,7\},\{4,6,7\},\{5,7,8\}\}$. The following table shows for each $F \in \mathcal{F}(\mathcal{E})$ values of *raw replication number* $c(F, \mathcal{E})$, class $\mathcal{F}(F, \mathcal{E})$, and *adjusted replication number* $d(F, \mathcal{E})$.

F	$\{4,2\}$	$\{5,2\}$	$\{2\}$	$\{4,5\}$	$\{4\}$	$\{5\}$	$\{4,7\}$	$\{5,7\}$	$\{7\}$	$\emptyset$
$c(F,\mathcal{E})$	2	2	3	2	4	4	2	2	3	6
$\mathcal{F}(F,\mathcal{E})$	$\emptyset$	$\emptyset$	{{4,2}, {5,2}}	$\emptyset$	{{4,2}, {4,5}, {4,7}}	{{5,2}, {4,5}, {5,7}}	$\emptyset$	$\emptyset$	{{4,7}, {5,7}}	$\mathcal{F}(\mathcal{E})\setminus\emptyset$
$d(F,\mathcal{E})$	1	1	0	1	0	0	1	1	0	0

Hierarchical class $\mathcal{H}(\mathcal{E})$ generated by $\mathcal{E}$ is the class:

{ ∅, {1}, {2}, {3}, {4}, {5}, {6}, {7}, {8}, {1,2}, {1,4}, {2,3}, {2,4}, {2,5}, {3,5}, {4,5}, {4,6}, {4,7}, {5,7}, {5,8}, {6,7}, {7,8}, {1,2,4}, {2,3,5}, {2,4,5}, {4,5,7}, {4,6,7}, {5,7,8} }.

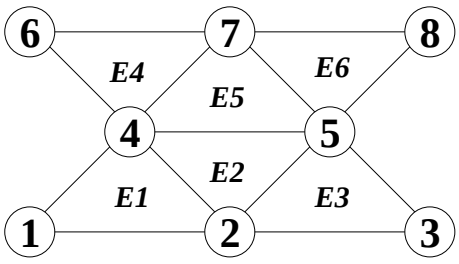


Figure 4.1: Generating class $\mathcal{E} = \{E_1, E_2, \ldots, E_6\}$

An important consequence of Lemma 2.8 based on orderings of $\mathcal{E}$ that meet *Running intersection property* (RIP) follows.

Theorem 4.1 *Let* $\{E_1, E_2, \ldots, E_s\}$ *be an ordering of a decomposable generating class* $\mathcal{E}$ *satisfying RIP and* $P \in \mathcal{P}$. *Define class* $\mathcal{S}_\mathcal{E} = \{S \in \mathcal{P} : S^E = P^E, E \in \mathcal{E}\} \neq \emptyset$. *Further let* $F_{j+1} = E_{j+1} \cap A_j, j = 1, 2, \ldots, s$, *where* $A_j = \cup_{k=1}^{j} E_k$. *Then*

$$\pi'_{\mathcal{R}_V} P(x^V) \quad = \quad \frac{\prod_{j=1,2,\ldots,s} P^{E_j}(x^{E_j})}{\prod_{j=2,\ldots,s} P^{F_j}(x^{F_j})} \cdot U_{V\setminus A_s}$$

Proof. The proof is nothing else than a successive application of Lemma 2.8 with

$$(\mathcal{E}_1)_{(0)} = \mathcal{E} \setminus E_s, (\mathcal{E}_2)_{(0)} = E_s, C = F_s,$$

for $i = 1, 2, \ldots, s-2 : (\mathcal{E}_1)_{(i)} = (\mathcal{E}_1)_{(i-1)} \setminus E_{s-i}$, $(\mathcal{E}_2)_{(i)} = E_{s-i}$, and $C = F_{s-i}$. □

The following Lemma 4.1 exploits decomposition $\mathcal{E}_1, \mathcal{E}_2$ of a generating class $\mathcal{E}$ to show that if an intersection set F belongs to hierarchical class $\mathcal{H}(\mathcal{E}_1) \setminus \mathcal{H}(\mathcal{E}_2)$ then all its supersets from *intersection class* belongs there as well. This property will be used in the proof of Theorem 4.2

Lemma 4.1 *Let $\mathcal{E}_1$ and $\mathcal{E}_2$ be a partition of $\mathcal{E}$, such that for some $E_1 \in \mathcal{E}_1$ and $E_2 \in \mathcal{E}_2$ it holds*

$$\cup_{E_i \in \mathcal{E}_1} E_i \cap \cup_{E_j \in \mathcal{E}_2} E_j = E_1 \cap E_2.$$

Further let F belong to the intersection class $\mathcal{F}(\mathcal{E})$.

$$\textit{If } F \in \mathcal{H}(\mathcal{E}_1) \setminus \mathcal{H}(\mathcal{E}_2) \textit{ then } \mathcal{F}(F, \mathcal{E}) \subset \mathcal{H}(\mathcal{E}_1) \setminus \mathcal{H}(\mathcal{E}_2).$$

Proof.

$$\mathcal{H}(\mathcal{E}_1) \setminus \mathcal{H}(\mathcal{E}_2) \quad = \quad \mathcal{H}(\mathcal{E}_1) \setminus (\mathcal{H}(\mathcal{E}_1) \cap \mathcal{H}(\mathcal{E}_2))$$

Since $(\cup_{E_i \in \mathcal{E}_1} E_i) \cap (\cup_{E_j \in \mathcal{E}_2} E_j) = E_1 \cap E_2 \quad \Rightarrow \quad \mathcal{H}(\mathcal{E}_1) \cap \mathcal{H}(\mathcal{E}_2) = \mathcal{H}(E_1 \cap E_2)$

$$\mathcal{H}(\mathcal{E}_1) \setminus \mathcal{H}(\mathcal{E}_2) \quad = \quad \mathcal{H}(\mathcal{E}_1) \setminus \mathcal{H}(E_1 \cap E_2)$$

$\mathcal{F}(F, \mathcal{E}) \subseteq \mathcal{F}(\mathcal{E})$. It is assumed that $F \in \mathcal{H}(\mathcal{E}_1) \setminus \mathcal{H}(E_1 \cap E_2)$ i.e. F is equal neither to $E_1 \cap E_2$ nor to any of its subsets. Therefore $E_1 \cap E_2 \notin \mathcal{F}(F, \mathcal{E})$. It follows that

$$\forall E' \in \mathcal{E}_1, E'' \in \mathcal{E}_2 : E' \cap E'' \notin \mathcal{F}(F, \mathcal{E}).$$

It is obvious that also

$$\forall E', E'' \in \mathcal{E}_2 : E' \cap E'' \notin \mathcal{F}(F, \mathcal{E}).$$

Consequently, $\mathcal{F}(F, \mathcal{E})$ consists only of $\{E' \cap E''\}$, where $E', E'' \in \mathcal{E}_1$ and therefore

$$\mathcal{F}(F, \mathcal{E}) \subset \mathcal{H}(\mathcal{E}_1) \setminus \mathcal{H}(\mathcal{E}_2).$$

□

The following theorem is of a fundamental importance for design of proofs given in this chapter. Hence we are going to give a revised version of its proof from [17, Theorem 5.1] in spite the fact it is very complex.

Theorem 4.2 (Closed form of I_2-projections on $\mathcal{R}_\mathcal{E}$)
Let $\mathcal{E}$ be a decomposable generating class. Further, let P, a distribution from $\mathcal{P}$, define class $\mathcal{S}_\mathcal{E} = \{S \in \mathcal{P} : S^E = P^E, E \in \mathcal{E}\} \neq \emptyset$. Then $\pi'_{\mathcal{R}_V} P$ exists and

$$\pi'_{\mathcal{R}_V} P(x) \quad = \quad \frac{\prod_{E \in \mathcal{E}} P^E(x^E)}{\prod_{F \in \mathcal{F}(\mathcal{E})} (P^F(x^F))^{d(F,\mathcal{E})}} \cdot U_D(x^D), \quad \text{where} \quad D = V \setminus \bigcup_{E \in \mathcal{E}} E.$$

Proof. The proof is by **induction** over the decomposition of $\mathcal{E}$.

If $\mathcal{E} = \{E\}$ then $\mathcal{F}(\mathcal{E}) = \emptyset$ and by Lemma 2.6

$$\pi'_{\mathcal{R}_V} P(x^V) = P^E(x^E) \cdot U_D(x^D).$$

Assume that the theorem holds for all decomposable generating classes properly included in $\mathcal{E}$. Suppose that $\mathcal{E}_1$ and $\mathcal{E}_2$ are disjoint decomposable classes such that $\mathcal{E} = \mathcal{E}_1 \cup \mathcal{E}_2$ and there exist $E_1 \in \mathcal{E}_1$ and $E_2 \in \mathcal{E}_2$: $(\bigcup_{E_1 \in \mathcal{E}_1} E_1) \cap (\bigcup_{E_2 \in \mathcal{E}_2} E_2) = E_1 \cap E_2$ (since $\mathcal{E}$ is decomposable there exists at least one such partition of $\mathcal{E}$.

The subsequent relationships between functions of $\mathcal{E}$ and functions of $\mathcal{E}_1$ and $\mathcal{E}_2$ will be used:

$$\begin{aligned}
&\mathcal{H}(\mathcal{E}) = \mathcal{H}(\mathcal{E}_1) \cup \mathcal{H}(\mathcal{E}_2) \\
&F \in \mathcal{F}(F, \mathcal{E}) \;\Rightarrow\; c(F, \mathcal{E}) = c(F, \mathcal{E}_1) + c(F, \mathcal{E}_2) \\
&\mathcal{F}(F, \mathcal{E}) = \emptyset \;\text{ and }\; F \in \mathcal{F}(\mathcal{E}) \Rightarrow \mathcal{F}(F, \mathcal{E}_1) = \mathcal{F}(F, \mathcal{E}_2) = \emptyset
\end{aligned}$$

In the rest of the proof we need to prove that relation between $d(F, \mathcal{E})$ and $d(F, \mathcal{E}_1), d(F, \mathcal{E}_2)$ properly coincides with composition of $\mathcal{E}$ from $\mathcal{E}_1, \mathcal{E}_2$. We are going to verify the relation step by step for all possible $F \in \mathcal{F}(\mathcal{E})$.

(A) $F \in \mathcal{F}(\mathcal{E}) : \mathcal{F}(F, \mathcal{E})$ is empty

$d(F, \mathcal{E}) = c(F, \mathcal{E}) - 1 = c(F, \mathcal{E}_1) + c(F, \mathcal{E}_2) - 1$ We will study situations $F \in \mathcal{H}(\mathcal{E}_1) \setminus \mathcal{H}(\mathcal{E}_2)$, $F \in \mathcal{H}(\mathcal{E}_2) \setminus \mathcal{H}(\mathcal{E}_1)$, and $F \in \mathcal{H}(\mathcal{E}_1) \cap \mathcal{H}(\mathcal{E}_2)$ separately.

(1) $F \in \mathcal{H}(\mathcal{E}_1) \setminus \mathcal{H}(\mathcal{E}_2) \;\Rightarrow\; c(F, \mathcal{E}_2) = 0 \;\Rightarrow\; d(F, \mathcal{E}) = c(F, \mathcal{E}_1) - 1 = d(F, \mathcal{E}_1)$.

(2) $F \in \mathcal{H}(\mathcal{E}_2) \setminus \mathcal{H}(\mathcal{E}_1)$ gives analogously to previous $d(F, \mathcal{E}) = d(F, \mathcal{E}_2)$.

(3) $F \in \mathcal{H}(\mathcal{E}_1) \cap \mathcal{H}(\mathcal{E}_2)$. Since $\mathcal{F}(F, \mathcal{E}_1) = \mathcal{F}(F, \mathcal{E}_2) = \emptyset$ we get

$$d(F, \mathcal{E}) = c(F, \mathcal{E}_1) + c(F, \mathcal{E}_2) - 1 = d(F, \mathcal{E}_1) + d(F, \mathcal{E}_2) + 1.$$

Thus we get relations for $d(F, \mathcal{E})$ in the three different cases. Next, we will proof these relations for more complicated case of $F \in \mathcal{F}(\mathcal{E})$ such that $\mathcal{F}(F, \mathcal{E})$ is not empty. We will study $F \in \mathcal{H}(\mathcal{E}_1) \setminus \mathcal{H}(\mathcal{E}_2)$, $F \in \mathcal{H}(\mathcal{E}_2) \setminus \mathcal{H}(\mathcal{E}_1)$, and $F \in \mathcal{H}(\mathcal{E}_1) \cap \mathcal{H}(\mathcal{E}_2)$ separately again.

(B) $F \in \mathcal{F}(\mathcal{E}) : \mathcal{F}(F, \mathcal{E})$ is not empty

(1) If $F \in \mathcal{H}(\mathcal{E}_1) \setminus \mathcal{H}(\mathcal{E}_2)$ then $c(F, \mathcal{E}_2) = 0 \;\Rightarrow\; c(F, \mathcal{E}) = c(F, \mathcal{E}_1)$. Define $F_{(i)} \in \mathcal{F}(F_{(i-1)}, i = 1, 2, \ldots$ and $F_{(0)} = F$. We can apply recursion over $i = 0, 1, 2, \ldots$ until $\mathcal{F}(F_{(i)}, \mathcal{E})$ is empty. Then $d(F_{(i)}, \mathcal{E}) = d(F_{(i)}, \mathcal{E}_1)$.
If $\mathcal{F}(F_{(i)}, \mathcal{E})$ is not empty then since $\mathcal{F}(F_{(i)}, \mathcal{E}) \subset \mathcal{H}(\mathcal{E}_1) \setminus \mathcal{H}(\mathcal{E}_2)$ we can apply Lemma 4.1 and write

$$\begin{aligned}
d(F_{(i)}, \mathcal{E}) &= c(F_{(i)}, \mathcal{E}) - 1 - \sum_{F_{(i+1)} \in \mathcal{F}(F_{(i)}, \mathcal{E})} d(F_{(i+1)}, \mathcal{E}) \\
&= c(F_{(i)}, \mathcal{E}_1) - 1 - \sum_{F_{(i+1)} \in \mathcal{F}(F_{(i)}, \mathcal{E})} d(F_{(i+1)}, \mathcal{E}).
\end{aligned}$$

It follows that $d(F,\mathcal{E}) = d(F,\mathcal{E}_1)$.

(2) $F \in \mathcal{H}(\mathcal{E}_2) \setminus \mathcal{H}(\mathcal{E}_1)$. Similarly to the previous case we get $d(F,\mathcal{E}) = d(F,\mathcal{E}_2)$.

(3a) $F \in \mathcal{H}(\mathcal{E}_2) \cap \mathcal{H}(\mathcal{E}_1)$ and $F = E_1 \cap E_2$

• $\mathcal{F}(F,\mathcal{E}_1) \subseteq \mathcal{F}(\mathcal{E}_1) \subseteq (\bigcup_{E_i \in \mathcal{E}_1} E_i)$ and $\mathcal{F}(F,\mathcal{E}_2) \subseteq \mathcal{F}(\mathcal{E}_2) \subseteq (\bigcup_{E_i \in \mathcal{E}_2} E_i)$ therefore $\mathcal{F}(F,\mathcal{E}_1) \cap \mathcal{F}(F,\mathcal{E}_2) \subseteq E_1 \cap E_2$

• $F \notin \mathcal{F}(F,\mathcal{E})$ therefore $\mathcal{F}(F,\mathcal{E}_1)$ and $\mathcal{F}(F,\mathcal{E}_2)$ are disjoint and $\mathcal{F}(F,\mathcal{E}) = \mathcal{F}(F,\mathcal{E}_1) \cup \mathcal{F}(F,\mathcal{E}_2)$.

$$\begin{aligned} d(F,\mathcal{E}) &= c(F,\mathcal{E}) - 1 - \sum_{F' \in \mathcal{F}(F,\mathcal{E})} d(F',\mathcal{E}) \\ &= c(F',\mathcal{E}_1) + c(F',\mathcal{E}_2) - 1 - \sum_{F' \in \mathcal{F}(F,\mathcal{E}_1)} d(F',\mathcal{E}) - \sum_{F' \in \mathcal{F}(F,\mathcal{E}_2)} d(F',\mathcal{E}) \end{aligned}$$

Since $\mathcal{F}(F,\mathcal{E}_1) \subset \mathcal{H}(\mathcal{E}_1) \setminus \mathcal{H}(\mathcal{E}_2)$ it follows from 1. that $d(F',\mathcal{E}) = d(F',\mathcal{E}_1)$ and similarly for $\mathcal{F}(F,\mathcal{E}_2)$.

$$\begin{aligned} &= c(F,\mathcal{E}_1) + c(F,\mathcal{E}_2) - 1 - \sum_{F' \in \mathcal{F}(F,\mathcal{E}_1)} d(F',\mathcal{E}) - \sum_{F' \in \mathcal{F}(F,\mathcal{E}_2)} d(F',\mathcal{E}) \\ &= d(F,\mathcal{E}_1) + d(F,\mathcal{E}_2) + 1 \end{aligned}$$

(3b) $F \in \mathcal{H}(\mathcal{E}_2) \cap \mathcal{H}(\mathcal{E}_1)$ and $F \subset E_1 \cap E_2$, $F \neq E_1 \cap E_2$

$$d(F,\mathcal{E}) = c(F,\mathcal{E}) - 1 - \sum_{F' \in \mathcal{F}(F,\mathcal{E})} d(F',\mathcal{E})$$

$\mathcal{F}(F,\mathcal{E})$ is the union of the following disjoint classes:

$$\begin{aligned} \mathcal{A} &= \{E_1 \cap E_2\}, \\ \mathcal{B} &= \mathcal{F}(F,\mathcal{E}) \cap (\mathcal{H}(\mathcal{E}_1) \cap \mathcal{H}(\mathcal{E}_2)), \\ \mathcal{C} &= \mathcal{F}(F,\mathcal{E}) \cap (\mathcal{H}(\mathcal{E}_2) \cap \mathcal{H}(\mathcal{E}_1)), \text{ and} \\ \mathcal{D} &= \mathcal{F}(F,\mathcal{E}) \cap \mathcal{H}(\mathcal{E}_1) \cap \mathcal{H}(\mathcal{E}_2) \setminus \{E_1 \cap E_2\}. \end{aligned}$$

Define $F_{(i)} \in \mathcal{F}(F_{(i-1)}\mathcal{E}), i = 1,2,\ldots$ and $F_{(0)} = F$ and $\mathcal{F}(F_{(i)},\mathcal{E}) = \mathcal{A}_{(i)} \cup \mathcal{B}_{(i)} \cup \mathcal{C}_{(i)} \cup \mathcal{D}_{(i)}$. Let $F_{(0)} = F$. We can apply recursion over $i = 0,1,2,...$ until $\mathcal{F}(F_{(i)},\mathcal{E})$ is empty.

• If $\mathcal{D}_{(i)}$ is empty then using equation from part 3. we get

$$\begin{aligned} d(F_{(i)},\mathcal{E}) &= c(F_{(i)},\mathcal{E}_1) + c(F_{(i)},\mathcal{E}_2) - 1 - \sum_{F_{(i+1)} \in \mathcal{A}_{(i)}} (d(F_{(i+1)},\mathcal{E}_1) + d(F_{(i+1)},\mathcal{E}_2) + 1) \\ &= d(F_{(i)},\mathcal{E}_1) + d(F_{(i)},\mathcal{E}_2). \end{aligned}$$

- If $\mathcal{D}_{(i)}$ is not empty then

$$\begin{aligned} d(F_{(i)},\mathcal{E}) &= c(F_{(i)},\mathcal{E}_1)+c(F_{(i)},\mathcal{E}_2)-1 \\ &\quad -\sum_{F_{(i+1)}\in\mathcal{A}_{(i)}} d(F_{(i+1)},\mathcal{E}) - \sum_{F_{(i+1)}\in\mathcal{B}_{(i)}} d(F_{(i+1)},\mathcal{E}) \\ &\quad -\sum_{F_{(i+1)}\in\mathcal{C}_{(i)}} d(F_{(i+1)},\mathcal{E}) - \sum_{F_{(i+1)}\in\mathcal{D}_{(i)}} d(F_{(i+1)},\mathcal{E}) \\ &= c(F_{(i)},\mathcal{E}_1)+c(F_{(i)},\mathcal{E}_2)-1 \\ &\quad -\sum_{F_{(i+1)}\in\mathcal{A}_{(i)}} (d(F_{(i+1)},\mathcal{E}_1)+d(F_{(i+1)},\mathcal{E}_2)+1) - \sum_{F_{(i+1)}\in\mathcal{B}_{(i)}} d(F_{(i+1)},\mathcal{E}_1) \\ &\quad -\sum_{F_{(i+1)}\in\mathcal{C}_{(i)}} d(F_{(i+1)},\mathcal{E}_2) - \sum_{F_{(i+1)}\in\mathcal{D}_{(i)}} d(F_{(i+1)},\mathcal{E}) \end{aligned}$$

Then for $F_{(0)}=F$ we get

$$\begin{aligned} d(F,\mathcal{E}) &= c(F_{(0)},\mathcal{E}_1)+c(F_{(0)},\mathcal{E}_2)-2 \\ &= -\sum_{F_{(1)}\in\mathcal{F}(F_{(0)},\mathcal{E}_1)} d(F_{(1)},\mathcal{E}_1) - \sum_{F_{(1)}\in\mathcal{F}(F_{(0)},\mathcal{E}_2)} d(F_{(1)},\mathcal{E}_2) \\ &= d(F,\mathcal{E}_1)+d(F,\mathcal{E}_2). \end{aligned}$$

The relations for $d(F,\mathcal{E})$ that were already proved are summarized in the following table.

$\mathcal{F}(\mathcal{E})$	$\mathcal{H}(\mathcal{E})$	$d(F,\mathcal{E})$
$F\in\mathcal{F}(\mathcal{E}_1)\setminus\mathcal{F}(\mathcal{E}_2)$	$F\in\mathcal{F}(\mathcal{E}_1)\cap\mathcal{H}(\mathcal{E}_2)$	$d(F,\mathcal{E}_1)$
$F\in\mathcal{F}(\mathcal{E}_1)\setminus\mathcal{F}(\mathcal{E}_2)$	$F\in\mathcal{F}(\mathcal{E})\cap\mathcal{H}(\mathcal{E}_1)\cap\mathcal{H}(\mathcal{E}_2)$	$d(F,\mathcal{E}_1)+d(F,\mathcal{E}_2)$
$F\in\mathcal{F}(\mathcal{E}_1)\cap\mathcal{F}(\mathcal{E}_2)$	$F\in\mathcal{F}(\mathcal{E})\cap\mathcal{H}(\mathcal{E}_1)\cap\mathcal{H}(\mathcal{E}_2)$	$d(F,\mathcal{E}_1)+d(F,\mathcal{E}_2)$
$F\in\mathcal{F}(\mathcal{E}_2)\setminus\mathcal{F}(\mathcal{E}_1)$	$F\in\mathcal{F}(\mathcal{E})\cap\mathcal{H}(\mathcal{E}_1)\cap\mathcal{H}(\mathcal{E}_2)$	$d(F,\mathcal{E}_1)+d(F,\mathcal{E}_2)$
$F\in\mathcal{F}(\mathcal{E}_2)\setminus\mathcal{F}(\mathcal{E}_1)$	$F\in\mathcal{F}(\mathcal{E}_2)\cap\mathcal{H}(\mathcal{E}_1)$	$d(F,\mathcal{E}_2)$

Using Lemma 2.8 and induction hypothesis we can write

$$\begin{aligned} \pi'_{\mathcal{R}_V}P &= \frac{\pi'_{\mathcal{R}_A}P\cdot\pi'_{\mathcal{R}_B}P}{P^C}\cdot U_D \\ &= \frac{\prod_{E\in\mathcal{E}_1}P^E}{\prod_{F\in\mathcal{F}(\mathcal{E}_1)}(P^F)^{d(F,\mathcal{E}_1)}}\cdot\frac{\prod_{E\in\mathcal{E}_2}P^E}{\prod_{F\in\mathcal{F}(\mathcal{E}_2)}(P^F)^{d(F,\mathcal{E}_2)}}\cdot\frac{1}{P^C}\cdot U_D \\ &= \frac{\prod_{E\in\mathcal{E}}P^E\cdot U_D}{\prod_{F\in\mathcal{F}(\mathcal{E}_1)\setminus\mathcal{F}(\mathcal{E}_2)}(P^F)^{d(F,\mathcal{E}_1)}\cdot\prod_{F\in\mathcal{F}(\mathcal{E}_2)\setminus\mathcal{F}(\mathcal{E}_1)}(P^F)^{d(F,\mathcal{E}_2)}}\cdot \end{aligned}$$

$$\frac{1}{\prod_{F\in(\mathcal{F}(\mathcal{E}_1)\cap\mathcal{F}(\mathcal{E}_2))\setminus C}(P^F)^{d(F,\mathcal{E}_1)+d(F,\mathcal{E}_2)}\cdot(P^C)^{d(C,\mathcal{E}_1)+d(C,\mathcal{E}_2)+1}}.$$

Finally, using the relations for $d(F,\mathcal{E})$, summarized in the table, and since $C\in\mathcal{F}(\mathcal{E})\cap\mathcal{H}(\mathcal{E}_1)\cap\mathcal{H}(\mathcal{E}_2)$ we get

$$\pi'_{\mathcal{R}_V}P \quad = \quad \frac{\prod_{E\in\mathcal{E}}P^E}{\prod_{F\in\mathcal{F}(\mathcal{E})}(P^F)^{d(F,\mathcal{E})}}\cdot U_D(x^D).$$

□

So far no simple procedure to test *strong consistency* is known. Application of methods of *linear programming* leads to the task of high dimensionality. In the case of *input set* with *decomposable generating class* the test of *weak consistency* is simple. Furthermore it is in this case equivalent to the *strong consistency*. Properties of *decomposable generating class* of this kind were already proved by several authors. At least, we refer to H. G. Kellerer [25]. The following theorem shows how the the test of *weak consistency* can be performed in the case of *decomposable generating class*.

Theorem 4.3 (Weak consistency test)
Let $\mathcal{P}_\mathcal{E}=\{P_1((X_i)_{i\in E_1}),\dots,P_s((X_i)_{i\in E_s})\}$ *be an input set of distributions and* E_k *for* $k=1,\dots,s$ *be an ordering of sets from* $\mathcal{E}$ *meeting the* RIP*, then the following two statements are equivalent:*

(a) $\mathcal{P}_\mathcal{E}$ *is weakly consistent*

(b) $\forall k=2,\dots,s \quad \exists\ell \quad (1\le\ell<k \quad \& \quad (E_k\cap\bigcup_{m=1}^{k-1}E_m)\subset E_\ell \quad \& \quad P_k^{E_k\cap E_\ell}=P_\ell^{E_k\cap E_\ell})$.

Proof. **(a)** $\Rightarrow$ **(b)** If the *input set* $\mathcal{P}_\mathcal{E}$ is weakly consistent then, according to definition, for any couple of different indices

$$k,\ell\in\{1,\dots,s\} \quad (P_k^{E_k\cap E_\ell}=P_\ell^{E_k\cap E_\ell}).$$

therefore this equality must be valid also for a subset of all couples defined by (b) (existence of ℓ with given properties is assured because of RIP).

(b) $\Rightarrow$ **(a)** Condition of weak consistency of *input set* $\mathcal{P}_\mathcal{E}$ can be rewritten :

$$\forall k=2,\dots,s \quad \forall j<k \quad (P_k^{E_k\cap E_j}=P_j^{E_k\cap E_j}). \tag{4.1}$$

For $k=2$ (b) and 4.1 coincide, as both these expressions can be written as

$$P_2^{E_2\cap E_1}=P_1^{E_2\cap E_1}.$$

Assume the theorem holds for $k<i$. Now, we shall prove it holds also for $k=i$, i.e. we are proving that

$$\forall j<i \quad P_i^{E_i\cap E_j}=P_j^{E_i\cap E_j}.$$

Statement (b) requires existence of $\ell<i$ such that $E_j\cap E_i\subseteq E_\ell$. P_j and P_ℓ are pairwise consistent because of the assumption of the mathematical induction (we assume the theorem is valid for all k ¡ i), pairwise consistency of P_ℓ and P_i is required by (b). Therefore

$$P_j^{E_j\cap E_\ell}=P_\ell^{E_j\cap E_\ell}, \quad P_\ell^{E_\ell\cap E_i}=P_i^{E_\ell\cap E_i}, \quad \text{and} \quad E_j\cap E_i\subseteq E_\ell$$

gives

$$P_j^{E_j \cap E_i} = P_i^{E_j \cap E_i}.$$

□

4.1 Convergence within one cycle

In Chapter 1 the mapping that associates to each conformal hypergraph its graph was described. Therefore, the study of *decomposable generating classes* that guarantee IPFP convergence within one cycle can exploits known facts about *chordal (triangulated) graphs.*

IPFP convergence rate depends on the choice of initial distribution, but we will not study this question. We are going to deal with initial distributions that *factorize* with respect to *generating class* and the uniform distribution only (that naturally factorize with respect to *generating class*), since they approve the best convergence rate of IPFP.

The fact that an explicit formula of $\pi'_{\mathcal{R}_V} P$ for orderings that meet Running Intersection Property is given in a simple form (Theorem 4.1) is used to prove the following theorem. The theorem together with its proof can be found in [17, Theorem 5.3].

Theorem 4.4 (Orderings that assure convergence within one cycle)
Let $\mathcal{E}$ be a decomposable generating class of input set $\mathcal{P}_\mathcal{E}$ and $P \in \mathcal{S}_\mathcal{E}, P \ll Q_{(0)}$. If $Q_{(0)} \in \mathcal{R}_\mathcal{E}$ then there exist an ordering $\mathcal{E} = \{E_i : i = 1, 2, \ldots, s\}$ such that

$$Q_{(s)} = Q_{(s+1)} = \ldots,$$

where $Q_{(i)}, i = 1, 2, \ldots$ are computed by formula 3.1 of Definition 3.2. We will say that IPFP converges within one cycle (s steps).

All the orderings that meet Running Intersection Property guarantee convergence in one cycle. But there are some other that can assure the convergence within one cycle, too. Theorem 4.5 proved by S. Haberman [17, Theorem 5.4] gives a sufficient (but not necessary) condition for a *family of generating classes* that guarantee IPFP convergence within one cycle no matter how is the *input set* ordered. The proof of Theorem 4.5 makes use of Lemma 4.2. This lemma gives an explicit formula for marginal distribution of the I_2-projection on a restriction of a *generating class* $\mathcal{E}$. Lemma 4.2 was also proved by S. Haberman [17, Lemma 5.14]. We are going to cite both these assertions since they constitute an important part of study of IPFP behavior for *decomposable generating classes.*

Lemma 4.2 *Let $\mathcal{E} = \{E_1, E_2, \ldots, E_s\}, s \geq 2$ be a decomposable generating class of input set $\mathcal{P}_\mathcal{E}$. Further, let $\mathcal{A}$ be the set of edges generated from $\mathcal{E}$ by set $A = \cup_{i=2}^{s} E_i$, i.e. $\mathcal{A} = \mathcal{E}(A)$ and $\mathcal{B}$ be the set of edges generated from $\mathcal{E} \setminus E_1$ by set E_1, i.e. $\mathcal{B} = (\mathcal{E} \setminus E_1)(E_1)$. If $\pi'_{\mathcal{R}_\mathcal{B}} P$ exists then*

$$(\pi'_{\mathcal{R}_{\mathcal{A},V}} P)^{E_1} = \pi'_{\mathcal{R}_\mathcal{B}} P \cdot \frac{1}{U_D}, \quad \text{where} \quad D = V \setminus \bigcup_{i=1,2,\ldots s} E_i.$$

Theorem 4.5 (Sufficient condition for IPFP convergence within one cycle - 1)
Let $\mathcal{E}$ be a totally decomposable generating class and $P \in \mathcal{P}$ be a probability distribution from $\mathcal{S}_\mathcal{E}$. Then IPFP starting with the uniform distribution $Q_{(0)} = U_V \in \mathcal{P}$ converges within one cycle (s steps) no matter how $\mathcal{E}$ is ordered and $Q_{(s)} = \pi'_{\mathcal{R}_V} P$.

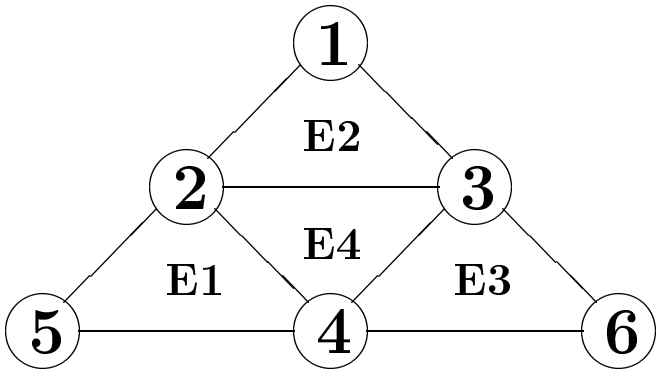


Figure 4.2: An example of generating class, which does not satisfy assumptions of Theorem 4.5. In spite of it IPFP converges in one cycle no matter how $\mathcal{E}$ is ordered (see Theorem 4.6).

We will exploit formulas for marginals of $Q_{(i)}$ given in Theorem 3.1 to prove another sufficient condition of IPFP convergence within s steps.

Theorem 4.6 (Sufficient condition for IPFP convergence within one cycle - 2) *Suppose there is a set* $E^* \in \mathcal{E} = \{E_i : i = 1, 2, \ldots, s\}$ *such that* $E_i \cap E_j \subset E^*, \forall i, j = 1, 2, \ldots s, i \neq j$. *If* $\mathcal{E}$ *is a generating class of input set* $\mathcal{P}_\mathcal{E}$ *and if there exist* $P \in \mathcal{S}_\mathcal{E}, P \ll Q_{(0)}$ *then IPFP starting with* $Q_{(0)} \in \mathcal{R}_\mathcal{F}$ *converges within* s *steps (one cycle) no matter how* $\mathcal{E}$ *is ordered.*

Proof. Denote r the index for which $E^* = E_r, 1 \leq r \leq s$. $\mathcal{E}$ is decomposable since its ordering starting with $E_{\delta(1)} = E_r$ satisfies RIP (see Lemma 1.1).

In the sequel i denotes iterative step, $j = (i-1) \bmod s + 1$ is index of P_{E_j} being fitted in step i. At first, we will simplify the formula of Theorem 3.1 separately for the cases $E_j \neq E_r$ and $E_j = E_r$.

(i) $E_j \neq E_r$

For any $E_j \in \mathcal{E}, E_j \neq E_r$ any ordering such that

$$\mathcal{E} = \{E_{\delta(1)}, E_{\delta(2)}, \ldots, E_{\delta(s)}\}, \text{ where } E_j = E_{\delta(1)} \text{ and } E_r = E_{\delta(2)} \text{ satisfies RIP.}$$

Let $A_j = \cup_{l=1}^{j-1} E_{\delta(l)}$. For $k = r$:

$$E_k \cap A_k = E_j \cap E_r$$

and for every $1 \leq k \leq s, k \neq j, k \neq r$:

$$E_k \cap A_k = E_k \cap E_r.$$

In this case the formula of Theorem 3.1 can be rewritten:

$$Q_{(i)}^{E_k} = \begin{cases} P_{E_j} & \text{for } k = j, \\ Q_{(i)}^{E_j \cap E_r} \dfrac{Q_{(i-1)}^{E_r}}{Q_{(i-1)}^{E_j \cap E_r}} & \text{for } k = r, \\ Q_{(i)}^{E_k \cap E_r} \dfrac{Q_{(i-1)}^{E_k}}{Q_{(i-1)}^{E_k \cap E_r}}, & \text{for } k \neq j, k \neq r, \end{cases}$$

From this formula it follows that $Q_{(i)}^{E_j \cap E_r} = P_{E_j}^{E_j \cap E_r}$ and consequently we can further rewrite the formula for $k = r$:

$$Q_{(i)}^{E_k} = P_{E_j}^{E_j \cap E_r} \frac{Q_{(i-1)}^{E_r}}{Q_{(i-1)}^{E_j \cap E_r}}. \tag{4.2}$$

By marginalization of 4.2 to $E_k \cap E_r$ we get

$$Q_{(i)}^{E_k \cap E_r} = P_{E_j}^{E_j \cap E_r \cap E_k} \frac{Q_{(i-1)}^{E_k \cap E_r}}{Q_{(i-1)}^{E_j \cap E_r \cap E_k}} = P_{E_j}^{E_k \cap E_r} \frac{Q_{(i-1)}^{E_k \cap E_r}}{Q_{(i-1)}^{E_k \cap E_r}} = P_{E_j}^{E_k \cap E_r},$$

so that we can finally rewrite the formula of Theorem 3.1

$$Q_{(i)}^{E_k} = \begin{cases} P_{E_j} & \text{for } k = j, \\ P_{E_j}^{E_j \cap E_r} \dfrac{Q_{(i-1)}^{E_r}}{Q_{(i-1)}^{E_j \cap E_r}} & \text{for } k = r, \\ P_{E_j}^{E_k \cap E_r} \dfrac{Q_{(i-1)}^{E_k}}{Q_{(i-1)}^{E_k \cap E_r}}, & \text{for } k \neq j, k \neq r, \end{cases} \tag{4.3}$$

(ii) $E_j = E_r$

All orderings

$$\mathcal{E} = \{E_{\delta(1)}, E_{\delta(2)}, \ldots, E_{\delta(s)}\}, \text{ such that } E_r = E_{\delta(1)},$$

meets RIP. Let $A_j = \cup_{l=1}^{j-1} E_{\delta(l)}$. For every $1 \leq k \leq s, k \neq r$:

$$E_k \cap A_k = E_k \cap E_r.$$

In this case the formula of Theorem 3.1 can be given in the form of:

$$Q_{(i)}^{E_k} = \begin{cases} P_{E_r} & \text{for } k = r, \\ Q_{(i)}^{E_k \cap E_r} \dfrac{Q_{(i-1)}^{E_k}}{Q_{(i-1)}^{E_k \cap E_r}} & \text{for } k \neq r. \end{cases}$$

Similarly to **(i)** this formula can be rewritten so that we get

$$Q_{(i)}^{E_k} = \begin{cases} P_{E_r} & \text{for } k = r, \\ P_{E_r}^{E_k \cap E_r} \dfrac{Q_{(i-1)}^{E_r}}{Q_{(i-1)}^{E_k \cap E_r}} & \text{for } k \neq r. \end{cases} \tag{4.4}$$

In the next, we will exploit the fact that for every $i = 2, 3, \ldots$

$$Q_{(i-1)}^{E_k} = P_{E_k} \quad \text{for} \quad k = i - 1.$$

Therefore it follows from equations 4.3 and 4.4 for $k = i - 1$

$$Q_{(i)}^{E_k} = \begin{cases} P_{E_j}^{E_k \cap E_r} \dfrac{P_{E_k}}{P_{E_k}^{E_k \cap E_r}} & \text{for } k \neq r, k \neq j, \\ P_{E_j}^{E_j \cap E_r} \dfrac{P_{E_r}}{P_{E_r}^{E_j \cap E_r}} & \text{for } k = r, k \neq j. \end{cases}$$

It is supposed that $\mathcal{S}_\mathcal{E} \neq \emptyset$ the *input set* is *consistent* which implies that

$$P_{E_j}^{E_k \cap E_r} = P_{E_k}^{E_k \cap E_r} \quad \text{and} \quad P_{E_j}^{E_j \cap E_r} = P_{E_r}^{E_j \cap E_r}.$$

Thus we get for $k = i - 1$

$$Q_{(i)}^{E_k} = P_{E_k}. \tag{4.5}$$

This procedure can be repeated also for $k = i - 2, i - 3, \ldots, 1$, so that equation 4.5 holds for $k = 1, 2, \ldots i$. Since i is arbitrary we can write for $i = s, k = 1, 2, \ldots s$

$$Q_{(s)}^{E_k} = P_{E_k}.$$

Since $Q_{(i)} \in \mathcal{R}_\mathcal{E}, i = 0, 1, 2, \ldots$ and $\mathcal{E}$ is decomposable, for any ordering

$$\mathcal{E} = \{E_{\delta(1)}, E_{\delta(2)}, \ldots, E_{\delta(s)}\},$$

that satisfies RIP we can write (using Lemma 2.8 and Theorem 4.1)

$$Q_{(s)} = \frac{\prod_{j=\delta(1)}^{\delta(s)} Q_{(s)}^{E_j}}{\prod_{j=\delta(2)}^{\delta(s)} Q_{(s)}^{E_j \cap A_j}} = \frac{\prod_{j=\delta(1)}^{\delta(s)} P_{E_j}}{\prod_{j=\delta(2)}^{\delta(s)} P_{E_j}^{E_j \cap A_j}} = \pi'_{\mathcal{R}_\mathcal{E}} P = \pi_{\mathcal{S}_\mathcal{E}} Q_{(0)},$$

which is the limit of joint probability distribution's sequence computed by IPFP. □

Conjecture 4.1 (Necessary and sufficient condition)
IPFP converges within one cycle (s steps) no matter how $\mathcal{E}$ is ordered and what are actual values of distribution in strongly consistent input set **iff** *either assumptions of Theorem 4.5 or Theorem 4.6 are satisfied.*

The argument in favor of the previous conjecture is based on the following consideration. Exploiting the correspondence between hypergraph and graphs notion so we can formulate the conjecture using *strong chordality* and *chordality* graph property (see Section 1.2.3). We could reject the previous conjecture if we had constructed a **counter example**. So we try to construct a graph that

(1) is chordal,

(2) is not strongly chordal,

(3) does not contain any clique E^* such that all pair-wise intersection of cliques are subsets of the clique E^*,

(4) IPFP converges within one cycle for all orderings of cliques.

We believe that all graphs satisfying (1) and (2) must contain a cycle of the length equal six without *strong chord*. It is because all *chordal graphs* that contain a cycle with even length greater than six without strong chord seems to contain the cycle of the length six as well. Consequently, every construction of a counter example must start with a cycle of length six having three chords (to assure (1)) but none of them being strong (Figure 4.2). Note that such a graph satisfies condition (4). Let us check all ways how this graph could be extended so that it remains *chordal* but not *strongly chordal* and does not contain any clique E^* such that all pair-wise intersection of cliques are subsets of the clique E^*. There are six pairs of nodes, having distance one in the cycle of the length six, where can be the graph extended so that conditions discussed above are satisfied. Using Figure 4.2 these pairs are:

$$\{1,3\},\{3,6\},\{6,4\},\{4,5\},\{5,2\},\{2,1\}.$$

Extension are constructed so that a new clique's intersection with the initial graph is equal to one of the pairs. Then all new graphs satisfies conditions (1), (2), and (3) but for all of them an ordering of their cliques can be found so that IPFP does not converge within one cycle.

The following theorem gives a sufficient condition for orderings that guarantee IPFP convergence within one cycle. The idea of its proof is due to Radim Jiroušek.

Theorem 4.7 (Orderings that guarantee convergence within one cycle)
Let $\{E_1, E_2, \dots E_s\}$ *be an ordering of generating class* $\mathcal{E}$*, initial probability distribution* $Q_{(0)} \in \mathcal{P}$ *be uniform and there exist a* $Q \in \mathcal{S}_{\mathcal{E}} = \{Q(x) \in \mathcal{P} : Q^{E_i} = P_{E_i}, i = 1, 2, \dots, s\}$*. If for every iteration* $1 < i \leq s$ *hypergraph*

$$H_{(i)} = (V_{(i)}, \mathcal{E}_{(i)}), \quad V_{(i)} = \cup_{j=1}^{i} E_j, \quad \mathcal{E}_{(i)} = \{E_1, E_2, \dots E_i\}$$

is decomposable then IPFP converges in one cycle.

Proof.
Using mathematical induction we will show that for all $i = 1, 2, \dots, s$ the distributions

$$Q_{(i)} \in \mathcal{S}_{\mathcal{E}_{(i)}}.$$

It is obvious that $Q_{(1)}^{E_1} = P_{E_1}$. Thus $Q_{(1)} \in \mathcal{S}_{\mathcal{E}_{(1)}}$.

Assume $Q_{(i-1)} \in \mathcal{S}_{\mathcal{E}_{(i-1)}}$. From Theorem 3.3 we know that $Q_{(i-1)} \in \mathcal{R}_{\mathcal{E}_{(i-1)}}$ therefore the computation of $Q_{(i)}$ can be performed in i "sub-steps" by the formula of Theorem 3.1. In this case we take any ordering $E_{\delta(1)}, E_{\delta(2)}, \dots E_{\delta(i)}$ of $\mathcal{E}_{(i)}$ meeting RIP such that $\delta(1) = i$. We will use mathematical induction again to prove that for

$$k = 1, 2, \dots i: \quad Q_{(i)}^{E_{\delta(k)}} = P_{E_\delta(k)}.$$

For $k = 1$

$$Q_{(i)}^{E_{\delta(1)}} = Q_{(i)}^{E_i} \quad = \quad P_{E_i} \frac{Q_{(i)}^{E_i}}{Q_{(i)}^{E_i}} = P_{E_i}. \tag{4.6}$$

Suppose for $n = 1, 2, \dots, m-1$, where $2 \leq m \leq i$, it holds that

$$Q_{(i)}^{E_{\delta(n)}} \quad = \quad P_{E_{\delta(n)}}. \tag{4.7}$$

Then it is assured by RIP that there exist l such that $1 \leq l < m$ and

$$E_{\delta(m)} \cap A_m \quad = \quad E_{\delta(m)} \cap E_{\delta(l)}, \quad \text{where} \quad A_m = \bigcup_{j=1,2,\dots m-1} E_{\delta(j)}.$$

Thus we can write using the formula of Theorem 3.1

$$Q_{(i)}^{E_{\delta(m)}} \quad = \quad Q_{(i)}^{E_{\delta(m)} \cap E_{\delta(l)}} \frac{Q_{(i-1)}^{E_{\delta(m)}}}{Q_{(i-1)}^{E_{\delta(m)} \cap E_{\delta(l)}}} \tag{4.8}$$

Since it was assumed that $Q_{(i-1)} \in \mathcal{S}_{\mathcal{E}_{(i-1)}}$ we get

$$Q_{(i)}^{E_{\delta(m)}} \quad = \quad Q_{(i)}^{E_{\delta(m)} \cap E_{\delta(l)}} \frac{P_{E_{\delta(m)}}}{P_{E_{\delta(m)}}^{E_{\delta(m)} \cap E_{\delta(l)}}}.$$

It follows from the induction hypothesis (equation 4.7) that $Q_{(i)}^{E_{\delta(l)}} = P_{E_{\delta(l)}}$. Therefore, we can further rewrite formula 4.8:

$$Q_{(i)}^{E_{\delta(m)}} \quad = \quad P_{E_{\delta(l)}}^{E_{\delta(m)} \cap E_{\delta(l)}} \frac{P_{E_{\delta(m)}}}{P_{E_{\delta(m)}}^{E_{\delta(m)} \cap E_{\delta(l)}}}.$$

Since the *input set* is *consistent* it holds that $P_{E_{\delta(m)}}^{E_{\delta(m)} \cap E_{\delta(l)}} = P_{E_{\delta(l)}}^{E_{\delta(m)} \cap E_{\delta(l)}}$ and finally, we get

$$Q_{(i)}^{E_{\delta(m)}} \quad = \quad P_{E_{\delta(m)}}^{E_{\delta(m)} \cap E_{\delta(l)}} \frac{P_{E_{\delta(m)}}}{P_{E_{\delta(m)}}^{E_{\delta(m)} \cap E_{\delta(l)}}} = P_{E_{\delta(m)}}.$$

□

However, Theorem 4.7 does not give a necessary condition since there exist orderings that does not satisfy the assumption of the theorem but the IPFP procedure converges within one cycle. An example of such ordering is the ordering $\{E_1, E_3, E_4, E_6, E_5, E_2\}$ of the *input set* whose *generating class* is described by Figure 4.3.

4.2 Convergence within more than one cycle

S. Haberman in [17] conjectures that if generating class $\mathcal{E} = \{E_i : i = 1, 2, \ldots, s\}$ is decomposable and $P_1, P_2, \ldots, P_s$ are consistent, then IPFP converges within $2 \cdot s$ steps (two cycles) no matter how $\mathcal{E}$ is ordered. **The conjecture does not hold.**

Assertion 4.1 *There exist decomposable generating classes $\mathcal{E} = \{E_i : i = 1, 2, \ldots, s\}$ such that for some of their orderings IPFP does not converge within $2 \cdot s$ steps (two cycles) even if input set $\{P_1, P_2, \ldots, P_s\}$ is consistent and the initial distribution $Q_{(0)}$ is uniform.*

The following example is a counterexample to the Haberman's conjecture which proves Assertion 4.1.

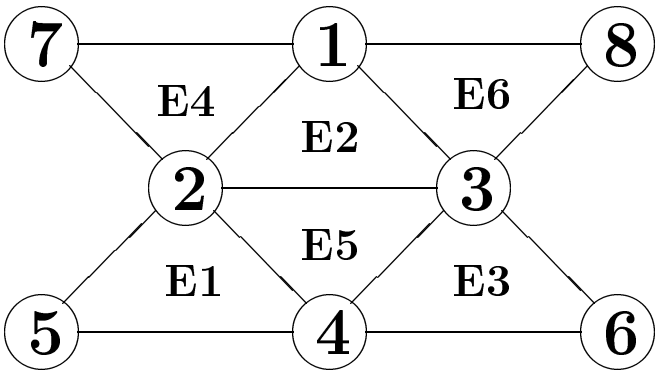


Figure 4.3: A decomposable generating class (conjectured to be the smallest) for which IPFP need not converge in two cycles for the ordering $\{E_1, E_2, E_3, E_4, E_5, E_6\}$

Table 4.1: Distribution P_{E_1}

	$X_4 = 1$		$X_4 = 2$	
P_{E_1}	$X_5 = 1$	$X_5 = 2$	$X_5 = 1$	$X_5 = 2$
$X_2 = 1$	5/100	8/100	5/100	11/100
$X_2 = 2$	13/100	17/100	31/100	10/100

The counter example was verified using implementation of IPFP written in **rational arithmetic** in Mathematica [43], an integrated system for technical computation, with extensive numerical, symbolic, graphical, programming, and interfacing capabilities. More details describing the implementation can be found in Appendix A, the source file is available from *World Wide Web* and attached to the thesis as Appendix A.1.

In Tables 4.1, 4.2, 4.3, 4.4, 4.5, and 4.6 probability distributions of a *consistent input set* are given. Initial probability distribution $Q_{(0)} \in \mathcal{P}$ was uniform. After *two cycles* the *total variances* between marginals of joint probability distribution $Q_{(12)}^{E_j}, j = 1, 2, 3, 4, 5$ and

Table 4.2: Distribution P_{E_2}

	$X_2 = 1$		$X_2 = 2$	
P_{E_2}	$X_3 = 1$	$X_3 = 2$	$X_3 = 1$	$X_3 = 2$
$X_1 = 1$	3/100	13/100	7/100	17/100
$X_1 = 2$	3/100	10/100	17/100	30/100

Table 4.3: Distribution P_{E_3}

	$X_4 = 1$		$X_4 = 2$	
P_{E_3}	$X_6 = 1$	$X_6 = 2$	$X_6 = 1$	$X_6 = 2$
$X_3 = 1$	7/100	5/100	11/100	7/100
$X_3 = 2$	23/100	8/100	27/100	12/100

Table 4.4: Distribution P_{E_4}

	$X_2 = 1$		$X_2 = 2$	
P_{E_4}	$X_7 = 1$	$X_7 = 2$	$X_7 = 1$	$X_7 = 2$
$X_1 = 1$	12/100	4/100	11/100	13/100
$X_1 = 2$	8/100	5/100	13/100	34/100

Table 4.5: Distribution P_{E_5}

	$X_3 = 1$		$X_3 = 2$	
P_{E_5}	$X_4 = 1$	$X_4 = 2$	$X_4 = 1$	$X_4 = 2$
$X_2 = 1$	2/100	4/100	11/100	12/100
$X_2 = 2$	10/100	14/100	20/100	27/100

Table 4.6: Distribution P_{E_6}

	$X_3 = 1$		$X_3 = 2$	
P_{E_6}	$X_8 = 1$	$X_8 = 2$	$X_8 = 1$	$X_8 = 2$
$X_1 = 1$	3/100	7/100	11/100	19/100
$X_1 = 2$	5/100	15/100	25/100	15/100

distributions from *input set* were different from zero. It means that IPFP does not converge within $2 \cdot s$ steps (two cycles). This contradicts the Haberman's conjecture. Rounded values of the *total variances* are given in Table 4.7.

j	$\lvert P_{E_j} - Q^{E_j}_{(12)} \rvert$
1	$2.5496 \cdot 10^{-12}$
2	$1.0533 \cdot 10^{-10}$
3	$1.8018 \cdot 10^{-13}$
4	$1.0533 \cdot 10^{-14}$
5	$2.5496 \cdot 10^{-12}$
6	0

Table 4.7: The total variances between marginals of $Q_{(12)}$ and the distributions from the input set

We conjecture that convergence of IPFP for decomposable *generating classes* within *finite* number of iterative steps can not be guaranteed if the structure of *generating class* corresponds to a *graph* containing as its subgraph graph from Figure 4.3.

Conjecture 4.2 *Let $\mathcal{E}$ be a decomposable generating class of input set $\mathcal{P}_{\mathcal{E}}$, and there exist a $P \in \mathcal{S}_{\mathcal{E}}, P \ll Q_{(0)}$. If IPFP starting with a joint probability distribution $Q_{(0)} \in \mathcal{R}_{\mathcal{E}}, j = 1, 2, \ldots s$ does not converge within finite number of iterative steps then the graph of hypergraph $H = (V, \mathcal{E})$ contains a cycle such that*

(1) its length is greater or equal eight,

(2) does not contain any strong chord, and

(3) there are no crossing chords in the cycle.

The idea behind the previous conjecture is based on the following considerations. If there is a *chordal graph* that does not contain any cycle such that its length is greater or equal eight (1) then we believe it must be a *strongly chordal* graph or a graph of the type given by Figure 4.2 (see the arguments in favor of Conjecture 4.1). For such graphs IPFP always converges within one cycle. The same holds for graphs containing a strong chord (2). The conditions (1) and (2) are not sufficient since all the graphs that satisfies them but contains a clique E^* such that all pair-wise intersection of cliques are subsets of the clique E^* must be excluded. his is the reason for the condition (3) to be involved. See Figure 1.5 for an example of graph that satisfies (1) and (2) but not (3).

5 Inconsistent knowledge integration

In practice, quite often we have to cope with incomplete or inconsistent information and make decisions on it. Intelligent systems should be capable of handling such information as well. Within probabilistic framework, the knowledge of a given domain can be represented by a joint probability distribution that embodies knowledge about a chosen domain. How can we construct such a distribution? What are possible solutions? What are their advantages and shortages? At the beginning of this chapter, we are going to present two basic approaches to *knowledge integration* having an *inconsistent input set* of probability distributions.

First approach exploits given criteria that **measure distances** from the marginals of a probability distribution to the probability distributions in a given *input set* (Section 5.1). Within this approach the probability distribution that minimizes that criteria is regarded to be the best solution of *inconsistent knowledge integration* task.

Second approach is based on the assumption that probability distributions from a given *input set* might be estimated as relative frequencies from data (Section 5.2). If the data are *incomplete* then the estimation may results into an *inconsistent input set* of probability distributions. In this case there are several methods in statistics that could be applied to the problem. They are usually referred as **missing data methods**. Later, we will show that these two approaches may, for a certain distance criterion, overlap and methods proposed within one approach can get a reasonable interpretation within the other.

Five iterative methods will be discussed in detail and the results of their empirical comparison will be presented. These methods are:

- Convex Conservative Modification of IPFP (CC) in Section 5.7,
- Log-Convex Conservative Modification of IPFP (LCC) in Section 5.7,
- Method of Iterative Arithmetic Averages (AA) in Section 5.8,

- Method of Iterative Geometric Averages (GA) in Section 5.9, and
- an instance of Generalized EM-algorithm (GEM) in Section 5.10.

5.1 Distance to a given input set

In this section, two reasonable criteria that can measure the distance between the marginals of the resulting distribution and the distributions of the *input set* will be presented. Suppose that to every probability distribution P_{E_j} from *input set* $\mathcal{P}_\mathcal{E}$ a nonnegative weight w_j is assigned such that $\sum_{j=1}^{s} w_j = 1$. The weights can be understood as a measure of belief of the corresponding probability distributions. Alternatively, the weights can be set to equal the ratio of number of data vectors used to estimation of values of the corresponding probability distribution N_{E_j} and total number of data vectors N, i.e. $w_j = \frac{N_{E_j}}{N}$.

Definition 5.1 (I_1-aggregate) Let $\mathcal{E} = \{E_1, \dots E_s\}$ be a *generating class* of an *input set* $\mathcal{P}_\mathcal{E}$, $Q \in \mathcal{P}$ be a probability distribution and w_j be nonnegative weights summing up to 1. Then an I_1-*aggregate* is defined to be the functional

$$\psi_1(Q) = \sum_{j=1}^{s} w_j \cdot I(P_{E_j} \parallel Q^{E_j}).$$

The set of all probability distributions minimizing the I_1-*aggregate* will be denoted $\mathcal{T}_1$, i.e.

$$\mathcal{T}_1 = \{Q \in \mathcal{P} : \ (\forall P \in \mathcal{P} : \psi_1(P) \geq \psi_1(Q)\}.$$

The following lemma shows that the I_1-aggregate is equivalent to $\sum_{j=1}^{s} w_j \cdot I(\pi_{\mathcal{S}_{\{E_j\}}} Q \parallel Q)$.

Lemma 5.1 *Let $P_{E_j} \in \mathcal{P}_{E_j}, j = 1, 2, \dots, s$, $\mathcal{S}_{\{E_j\}} = \{S \in \mathcal{P}, S^{E_j} = P_{E_j}\}$, $Q \in \mathcal{P}$, such that $P_{E_j} \ll Q^{E_j}$, and w_j be nonnegative weights summing up to 1. Then*

$$\sum_{j=1}^{s} w_j \cdot I(\pi_{\mathcal{S}_{\{E_j\}}} Q \parallel Q) = \sum_{j=1}^{s} w_j \cdot I(P_{E_j} \parallel Q^{E_j}).$$

Proof.

$$\begin{aligned}
&\sum_{j=1}^{s} w_j \cdot I(\pi_{\mathcal{S}_{\{E_j\}}} Q \parallel Q) = \\
&= \sum_{j=1}^{s} w_j \cdot \sum_{x,\ Q(x)>0} Q(x) \frac{P_{E_j}(x^{E_j})}{Q^{E_j}(x^{E_j})} \cdot \log \frac{Q(x) \frac{P_{E_j}(x^{E_j})}{Q^{E_j}(x^{E_j})}}{Q(x)} \\
&= \sum_{j=1}^{s} w_j \cdot \sum_{x,\ Q(x)>0} Q(x) \frac{P_{E_j}(x^{E_j})}{Q^{E_j}(x^{E_j})} \cdot \log \frac{P_{E_j}(x^{E_j})}{Q^{E_j}(x^{E_j})}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{j=1}^{s} w_j \cdot \sum_{x^{E_j},\ Q^{E_j}(x^{E_j})>0} \frac{P_{E_j}(x^{E_j})}{Q^{E_j}(x^{E_j})} \cdot \log \frac{P_{E_j}(x^{E_j})}{Q^{E_j}(x^{E_j})} \sum_{y^{V\setminus E_j}=x^{V\setminus E_j}} Q(y) \\
&= \sum_{j=1}^{s} w_j \cdot \sum_{x^{E_j},\ Q^{E_j}(x^{E_j})>0} P_{E_j}(x^{E_j}) \cdot \log \frac{P_{E_j}(x^{E_j})}{Q^{E_j}(x^{E_j})} \\
&= \sum_{j=1}^{s} w_j \cdot I(P_{E_j} \parallel Q^{E_j}).
\end{aligned}
$$

□

Since $I(P \parallel Q) \geq 0$ and $I(P \parallel Q) = 0$ if and only if $P = Q$ it can be easily seen that minimization of I_1-*aggregate* in *consistent case* ensure that a resulting distribution belongs to $\mathcal{S}_\mathcal{E}$. Similarly to the I_1-*aggregate*, we can define the I_2-*aggregate* and set of its minimizers $\mathcal{T}_2$.

Definition 5.2 (I_2-aggregate) Let $\mathcal{E} = \{E_1, \dots E_s\}$ be a *generating class* of an *input set* $\mathcal{P}_\mathcal{E}$, $Q \in \mathcal{P}$ be a probability distribution and w_j be nonnegative weights summing up to 1. Then an I_2-*aggregate* is defined by

$$\psi_2(Q) = \sum_{j=1}^{s} w_j \cdot I(Q^{E_j} \parallel P_{E_j}).$$

The set of all probability distributions minimizing the I_2-*aggregate* will be denoted $\mathcal{T}_2$, i.e.

$$\mathcal{T}_2 = \{Q \in \mathcal{P} : \ (\forall P \in \mathcal{P} : \psi_2(P) \geq \psi_2(Q)\}.$$

Next, we will show that I_2-*aggregate* is equivalent to $\sum_{j=1}^{s} w_j \cdot I(Q \parallel \pi_{\mathcal{S}_{\{E_j\}}} Q)$.

Lemma 5.2 *Let $P_{E_j} \in \mathcal{P}_{E_j}, j = 1, 2, \dots, s$, $\mathcal{S}_{\{E_j\}} = \{S \in \mathcal{P}, S^{E_j} = P_{E_j}\}$, $Q \in \mathcal{P}$ such that $Q^{E_j} \ll P_{E_j}$ and w_j be nonnegative weights summing up to 1. Then*

$$\sum_{j=1}^{s} w_j \cdot I(Q \parallel \pi_{\mathcal{S}_{\{E_j\}}} Q) = \sum_{j=1}^{s} w_j \cdot I(Q^{E_j} \parallel P_{E_j}).$$

Proof.

$$
\begin{aligned}
&\sum_{j=1}^{s} w_j \cdot I(Q \parallel \pi'_{\mathcal{S}_{\{E_j\}}} Q) = \\
&\quad = \sum_{j=1}^{s} w_j \cdot \sum_{x,\ Q(x)>0} Q(x) \cdot \log \frac{Q(x)}{Q\frac{P_{E_j}(x^{E_j})}{Q^{E_j}(x^{E_j})}} \\
&\quad = \sum_{j=1}^{s} w_j \cdot \sum_{x,\ Q(x)>0} Q(x) \cdot \log \frac{Q^{E_j}(x^{E_j})}{P_{E_j}(x^{E_j})} \\
&\quad = \sum_{j=1}^{s} w_j \cdot \sum_{x^{E_j},\ P_{E_j}(x^{E_j})>0} \log \frac{Q^{E_j}(x^{E_j})}{P_{E_j}(x^{E_j})} \sum_{y^{V\setminus E_j}=x^{V\setminus E_j}} Q(y)
\end{aligned}
$$

$$= \sum_{j=1}^{s} w_j \cdot \sum_{x^{E_j},\, P_{E_j}(x^{E_j})>0} Q^{E_j}(x^{E_j}) \cdot \log \frac{Q^{E_j}(x^{E_j})}{P_{E_j}(x^{E_j})}$$

$$= \sum_{j=1}^{s} w_j \cdot I(Q^{E_j} \parallel P_{E_j})$$

□

Either of the criteria need not be minimized by a unique probability distribution. If there exists a probability distribution Q that minimize I_1-*aggregate* or I_2-*aggregate* then all probability distributions having the same marginals $Q^E, E \in \mathcal{E}$ minimize the corresponding *aggregate*, as well. Thus, the set of all probability distributions minimizing the *aggregate* is either an *additively constrained set* or a union of *additively constrained sets*. All experiments we have performed with the proposed methods and minimization of ψ_1 using an optimization algorithm always gave unique marginals $Q^E, E \in \mathcal{E}$ so that we lay down the following conjecture.

Conjecture 5.1 (Set of ψ_1 minimizers is *additively constrained*)
Let $\mathcal{E} = \{E_1, \ldots E_s\}$ be a generating class *of an* input set *$\mathcal{P}_\mathcal{E}$ and w_j be nonnegative weights summing up to 1. Then for all probability distribution $Q_1, Q_2 \in \mathcal{P}$ that minimize* I_1-aggregate *it holds that*

$$Q_1^{E_j} = Q_2^{E_j}, \quad \text{for} \quad j = 1, 2, \ldots, s.$$

For *I-divergence* it holds that

$$I(P \parallel Q) \geq 0 \quad \text{and} \quad (\, I(P \parallel Q) = 0 \;\Leftrightarrow\; P = Q)$$

and the minimum value of *I_1-aggregate* is equal zero in the case of *consistent input set*. Therefore for all $Q_1, Q_2 \in \mathcal{P}$ that minimize *I_1-aggregate* it holds that

$$\text{for } j = 1, 2, \ldots, s: \quad Q_1^{E_j} = P_{E_j}, \quad Q_2^{E_j} = P_{E_j}.$$

The previous proves that if the *input set* is *consistent* the conjecture holds. Note that, in a special case, there might be only one probability distribution in an *additively constrained set*. However, generally, it is a rare case.

If someone wishes to represent knowledge by a single probability distribution then an additional criterion selecting a unique probability distribution should be defined. In the case of an *consistent input set*, I_1-projection $\pi_{\mathcal{S}_\mathcal{E}} Q_{(0)}$ was used to select a probability distribution from an *additively constrained set* $\mathcal{S}_\mathcal{E}$. It was shown that IPFP converges to this distribution. Furthermore, if the initial probability distribution $Q_{(0)}$ is uniform then the I_1-projection factorizes with respect to a generating class $\mathcal{E}$.

If $\mathcal{T}_1$ is an *additively constrained set* then *I_1-projection* can be used to select a unique distribution from $\mathcal{T}_1$. Following the previous notation it will be denoted $\pi_{\mathcal{T}_1} Q$, i.e.

$$\pi_{\mathcal{T}_1} Q = \arg \min_{\{P \in \mathcal{T}_1\}} I(Q \parallel P).$$

In the case of a *consistent input set* $\mathcal{P}_\mathcal{E}$, it is obvious that $\mathcal{T}_1 = \mathcal{S}_\mathcal{E}$. Therefore, in this case, $\pi_{\mathcal{T}_1} Q$ is equivalent to $\pi_{\mathcal{S}_\mathcal{E}} Q$. Thus any procedure that converges to $\pi_{\mathcal{T}_1} Q_{(0)}$ can be understood as a generalization of IPFP.

5.2 Missing data

The goal of this section is to give a brief overview of methods that are used to solve *missing data problem* in statistics. The methods usually suppose that data are given in the form of a list of data vectors. If every vector has given values for all of the variables under the study then we say, that the data are *complete*. Otherwise, if there is a vector that has at least one unknown value of a variable, it is we deal with *incomplete* or *missing data* [14].

An example of *missing data* that is specified by Table 5.1. Data consists of ten vectors, first three vectors are *complete*, while remaining seven vectors are *incomplete*.

X_1	1	1	2	?	?	?	1	1	2	2
X_2	1	2	2	1	1	2	?	?	?	?
X_3	1	2	2	1	2	1	1	1	1	2

Table 5.1: An example of *missing data*

The following discussion suppose that data are *missing completely at random* (MCAR) which means that probability that a value is missing depends on neither the values that were observed nor the missing value itself. A more general situation, when probability that a value is missing may depend on the values that were observed is called *missing at random* (MAR). If the parameter to be estimated and the parameter of the *missing data mechanism* are distinct then MAR condition is sufficient [35].

There are several basic MCAR methods used in Statistics [31]:

Complete-Case Analysis. In this case, the subsample of the original data set containing complete data vectors is chosen. All incomplete observations are omitted. This approach is often unsatisfactory, particularly, if there is a lot of incomplete observations.

Available-Case Method. Within this method all cases having values of variables of interest are present (other variables can be absent). Suppose that values of probability distributions defined for variables $(X_i)_{i\in E}$ having their indices from a set $E \subset V$ are estimated. Then, within this approach, the subsample of the original data set containing all data vectors that have observed all values of variables $X_{ii\in E}$ are used for estimation. A disadvantage of this approach is that the sample base changes from variable to variable according to the pattern of missing data. In addition, estimations of some model parameters using this method can yield contradictory results.

Fill-In Methods. Fill-In methods tempt to predict missing values from the values of other variables supposing they are correlated. Then these imputed values are included for further computations. There are two basic fill-in methods:

- **Imputing unconditional means**. Values filled-in for a variable are means of the present values of that variable. This method is not generally satisfactory as, for example, it systematically underestimates dependence between variables.

- **Imputing conditional means**. This method substitutes means that are conditioned on the variables recorded in an incomplete case. An example of this approach is the

Buck's method [6] used in the case of multivariate normally distributed variables. The method first estimates the mean and the covariance matrix from the sample mean and the covariance matrix based on the complete cases. Then it uses these estimates to calculate the linear regressions of the missing variables on the present variables, case by case.

If the missing values are filled according to a probabilistic model which is than recomputed using the new filled-in data and if this process is repeated till a stable values are reached then we are getting a well known procedure called *EM-algorithm*.

EM-algorithm was firstly introduced in [14] as a broadly applicable algorithm for computing maximum likelihood estimates from incomplete data. As a reference for the detail description, its methodology, and applications of the algorithm may serve the book of McLachlan and Krishnan [34].

Every iteration of the *EM-algorithm* consists of two steps: *expectation step*(*E-step*) and *maximization step*(*M-step*). At the *expectation step* the algorithm forms the conditional expectation of the log-likelihood function for complete data given the observed data. At the *maximization step* values of model parameters maximizing *conditional expectation of the log-likelihood* are determined. The *EM-algorithm* alternates the *E-step* and the *M-step* using parameters' estimates from the previous steps. If the searched joint probability distribution belongs to an exponential family the EM algorithm can be simplified. Within the *E-step* this simplified version adjusts the values of the sufficient statistics of a model, given the incomplete data and the current value of the parameters. Within the *M-step* this version uses the adjusted values of the sufficient statistics to find the *maximum likelihood estimates*.

A generalization of *EM-algorithm* (called *GEM-algorithm*) does not maximize the *log-likelihood* within *M-step* but only finds values of model parameters such that the *log-likelihood* is a strictly increasing function of the iteration i. An application of this algorithm to our problem will be discussed in Section 5.10.

Multiple Imputation was proposed by Rubin (see [36] for details). During this procedure every *missing value* is consequently being replaced by all possible *fill-in* values so that for every possible combination of *fill-in* values a *complete data set* is created. Then to every *completed data set* a standard *complete data method* is applied. Finally, all results are combined. The resulting model is able to reflect the uncertainty of parameters introduced by *incomplete data*. This method can be also used if there is more than one model of nonresponse.

5.3 Definition of methods

Before presenting the definitions, let us recall that I_1-projection of a probability distribution Q on an *additively constrained set* $\mathcal{S}_{\{E_j\}}$ can be computed as (see formula 2.1)

$$\pi_{\mathcal{S}_{\{E_j\}}} Q(x) \quad = \quad \begin{cases} 0, & \text{for } Q^{E_j}(x^{E_j}) = 0, \\ Q(x)\dfrac{S_E(x^{E_j})}{Q^{E_j}(x^{E_j})}, & \text{for } Q^{E_j}(x^{E_j}) > 0, \end{cases} \tag{5.1}$$

which corresponds to one computational step of IPFP (see Definition 3.2).

Definition 5.3 (Methods for inconsistent knowledge integration)
Let $\mathcal{E} = \{E_1, \dots E_s\}$ be a *generating class* of an *input set* $\mathcal{P}_\mathcal{E}$, $\mathcal{F} = \{F_1, \dots F_t\}$ be a class of subsets of V, $Q_{(0)} \in \mathcal{P}$ be an initial probability distribution such that $P_{E_j} \ll Q_{(0)}^{E_j}$ for all $E_j \in \mathcal{E}$, and $j = ((i-1) \mod s+1)$. Further, let $\{\alpha_i\}_{i=1}^{\infty}$ be a *positive monotonous decreasing sequence* and w_j be nonnegative weights allocated to every distribution P_{E_j}, such that $\sum_{j=1}^{s} w_j = 1$. We define the following computational processes.

- **(CC)**, conservative modification of IPFP with convex mixture [24]:

$$Q_{(i)} \;=\; (1-\alpha_i)\cdot Q_{(i-1)} \;+\; \alpha_i \cdot \left(\pi_{\mathcal{S}_{\{E_j\}}} Q_{(i-1)}\right) \tag{5.2}$$

- **(LCC)**, conservative modification of IPFP with log-convex mixture:

$$Q_{(i)} \;=\; Q_{(i-1)}^{1-\alpha_i} \;\cdot\; \left(\pi_{\mathcal{S}_{\{E_j\}}} Q_{(i-1)}\right)^{\alpha_i} \tag{5.3}$$

- **(AA)**, method of iterative arithmetic averages [32, 24]:

$$Q_{(i)} \;=\; \sum_{j=1}^{s} w_j \cdot \pi_{\mathcal{S}_{\{E_j\}}} Q_{(i-1)} \tag{5.4}$$

- **(GA)**, method of iterative geometric averages [32]:

$$Q_{(i)} \;=\; \frac{1}{c} \cdot \prod_{j=1}^{s} \left(\pi_{\mathcal{S}_{\{E_j\}}} Q_{(i-1)}\right)^{w_j}, \quad \text{where } c = \sum_{x} \prod_{j=1}^{s} \left(\pi_{\mathcal{S}_{\{E_j\}}} Q_{(i-1)}(x)\right)^{w_j} \tag{5.5}$$

- **(GEM)**, an instance of generalized EM-algorithm:

$$\begin{array}{lllcl}
\textbf{E-step:} & \text{for } l = 1,2,\dots,t: & R_{(i)}^{F_l} & = & \left(\sum_{k=1}^{s} w_k \cdot \pi_{\mathcal{S}_{\{E_k\}}} Q_{(i-1)}\right)^{F_l} \\
 & \text{for } l = 1,2,\dots,t: & \mathcal{S}'_{\{F_l\}} & = & \{P \in \mathcal{P} : P^{F_l} = R_{(i)}^{F_l}\} \\
\textbf{M-step:} & & Q_{(i,0)} & = & Q_{(i-1)} \\
 & \text{for } j = 1,2,\dots,t: & Q_{(i,j)} & = & \pi_{\mathcal{S}'_{\{F_j\}}} Q_{(i,j-1)}, \\
 & & Q_{(i)} & = & Q_{(i,t)}
\end{array} \tag{5.6}$$

Note that formulae for $\mathcal{S}'_{\{F_l\}}$ does not require any computation. The only reason it is defined is that the notation of I_1-*projection* requires a definition of set to which the I_1-*projection* is performed.

5.4 Methods from missing data perspective

In this section a "bridge" between *missing data methods* and *inconsistent knowledge integration* is going to be built. Having described the relation between these two approaches we can apply *missing data methods* to the problem of *inconsistent knowledge integration*. Later, in this section, it will be shown how two of proposed methods from Section 5.3 (AA and GEM) fit into the *missing data* approach.

Suppose $\mathbb{Y}$ denotes *incomplete (missing) data* consisting of $N = |\mathbb{Y}|$ *incomplete vectors* $\mathbb{Y} = (y_i)_{i=1}^N$. Every vector y_i consists of two parts defined by:

- E_i - the index set of variables having their values **observed** in vector y_i and
- $V \setminus E_i$ - the index set of variables having their values **unobserved** in vector y_i.

Usually, there is a lot of vectors with the same set of observed variables. Therefore it is convenient to partition *incomplete data* $\mathbb{Y}$ into subsets, each having the same variables observed:

$$\mathbb{Y} = \mathbb{Y}_{E_1} \cup \mathbb{Y}_{E_2} \cup \ldots \cup \mathbb{Y}_{E_s}, \quad \forall i \neq j : \mathbb{Y}_{E_i} \cap \mathbb{Y}_{E_j} = \emptyset.$$

We note that the partitioning of missing data presented above is more general than the missing patterns of S. Haberman regarded in [18], [19], since we do not require index sets E_i of complete subsamples to be disjoint.

Denote the set of all *missing data patterns* by $\mathcal{E} = \{E_1, E_2, \ldots E_s\}$. It is our intention to use the symbol $\mathcal{E}$ which has been already used to denote to a *generating class*. The reason is that, in the following, the *generating class* will be defined by the *missing data patterns*.

In order to be able to fit *missing data approach* into *knowledge integration* framework let us suppose that probability distributions that form an *input set* are estimated as relative frequencies from a given list of data vectors. Having an *input set* $\mathcal{P}_{\mathcal{E}} = \{P_{E_1}, P_{E_2}, \ldots P_{E_s}\}$ (*consistent* or *inconsistent*) we can always create *missing data* $\mathbb{Y}$ which correspond to this set so that all probabilities from *input set* are defined by

$$\forall j \in \{1, 2, \ldots, s\} : \quad P_{E_j}(x^{E_j}) \quad = \quad \frac{|\{y_j \in \mathbb{Y}_{E_j} : y_j^{E_j} = x^{E_j}\}|}{|\mathbb{Y}_{E_j}|} \tag{5.7}$$

i.e. equal to the ratio between number of incomplete data vectors from $\mathbb{Y}_{E_j}$ taking value $y^E = x^E$ and total of data vectors from $\mathbb{Y}_{E_j}$. It will be used in the sequel that to every distribution P_{E_j} from *input set* a weight corresponding to proportion of $|\mathbb{Y}_{E_j}|$ with respect to the dimension of set of all missing data $|\mathbb{Y}|$ is assigned, i.e.

$$\forall j \in \{1, 2, \ldots, s\} : \quad w_j \quad = \quad \frac{|\mathbb{Y}_{E_j}|}{|\mathbb{Y}|}. \tag{5.8}$$

Assume that the probability distribution Q defining probabilities for all possible values x of *complete data*) is known. Then the question is, what is the expected conditional probability given the *incomplete data* $\mathbb{Y}$ and the distribution Q, i.e. $E_Q\{Q(x|\mathbb{Y})\}$. The probability of a *fully observed vector* x can be computed as a sum over all *incomplete vectors* $y_i \in \mathbb{Y}$ having

the observed part equivalent to x of conditional probabilities $Q(x|y_i)$. Thus we can write

$$\begin{aligned} E_Q\{Q(x|\mathbb{Y})\} &= \frac{1}{|\mathbb{Y}|} \sum_{y_i \in \mathbb{Y}:\ y_i^{E_i} = x^{E_i}} Q(x|y_i) \\ &= \frac{1}{|\mathbb{Y}|} \sum_{E_j \in \mathcal{E}} \sum_{y_j \in \mathbb{Y}_{E_j}:\ y_j^{E_j} = x^{E_j}} Q(x|y_j) \end{aligned}$$

Using equations 5.7 and 5.8 we can write

$$\begin{aligned} E_Q\{Q(x|\mathbb{Y})\} &= \frac{1}{|\mathbb{Y}|} \sum_{E_j \in \mathcal{E}} |\mathbb{Y}_{E_j}| \cdot P_{E_j}(x^{E_j}) \cdot Q(x|x^{E_j}) \\ &= \sum_{E_j \in \mathcal{E}} w_{E_j} \cdot P_{E_j}(x^{E_j}) \cdot Q(x|x^{E_j}) \\ &= \sum_{E_j \in \mathcal{E}} w_{E_j} \cdot P_{E_j}(x^{E_j}) \cdot \frac{Q(x)}{Q^{E_j}(x^{E_j})} \end{aligned} \tag{5.9}$$

We assumed that the probability distribution Q had been known. If, instead of a given Q, the resulting distributions $E_Q\{Q(x|\mathbb{Y})\}$ are used for Q and the expected conditional probabilities given the *incomplete data* $E_Q\{Q(x|\mathbb{Y})\}$ are iteratively beeing recomputed then we get to the AA method (defined already in Section 5.3).

In addition to the previous, let the class $\mathcal{F}$ of subsets of V define a multiplicatively constrained set $\mathcal{R}_\mathcal{F}$ and let the resulting distribution be required to belong to $\mathcal{R}_\mathcal{F}$. A classical statistical approach to such a task is *maximum likelihood estimation from missing data.*

Definition 5.4 (Maximum likelihood estimation from missing data)
Let $\mathbb{Y}$ be a given *incomplete data* and set $\mathcal{F}$ define a multiplicatively constrained set $\mathcal{R}_\mathcal{F}$. Then a probability distribution $Q \in \mathcal{R}_\mathcal{F}$ that maximizes

$$\arg \max_{Q \in \mathcal{R}_\mathcal{F}} \prod_x Q(x|\mathbb{Y})$$

is defined to be a *maximum likelihood estimate from missing data.*

It is a consequence of a result proved by R. Sundberg [38] for *exponential families* that for a *maximum likelihood estimate* $\hat{Q}$ it must hold that for $k = 1, 2, \ldots, t$

$$E_{\hat{Q}}\{\hat{Q}^{F_k}(x^{F_k}|\mathbb{Y})\} = \hat{Q}^{F_k}(x^{F_k}).$$

Substituting equation 5.9 for $E_{\hat{Q}}\{\hat{Q}(x|\mathbb{Y})\}$ we get for $k = 1, 2, \ldots, t$

$$\begin{aligned} \hat{Q}^{F_k} &= \left(\sum_{E_j \in \mathcal{E}} w_{E_j} \cdot P_{E_j}(x^{E_j}) \cdot \frac{\hat{Q}(x)}{\hat{Q}^{E_j}(x^{E_j})} \right)^{F_k} \\ &= \left(\sum_{E_j \in \mathcal{E}} w_{E_j} \cdot P_{E_j}(x^{E_j}) \cdot \frac{\hat{Q}^{E_j \cup F_k}(x)}{\hat{Q}^{E_j}(x^{E_j})} \right)^{F_k}. \end{aligned} \tag{5.10}$$

Next, we will show a relation between *minimization of I_1-aggregate* and *maximum likelihood estimation* from missing data. Let us restrict the set $\mathcal{F}$ so that $\mathcal{F} = \{V\}$. Then all equations given by 5.10 reduce to one

$$\hat{Q} \quad = \quad \sum_{E_j \in \mathcal{E}} w_{E_j} \cdot P_{E_j}(x^{E_j}) \cdot \frac{\hat{Q}(x)}{\hat{Q}^{E_j}(x^{E_j})}. \tag{5.11}$$

We have shown above that within the AA method the recomputation of expected conditional probability given the *incomplete data* $E_Q\{Q(x|\mathbb{Y})\}$ is iteratively repeated until a convergence to a limit distribution Q^* is reached. It was proved in [32] that the AA method minimizes

$$\sum_{j=1}^{s} w_j \cdot I(\pi_{\mathcal{S}_{\{E_j\}}} Q \parallel Q),$$

which we have shown to be equivalent to *I_1-aggregate* (Lemma 5.1). For a limit distribution Q^* of the AA method it must hold that

$$Q^* \quad = \quad \sum_{E_j \in \mathcal{E}} w_{E_j} \cdot P_{E_j}(x^{E_j}) \cdot \frac{Q^*(x)}{Q^{*E_j}(x^{E_j})}. \tag{5.12}$$

Comparing equations 5.11 and 5.12 we can see that the conditions for a *maximum likelihood* estimate and a minimizer of the *I_1-aggregate* are equivalent.

5.5 Properties of iterative methods

We have presented five iterative methods that can be applied to both *consistent and inconsistent input sets*. In this section we summarize properties of the described methods. The summary is based not only on proven results but also on observations that were not proven yet. We will start with Table 5.2 where all six procedures are compared with respect to nine basic properties. In the following text we will describe the results in detail. For an easy access to the text where a given property is discussed we recommend to use **Index** attached at the end of the thesis. Note that boldface answers refer to results already proven while the other (written in *italics*) are based on our observations only. Limit distributions of corresponding algorithms are denoted by Q^*.

IPFP applied to an *inconsistent input set* cycles, but after a certain number of cycles the procedure stabilizes in a cycle consisting of j distributions, let us denote them $Q^*_j, j = 1, 2, \ldots s$. Than one can simply compute the *arithmetic* or *normalized geometric average* of the respective "limit" distributions (see Section 3.3). In Table 5.3 such distributions are denoted Q^*_a and Q^*_g, respectively.

In the next sections we are going to give details for every cell of both tables. It will be either a proof of the property for a corresponding method, reference to a proof presented somewhere else, or the reasons why we conjecture that the respective property holds. The text is organized so that every property of a given method is referred by the abbreviation of the corresponding property and by the abbreviation of the corresponding method. For example, **C1.GA** stands for a *convergence* property of the *method of geometric averages* on a *consistent input set*. The last section will be devoted to the *convergence rates* and the *computational complexity* of the proposed methods.

Table 5.2: Methods' properties

Property description		IPFP	CC	LCC	AA	GA	GEM
G1	$Q_{(0)} \in \mathcal{R}_{\mathcal{E}} \quad \Rightarrow \quad i = 1, 2, \ldots : \; Q_{(i)} \in \mathcal{R}_{\mathcal{E}}$	**yes**	**no**	**yes**	**no**	**yes**	**yes**
consistent $\mathcal{P}_{\mathcal{E}}$							
C1	Convergence	**yes**	*yes*	*yes*	**yes**	**yes**	**yes**
C2	Q^* is indep. of the ordering of $\mathcal{P}_{\mathcal{E}}$	**yes**	*yes*	*yes*	**yes**	**yes**	*yes*
C3	$Q^* = \pi_{\mathcal{E}} Q_{(0)}$	**yes**	*yes*	*yes*	*no*	**yes**	*yes*
inconsistent $\mathcal{P}_{\mathcal{E}}$							
I1	Convergence	**no**	*yes*	*yes*	**yes**	**yes**	**yes**
I2	Q^* is indep. of the ordering of $\mathcal{P}_{\mathcal{E}}$		*yes*	*yes*	**yes**	**yes**	*yes*
I3	$Q^* \in \mathcal{T}_1$		*yes*	*no*	**yes**	*no*	*yes*
I4	$Q^* \in \mathcal{T}_2$		*no*	*yes*	*no*	**yes**	*no*

Table 5.3: IPFP on inconsistent input

Property description		$Q \in \{Q_j^*, j = 1, 2, \ldots s\}$	$Q = Q_a^*$	$Q = Q_g^*$
I2	Indep. of the ordering of $\mathcal{P}_{\mathcal{E}}$	*no*	*no*	*no*
I3	$Q \in \mathcal{T}_1$	**no**	*no*	**no**
I4	$Q \in \mathcal{T}_2$	**no**	**no**	**no**

5.6 On some properties of IPFP limit cycles

Substantial part of this section analyzes an example of an application of IPFP to the *input set* that consists of two distributions. The values of these distributions are treated symbolically so that general properties of I_1-*aggregate* and I_2-*aggregate* for distributions of the limit cycle, their arithmetic average, and their normalized geometric average can be derived. This example is a proof of the assertion that neither the *geometric average* of the limit distributions nor the distributions from a limit cycle themselves are generally minimizers of the I_1-*aggregate.* and the I_2-*aggregate.* The *arithmetic average* of the limit distributions is shown that it is not minimizer of I_2-*aggregate* while this example does not disprove it to be a minimizer of the I_1-*aggregate.*

Example 5.1 (IPFP on inconsistent input - continued Examples 3.2 and 3.4) Suppose that the *input set* consists of two distributions $P_{\{1,2\}}(X_1, X_2)$ and $P_{\{2,3\}}(X_2, X_3)$, whose values were defined by Figure 3.1 in Example 3.2. Recall, that after three iterations IPFP stabilized in the cycle of two joint probability distributions $Q_{(2i)} = Q_{(2)}$ and $Q_{(2i+1)} = Q_{(3)}$ for $i = 2, 3, \ldots$. In Table 5.4 their values are displayed in the symbolic form. Thus, we

Table 5.4: Values of $Q_{(2)}$ and $Q_{(3)}$

X_1	X_2	X_3	$Q_{(2)}$	$Q_{(3)}$
1	1	1	$\frac{ae}{(a+b)}$	$\frac{ae}{(e+f)}$
1	1	2	$\frac{af}{(a+b)}$	$\frac{af}{(e+f)}$
1	2	1	$\frac{cg}{(c+d)}$	$\frac{cg}{(g+h)}$
1	2	2	$\frac{ch}{(c+d)}$	$\frac{ch}{(g+h)}$
2	1	1	$\frac{be}{(a+b)}$	$\frac{be}{(e+f)}$
2	1	2	$\frac{bf}{(a+b)}$	$\frac{bf}{(e+f)}$
2	2	1	$\frac{dg}{(c+d)}$	$\frac{dg}{(g+h)}$
2	2	2	$\frac{dh}{(c+d)}$	$\frac{dh}{(g+h)}$

can compute two *unique joint probability distributions*:

$$Q_a^* = \frac{1}{2}(Q_{(2)} + Q_{(3)})$$

and the normalized geometric average

$$Q_g^* = \frac{\sqrt{Q_{(2)}Q_{(3)}}}{\sum_x \sqrt{Q_{(2)}(x)Q_{(3)}(x)}}.$$

In Table 5.5 the values of these distributions are displayed using the symbolic form.

Table 5.5: Values of Q_a^* and Q_g^*

X_1	X_2	X_3	Q_a^*	Q_g^*
1	1	1	$\frac{ae((e+f)+(a+b))}{2(e+f)(a+b)}$	$\frac{ae}{\sqrt{(a+b)(e+f)}\left(\sqrt{(a+b)(e+f)}+\sqrt{(c+d)(g+h)}\right)}$
1	1	2	$\frac{af((e+f)+(a+b))}{2(e+f)(a+b)}$	$\frac{af}{\sqrt{(a+b)(e+f)}\left(\sqrt{(a+b)(e+f)}+\sqrt{(c+d)(g+h)}\right)}$
1	2	1	$\frac{cg((g+h)+(c+d))}{2(g+h)(c+d)}$	$\frac{cg}{\sqrt{(c+d)(g+h)}\left(\sqrt{(a+b)(e+f)}+\sqrt{(c+d)(g+h)}\right)}$
1	2	2	$\frac{ch((g+h)+(c+d))}{2(g+h)(c+d)}$	$\frac{ch}{\sqrt{(c+d)(g+h)}\left(\sqrt{(a+b)(e+f)}+\sqrt{(c+d)(g+h)}\right)}$
2	1	1	$\frac{be((e+f)+(a+b))}{2(e+f)(a+b)}$	$\frac{be}{\sqrt{(a+b)(e+f)}\left(\sqrt{(a+b)(e+f)}+\sqrt{(c+d)(g+h)}\right)}$
2	1	2	$\frac{bf((e+f)+(a+b))}{2(e+f)(a+b)}$	$\frac{bf}{\sqrt{(a+b)(e+f)}\left(\sqrt{(a+b)(e+f)}+\sqrt{(c+d)(g+h)}\right)}$
2	2	1	$\frac{dg((g+h)+(c+d))}{2(g+h)(c+d)}$	$\frac{dg}{\sqrt{(c+d)(g+h)}\left(\sqrt{(a+b)(e+f)}+\sqrt{(c+d)(g+h)}\right)}$
2	2	2	$\frac{dh((g+h)+(c+d))}{2(g+h)(c+d)}$	$\frac{dh}{\sqrt{(c+d)(g+h)}\left(\sqrt{(a+b)(e+f)}+\sqrt{(c+d)(g+h)}\right)}$

Let us start by solving the necessary conditions for the minima. There are seven independent variables and one dependent variable in the minimization task. Every variable corresponds to one value of resulting *joint probability distribution*. Let them be defined by the following relations:

$$\begin{aligned}
&x_1 = Q(X_1 = 1, X2 = 1, X3 = 1), \quad x_2 = Q(X_1 = 1, X2 = 1, X3 = 2),\\
&x_3 = Q(X_1 = 2, X2 = 1, X3 = 1), \quad x_4 = Q(X_1 = 2, X2 = 1, X3 = 2),\\
&x_5 = Q(X_1 = 1, X2 = 2, X3 = 1), \quad x_6 = Q(X_1 = 1, X2 = 2, X3 = 2),\\
&x_7 = Q(X_1 = 2, X2 = 2, X3 = 1),\\
&x_8 = 1 - x_1 - x_2 - x_3 - x_4 - x_5 - x_6 - x_7 = Q(X_1 = 2, X2 = 2, X3 = 2).
\end{aligned}$$

For all the distributions minimizing the I_1-*aggregate* $\psi_1(x_1, x_2, \ldots, x_7)$ it must hold that their partial derivations are equal to zero, i.e.

$$\frac{\partial}{\partial x_i}\ \psi_1(x_1, x_2, \ldots, x_7) = 0, \quad i = 1, 2, \ldots, 7.$$

By performing the derivations we get

$$\frac{a}{x_1 + x_2} + \frac{e}{x_1 + x_3} = \frac{h}{x_6 + x_8} + \frac{d}{x_7 + x_8} \tag{5.13}$$

$$\frac{a}{x_1 + x_2} + \frac{f}{x_2 + x_4} = \frac{h}{x_6 + x_8} + \frac{d}{x_7 + x_8} \tag{5.14}$$

$$\frac{b}{x_3 + x_4} + \frac{e}{x_1 + x_3} = \frac{h}{x_6 + x_8} + \frac{d}{x_7 + x_8} \tag{5.15}$$

$$\frac{b}{x_3 + x_4} + \frac{f}{x_2 + x_4} = \frac{h}{x_6 + x_8} + \frac{d}{x_7 + x_8} \tag{5.16}$$

$$\frac{c}{x_5 + x_6} + \frac{g}{x_5 + x_7} = \frac{h}{x_6 + x_8} + \frac{d}{x_7 + x_8} \tag{5.17}$$

$$\frac{d}{x_7 + x_8} = \frac{c}{x_5 + x_6} \tag{5.18}$$

$$\frac{g}{x_5 + x_7} = \frac{h}{x_6 + x_8} \tag{5.19}$$

These equations can be combined and the dependent ones eliminated so that the resulting system is defined by

$$\begin{aligned}
&\text{from 5.13 and 5.15 or from 5.24 and 5.26:} && \frac{a}{b} = \frac{x_1 + x_2}{x_3 + x_4}\\
&\text{from 5.18 or from 5.29:} && \frac{c}{d} = \frac{x_5 + x_6}{x_7 + x_8}\\
&\text{from 5.19 or from 5.30} && \frac{g}{h} = \frac{x_5 + x_7}{x_6 + x_8}\\
&\text{from 5.13 and 5.14 or from 5.13 and 5.14:} && \frac{e}{f} = \frac{x_1 + x_3}{x_2 + x_4}
\end{aligned}$$

If we return to the notation of probability distributions then these conditions can be rewritten as

$$\text{for } j = 1, 2\text{:} \quad \frac{Q^{\{1,2\}}(X_1 = 1, X_2 = j)}{Q^{\{1,2\}}(X_1 = 2, X_2 = j)} = \frac{P_{\{1,2\}}(X_1 = 1, X_2 = j)}{P_{\{1,2\}}(X_1 = 2, X_2 = j)} \tag{5.20}$$

$$\text{for } j=1,2\text{:}\quad \frac{Q^{\{2,3\}}(X_2=j,X_3=1)}{Q^{\{2,3\}}(X_2=j,X_3=2)} \;=\; \frac{P_{\{2,3\}}(X_2=j,X_3=1)}{P_{\{2,3\}}(X_2=j,X_3=2)} \tag{5.21}$$

The following table shows that these conditions are satisfied for distributions from the limit cycle and their arithmetic and geometric averages as well.

	$\frac{x_1+x_2}{x_3+x_4}$	$\frac{x_5+x_6}{x_7+x_8}$	$\frac{x_5+x_7}{x_6+x_8}$	$\frac{x_1+x_3}{x_2+x_4}$
$Q_{(2)}$	$\frac{a}{b}$	$\frac{c}{d}$	$\frac{g}{h}$	$\frac{e}{f}$
$Q_{(3)}$	$\frac{a}{b}$	$\frac{c}{d}$	$\frac{g}{h}$	$\frac{e}{f}$
Q_a^*	$\frac{a}{b}$	$\frac{c}{d}$	$\frac{g}{h}$	$\frac{e}{f}$
Q_g^*	$\frac{a}{b}$	$\frac{c}{d}$	$\frac{g}{h}$	$\frac{e}{f}$

Assertion 5.1 (I3.IPFP) *For the given example the distributions of the limit cycle and the distribution computed as their normalized geometric average does not minimize the I_1-aggregate.*

Proof. Let us compare distributions' values of I_1-*aggregate* ψ_1. Since the resulting values of this criterion depends only on values of marginals of distributions from the *input set* marginalized to the intersection $\{2\}=\{1,2\}\cap\{2,3\}$ we can substitute

$$P^{\{2\}}_{\{1,2\}}(X_2=1)=a+b=u, \qquad P^{\{2\}}_{\{1,2\}}(X_2=2)=c+d=1-u,$$
$$P^{\{2\}}_{\{2,3\}}(X_2=1)=e+f=v, \qquad P^{\{2\}}_{\{2,3\}}(X_2=2)=g+h=1-v.$$

Thus the resulting probability distributions evaluated by the criterion ψ_1 gives the following values

$$\begin{aligned}
\psi_1(Q_{(2)}) &= -H(P^{\{2\}}_{\{1,2\}}) - (1-u)\log(1-v) - u\log(v)\\
\psi_1(Q_{(3)}) &= -H(P^{\{2\}}_{\{2,3\}}) - (1-v)\log(1-u) - v\log(u)\\
\psi_1(Q_a^*) &= -H(P^{\{2\}}_{\{2,3\}}) - H(P^{\{2\}}_{\{1,2\}}) \;+\; 2\log 2\\
&\quad -\; (u+v)\log(u+v) \;-\; (2-u-v)\log(2-u-v)\\
\psi_1(Q_g^*) &= -H(P^{\{2\}}_{\{2,3\}}) - H(P^{\{2\}}_{\{1,2\}}) \;+\; 2\log\left(\sqrt{uv}+\sqrt{(1-u)(1-v)}\right)\\
&\quad -\; (u+v)\log\sqrt{u+v} \;-\; (2-u-v)\log\sqrt{2-u-v}
\end{aligned}$$

Comparing the functions and using information theoretical inequalities

$$\left(\sum_x P(x)=1 \;\text{ and }\; \sum_x Q(x)=1\right) \;\Rightarrow\; \left(P(x)\log P(x)>0,\quad P(x)\log\frac{Q(x)}{P(x)}\geq 0\right)$$

it can be shown (the computations are rather space consuming) that some resulting distributions can not be better (measured by ψ_1) than others. Preferably to the details of the computations we use 3-dimensional plots to compare values of $\psi_1(Q_{(2)})$, $\psi_1(Q_{(3)})$, $\psi_1(Q_a^*)$, and $\psi_1(Q_g^*)$, being a function of u and v.

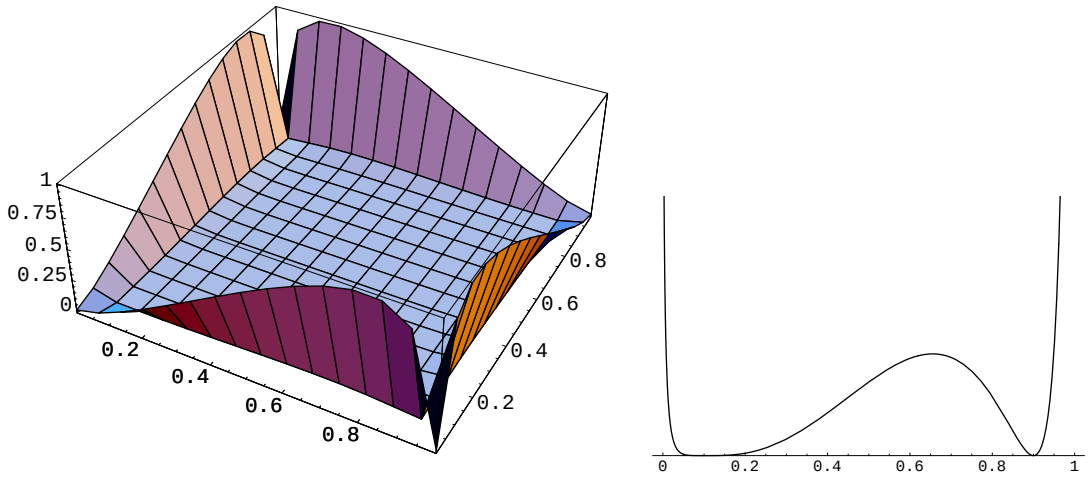


Figure 5.1: Values of $\psi_1(Q_g^*) - \psi_1(Q_a^*)$ together with a slice for value of $u = 1/10$.

First, let us compare values of two distributions that were computed as averages of distributions in the limit cycle. On Figure 5.1 it can be seen that value of $\psi_1(Q_g^*)$ is never lower than $\psi_1(Q_a^*)$

$$\psi_1(Q_g^*) = \psi_1(Q_a^*) \quad \Leftrightarrow \quad u = v \ \vee \ u = 1 - v \tag{5.22}$$

$$\psi_1(Q_g^*) > \psi_1(Q_a^*) \quad \Leftrightarrow \quad u \neq v \ \wedge \ u \neq 1 - v \tag{5.23}$$

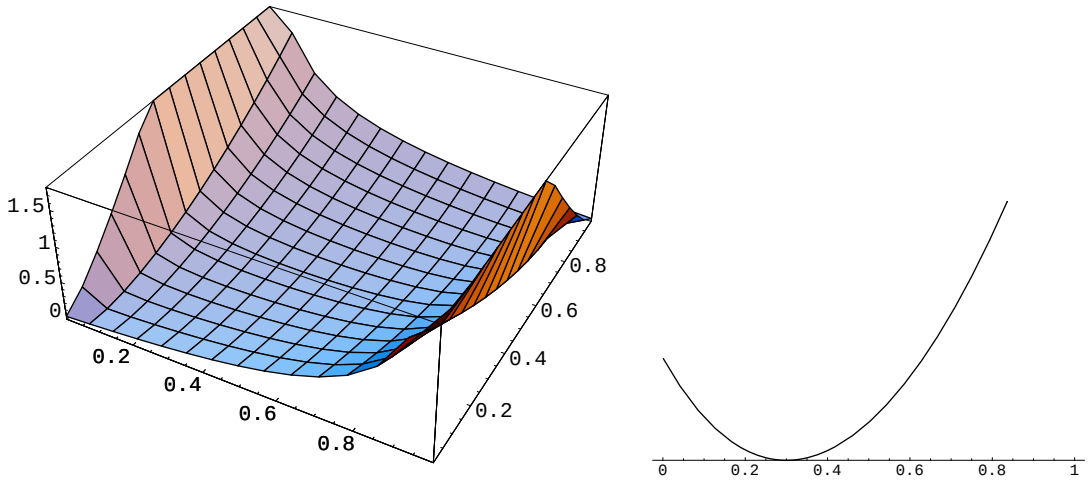


Figure 5.2: Values of $\psi_1(Q_{(3)}) - \psi_1(Q_a^*)$ together with a slice for value of $u = 3/10$.

Next, we will compare values of distributions from the limit cycle with the one computed as their arithmetic average. On Figure 5.2 it can be seen that value of $\psi_1(Q_{(3)})$ is never lower then $\psi_1(Q_a^*)$. The comparison of $\psi_1(Q_{(2)})$ with $\psi_1(Q_a^*)$ gives plot that differs only in

transposed axes.

$$\text{for } j = 2,3: \quad \psi_1(Q_{(j)}) = \psi_1(Q_a^*) \quad \Leftrightarrow \quad u = v$$
$$\text{for } j = 2,3: \quad \psi_1(Q_{(j)}) > \psi_1(Q_a^*) \quad \Leftrightarrow \quad u \neq v$$

Thus we have shown (equation 5.22) that $u = v$ implies $\psi_1(Q_a^*) = \psi_1(Q_g^*)$, but the opposite implication does not hold. Since for any $j \in \{2,3\}$ there exist two pairs of input distributions values $(u = u_1, v = v_1)$, $(u = u_2, v = v_2)$ such that

$$\text{for } u = u_1, v = v_1: \ \psi_1(Q_g^*) > \psi_1(Q_{(j)}) \ \text{ and for } u = u_2, v = v_2: \ \psi_1(Q_g^*) < \psi_1(Q_{(j)})$$

it proves that it does not hold generally, that neither Q_g^* nor $\psi_1(Q_{(j)}), j = 2,3$ is a minimizer of the I_1-*aggregate* □

Values of $\psi_1(Q_{(2)})$, $\psi_1(Q_{(3)})$, and $\psi_1(Q_g^*)$ does not satisfy any relations similar to inequalities 5.23 and 5.24. On Figures 5.3 and 5.4 it can be seen that it depends on values of u and v which of the two distributions achieved higher value of ψ_1. The only exception is the following statement.

$$\psi_1(Q_g^*) = \psi_1(Q_{(2)}) = \psi_1(Q_{(3)}) \quad \Leftrightarrow \quad u = v.$$

It means that if probability distributions from *input set* are *weakly consistent* then the values of ψ_1 are equivalent for all these distributions. This is nothing surprising since, in this case, I_1-*aggregate* equals zero.

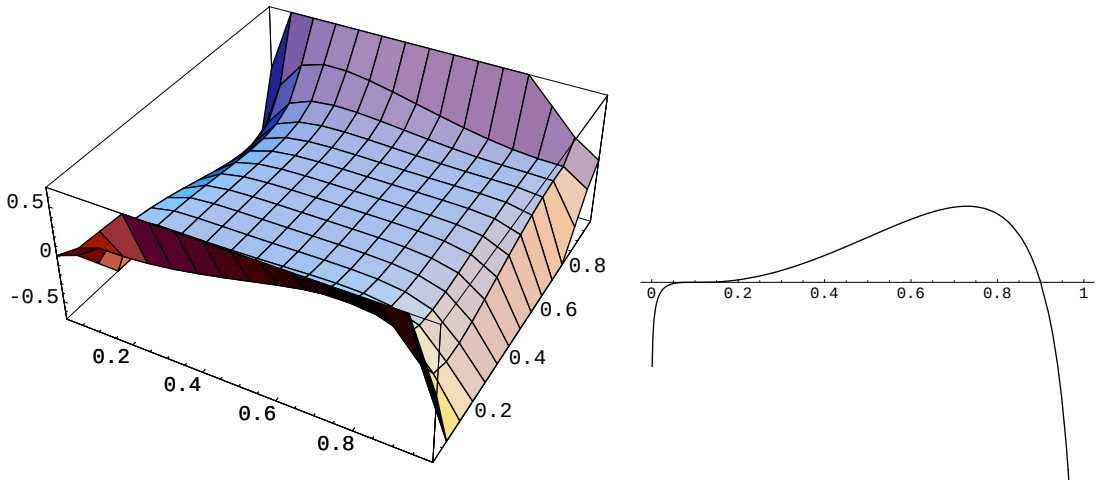


Figure 5.3: Values of $\psi_1(Q_{(2)}) - \psi_1(Q_{(3)})$ being a function of u and v, together with a slice for value of $v = 1/10$.

Note that we have not proved that Q_a^* does not generally minimize ψ_1, since it seems to be a global minimizer of ψ_1 in this example.

For all the distributions minimizing the I_2-*aggregate* $\psi_2(x_1, x_2, \ldots, x_7)$ it must hold that partial derivations of ψ_2 are equal to zero, i.e.

$$\frac{\partial}{\partial x_i} \psi_2(x_1, x_2, \ldots, x_7) = 0, \quad i = 1, 2, \ldots, 7.$$

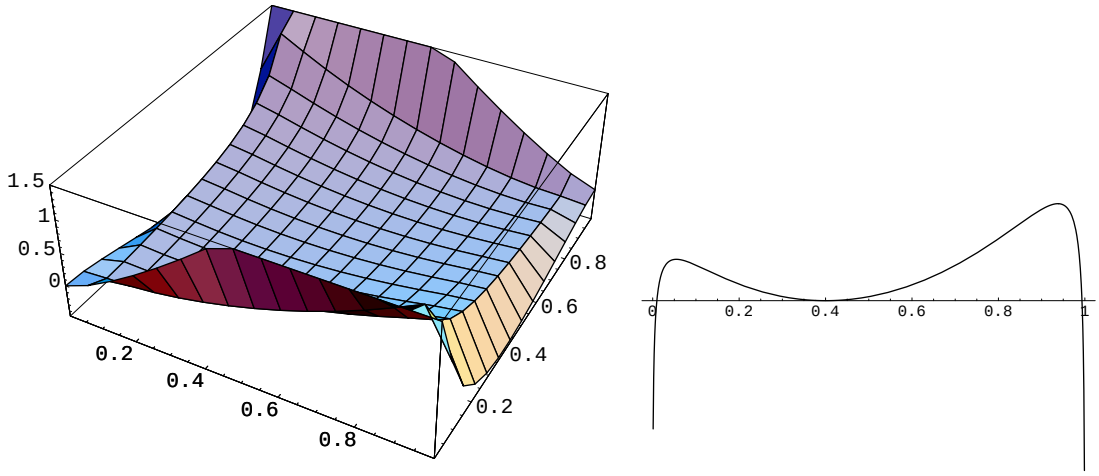


Figure 5.4: Values of $\psi_1(Q_{(2)}) - \psi_1(Q_g^*)$ being a function of u and v, together with a slice for value of $v = 4/10$.

By performing the derivations we get

$$\log\left(\frac{a}{x_1+x_2}\right) + \log\left(\frac{e}{x_1+x_3}\right) = \log\left(\frac{h}{x_6+x_8}\right) + \log\left(\frac{d}{x_7+x_8}\right) \quad (5.24)$$

$$\log\left(\frac{a}{x_1+x_2}\right) + \log\left(\frac{f}{x_2+x_4}\right) = \log\left(\frac{h}{x_6+x_8}\right) + \log\left(\frac{d}{x_7+x_8}\right) \quad (5.25)$$

$$\log\left(\frac{b}{x_3+x_4}\right) + \log\left(\frac{e}{x_1+x_3}\right) = \log\left(\frac{h}{x_6+x_8}\right) + \log\left(\frac{d}{x_7+x_8}\right) \quad (5.26)$$

$$\log\left(\frac{b}{x_3+x_4}\right) + \log\left(\frac{f}{x_2+x_4}\right) = \log\left(\frac{h}{x_6+x_8}\right) + \log\left(\frac{d}{x_7+x_8}\right) \quad (5.27)$$

$$\log\left(\frac{c}{x_5+x_6}\right) + \log\left(\frac{g}{x_5+x_7}\right) = \log\left(\frac{h}{x_6+x_8}\right) + \log\left(\frac{d}{x_7+x_8}\right) \quad (5.28)$$

$$\log\left(\frac{d}{x_7+x_8}\right) = \log\left(\frac{c}{x_5+x_6}\right) \quad (5.29)$$

$$\log\left(\frac{g}{x_5+x_7}\right) = \log\left(\frac{h}{x_6+x_8}\right) \quad (5.30)$$

Involving equations 5.24-5.30 we can see that the conditions for the solution of the previous system are equivalent to the necessary conditions for ψ_1 minimizers given by Equations 5.20 and 5.21.

Assertion 5.2 (I4.IPFP) *For the given example the distributions of the limit cycle, the distribution computed as their arithmetic average, and the distribution computed as their normalized geometric average does not minimize the I_2-aggregate.*

Proof. The comparisons on distributions $Q_{(2)}$, $Q_{(3)}$, Q_a^*, and Q_q^* evaluated by the criterion ψ_2 gives the following values

$$\psi_2(Q_{(2)}) = v \cdot \log\frac{v}{u} + (1-v)\cdot\log\frac{1-v}{1-u} = I(P^{\{2\}}_{\{2,3\}} || P^{\{2\}}_{\{1,2\}})$$

$$
\begin{aligned}
\psi_2(Q_{(3)}) &= u \cdot \log \frac{u}{v} + (1-u) \cdot \log \frac{1-u}{1-v} = I(P^{\{2\}}_{\{1,2\}} || P^{\{2\}}_{\{2,3\}}) \\
\psi_2(Q^*_a) &= \frac{u+v}{2} \cdot \log \frac{\left(\frac{u+v}{2}\right)^2}{uv} + \frac{(1-u)+(1-v)}{2} \cdot \log \frac{\left(\frac{(1-u)+(1-v)}{2}\right)^2}{(1-u)(1-v)} \\
&= 2 \cdot I\left(P^{\{2\}}_a || \sqrt{P^{\{2\}}_{\{1,2\}} \cdot P^{\{2\}}_{\{2,3\}}}\right) \\
\psi_2(Q^*_g) &= -2 \log\left(\sqrt{(1-u)(1-v)} + \sqrt{uv}\right) \\
&\quad - \frac{\sqrt{(1-u)(1-v)}}{\sqrt{(1-u)(1-v)} + \sqrt{uv}} \cdot \log\left(\sqrt{(1-u)(1-v)}\right) \\
&\quad - \frac{\sqrt{uv}}{\sqrt{(1-u)(1-v)} + \sqrt{uv}} \cdot \log \sqrt{u * v} \\
&= H(P^{\{2\}}_g) - 4 \log\left(\sqrt{(1-u)(1-v)} + \sqrt{uv}\right)
\end{aligned}
$$

where $P^{\{2\}}_a$ and $P^{\{2\}}_g$ denote *arithmetic* average and *normalized geometric* average of the marginals $P^{\{2\}}_{\{1,2\}}$ and $P^{\{2\}}_{\{2,3\}}$, respectively.

Using these formula we can show that none of the four distributions $Q_{(2)}$, $Q_{(3)}$, Q^*_a, and Q^*_q is generally minimizer of I_2-*aggregate* ψ_2. See the next table where three examples proving the assertion are given. The minima of the criterion values are highlighted.

u	v	$\psi_2(Q_{(2)})$	$\psi_2(Q_{(3)})$	$\psi_2(Q^*_a)$	$\psi_2(Q^*_g)$
0.500	0.001	**0.68524**	2.76123	1.08483	1.29852
0.500	0.100	0.36806	0.51082	0.89705	**0.23594**
0.001	0.500	2.76123	**0.68524**	1.08483	1.29852

This proves the assertion. □

In Example 3.6 it appears that the procedure converges to the cycles which are dependent on the ordering of the *input set*. We observe that *arithmetic and geometric averages* are different as well. If we evaluate the *arithmetic and geometric averages* of the distributions from two different cycles (that appears to be limit) we observe that the values of ψ_1 are different and thus they both can not be minimizers of the respective criterion (the **I3.IPFP** property). See Table 5.6.

It looks like that if the *generating class* of an *input set* is not *decomposable* then different ordering of the *input set* yields to different *limit cycles* with different *arithmetic* and *geometric average* of distributions from *limit cycles* (the **I2.IPFP** property). See Figure 3.8 on page 52.

5.7 Conservative modifications of IPFP

Convergence of *conservative modifications* of IPFP, defined by formulae 5.2 and 5.3 and abbreviated by CC and LCC, respectively, is distinctively dependent on a type of sequence

Order	$\psi_1(Q_a^*)$	$\psi_1(Q_g^*)$	$\psi_2(Q_a^*)$	$\psi_2(Q_g^*)$
$\{E_1, E_2, E_3, E_4\}$	0.614710	0.622487	0.675455	0.677116
$\{E_4, E_2, E_1, E_3\}$	0.594673	0.620673	0.673036	0.684624

Table 5.6: Values of ψ_1 and ψ_2 for distribution that are *arithmetic averages* and *geometric averages* of distributions from *limit cycles* of IPFP applied to inconsistent input for the different orderings of *input set*

$\{\alpha_i\}$. Every α_i is a weight used to mixture actually computed joint probability distribution with one from previous step. For example, if for all $i = 1, 2, \ldots$ it holds that $\alpha_i = 1$ then both methods are equivalent to regular IPFP. On the opposite end of the scale are sequences for which there exists a finite j such that for $i > j$: $\alpha_i = 0$, i.e. the procedure stops in a fixed point j. It is obvious that neither of these sequences may generally yields satisfactory results on *inconsistent input set*. Intuitively speaking, it must be somehow guaranteed that the procedure will be "controlled" by the values of the probability distributions from an *input set* rather then the type of the sequence. Furthermore, the ordering of the *input set* should not matter. We have made experiments with different sequences of coefficients $\{\alpha_i\}$. In order to fulfill the requirements discussed above it appears that a sequence $\{\alpha_i\}$ must obey the following five requirements.

- $\{\alpha_i\}$ is monotonously decreasing, (5.31)
- for $i = 1, 2, \ldots \quad 0 \leq \alpha_i < 1$, (5.32)
- $\lim_{i \to \infty} \alpha_i = 0$, (5.33)
- for $i = 1, 2, \ldots \quad \sum_{n=i}^{\infty} \alpha_n = +\infty$, (5.34)
- $\{\alpha_i\}$ is stable within one cycle, i.e.
  if $((i-1) \text{ div } s) = ((j-1) \text{ div } s)$ then $\alpha_i = \alpha_j$, (5.35)

where m div n denotes integer part of $\frac{m}{n}$.

F. Matúš [33] has proved the convergence of LCC in *consistent case* for certain types of sequences $\{\alpha_i\}$ called *(under-)relaxation sequences* (see [8]), i.e.

$$0 < \delta \leq 1, \quad \delta \leq \alpha_i, \quad 1 \leq i.$$

Theorem 5.1 (C1.LCC) *If the initial probability distribution $Q_{(0)}$ dominates at least one probability distribution from $\mathcal{S}_\mathcal{E}$ and the sequence $\{\alpha_i\}$ is an (under-)relaxation sequence then LCC procedure converge to $\pi_{\mathcal{S}_\mathcal{E}} Q_{(0)}$.*

Note, that the sequence of coefficients which we have used in our experiments (given by equation 5.36) is not an *(under-)relaxation sequence*, since there is no $0 < \delta \leq 1$ and the proof of LCC convergence for a sequence $\{\alpha_i\}$ satisfying properties 5.31, 5.32, 5.33, 5.34, and 5.35 is still missing. However, after a number of computational experiments with CC

and LCC (see e.g. Figure 5.7) we conjecture that both methods converge to the same joint probability distribution as IPFP and the ordering of *input set* is not important with respect to limit distribution (**C1.CC**, **C1.LCC**, **C2.CC**, **C2.LCC**, **C3.CC**, and **C3.LCC**). On the other hand, the ordering of *input set* may be important with respect to the speed of convergence.

If $Q_{(0)}$ factorize with respect to the same $\mathcal{E}$ then the iterated joint probability distributions $Q_{(i)}, i = 1, 2, \ldots$ computed by LCC factorize with respect to $\mathcal{E}$ as well. If two distributions factorize with respect to the set $\mathcal{E}$ then their product factorizes with respect to $\mathcal{E}$ as well. The property **G1.LCC**) is an immediate consequence of this fact and Theorem 3.3. Summation of probability distribution factorizing with respect to $\mathcal{E}$ does not generally result into a probability distribution factorizing with respect to $\mathcal{E}$ (property referred as **G1.CC**).

Results of application of CC and LCC to an *inconsistent input set* are shown in Figure 5.5 and compared in Table 5.7. Note that, for simplicity, the values of

$$Q_{(800)}(X_1 = 1,\ X_2 = 1,\ X_3 = 2,\ X_4 = 1)$$

are displayed only. Two orderings of the *input set* were used:

$$\mathcal{E}_1 = \{E_1, E_2, E_3, E_4\} \ \text{ and } \ \mathcal{E}_2 = \{E_4, E_2, E_1, E_3\}.$$

We used the sequence of coefficients

$$\alpha_i \ = \ \frac{1}{1+k}, \ \text{where } k = ((i-1) \text{ div } s) + 1. \tag{5.36}$$

Remark. If it is not stated otherwise then all the remaining plots of this text display a history of a joint probability distribution $Q_{(i)}$ with respect to an iteration step i (horizontal axis). The vertical axis refers to the probability values for one particular combination of random variables' values given by $X_1 = 1,\ X_2 = 1,\ X_3 = 2,\ X_4 = 1$. If it is stated that the *input set* is *consistent* then *input set* $\{P_{E_1}, P_{E_2}, P_{E_3}, P_{E_4}\}$ given in Table 3.1 was used. If it is stated that the *input set* is *inconsistent* then *input set* $\{P_{E_1}, P_{E_2}, P_{E_3}, P_{E_4}\}$ is given in Table 3.6.

The Figure 5.5 may be seen as a support for the next two conjectures.

Conjecture 5.2 (I1.CC,I2.CC, and I3.CC)
Let $\{\alpha_i\}$ be a monotonous sequence that obey properties 5.31, 5.32, 5.33, 5.34, and 5.35 then CC converges, the respective limit distribution does not depend on the ordering of input set and minimizes the I_1-aggregate.

The situation with convergence of LCC applied to *inconsistent input set* is more complicated. Since it may happen during computational process that for some $i = 1, 2, \ldots, j = (i-1) \bmod s + 1$:

$$P_{E_j} \not\ll Q_{(i)}^{E_j},$$

the convergence can not be generally guaranteed. However, we observe that in many cases *log-convex conservative modification* does converge. (**I1.LCC**).

Conjecture 5.3 (I2.LCC and I4.LCC) *If $\{\alpha_i\}$ is a monotonous sequence that obey properties 5.31, 5.32, 5.33, 5.34, 5.35 and LCC converges, then the respective limit distribution does not depend on the ordering of input set and minimizes the* I_2-aggregate.

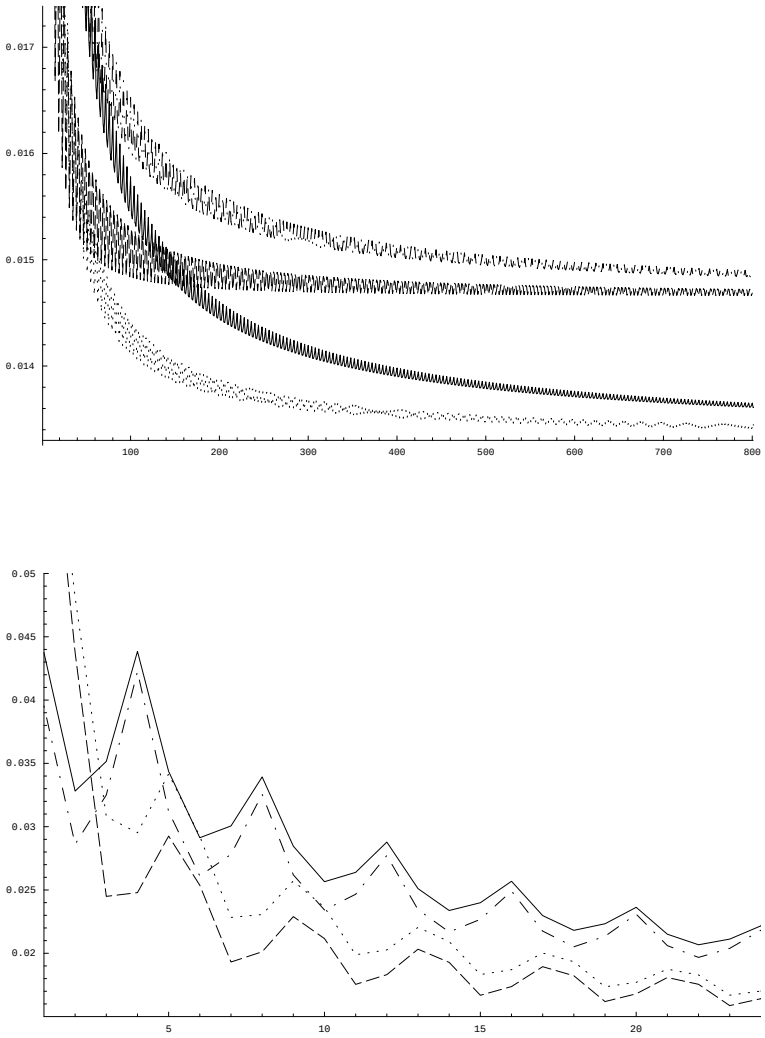


Figure 5.5: CC and LCC applied to *inconsistent input set*. The curves of first plot corresponds to the following procedures (the most upper curve is given first): **(1)** LCC on $\mathcal{E}_1$, **(2)** LCC on $\mathcal{E}_2$, **(3)** CC on $\mathcal{E}_1$, **(4)** CC on $\mathcal{E}_2$. Since the oscillations of the joint probability distributions' values within one cycle are significantly high and graphics resolution is limited, the curves are displayed as bands. Second plot displays details of first 24 iterations. **(1)** LCC on $\mathcal{E}_1$ is drawn by the full line, **(2)** LCC on $\mathcal{E}_2$ by the dotted line, **(3)** CC on $\mathcal{E}_1$ by the dash and dot line, and **(4)** CC on $\mathcal{E}_2$ by the dashed line.

The limit distributions of CC and LCC applied to *inconsistent input set* are generally different and minimizing different criterion (see Tables 5.7 and 5.9). The difference of the limit distributions is related to the fact that all distributions computed during the process of LCC in contrary to CC factorize with respect to a *generating class*. Therefore we can not expect that limit distributions of CC and LCC would minimize both criteria at the same time. Since we have conjectured that limit distributions of CC minimizes the I_1-*aggregate* it should not minimize the I_2-*aggregate* (**I4.CC**). Similarly, we have conjectured that if LCC converges then limit distributions of LCC minimizes the I_2-*aggregate*. Thus it should not minimize the I_1-*aggregate* (**I3.LCC**). It seems that for both methods the ordering of *input set* is not important with respect to limit distribution even if the *input set* is *inconsistent* (see Table 5.7 and Figure 5.5).

Table 5.7: Comparison of CC and LCC on *inconsistent input set*.

	CC		LCC	
	$\mathcal{E}_1$	$\mathcal{E}_2$	$\mathcal{E}_1$	$\mathcal{E}_2$
$Q_{(1000)}(1,1,2,1)$	0.013364	0.013317	0.016908	0.016860
$\psi_1(Q_{(1000)})$	0.484556	0.484562	0.507590	0.507530
$\psi_2(Q_{(1000)})$	0.586154	0.586151	0.560854	0.560854

5.8 Arithmetic mean of I_1-projections

The *method of iterative arithmetic averages* (abbreviated by AA) was proposed by F. Matúš. As we have shown it is in fact an instance of *GEM-algorithm*. Similar algorithm was used by Chen [9] for obtaining *maximum likelihood estimates* in case of double sampling with misclassification. Looking to the definition of method (formula 5.4) one can see that it is based on iterating arithmetic averages of I_1-projections to $\mathcal{S}_{E_j}, j = 1, 2, \ldots, s$. Results of several experiments with this modification were already published in [24].

[**G1.AA**] *Arithmetic average* of probability distribution factorizing with respect to $\mathcal{E}$ does not generally result into a probability distribution *factorizing* with respect to $\mathcal{E}$. Therefore an iterated joint probability distributions $Q_{(i)}, i = 1, 2, \ldots$ do not generally factorize with respect to $\mathcal{E}$.

[**C2.AA** and **I2.AA**] It follows from the fact that probability distributions $Q_{(i)}, i = 1, 2, \ldots$ are computed by *arithmetic average* that resulting probability distributions does not depend on the ordering of the *input set*.

It was proved by F. Matúš (see [32, Consequence 1]) that the AA method converges

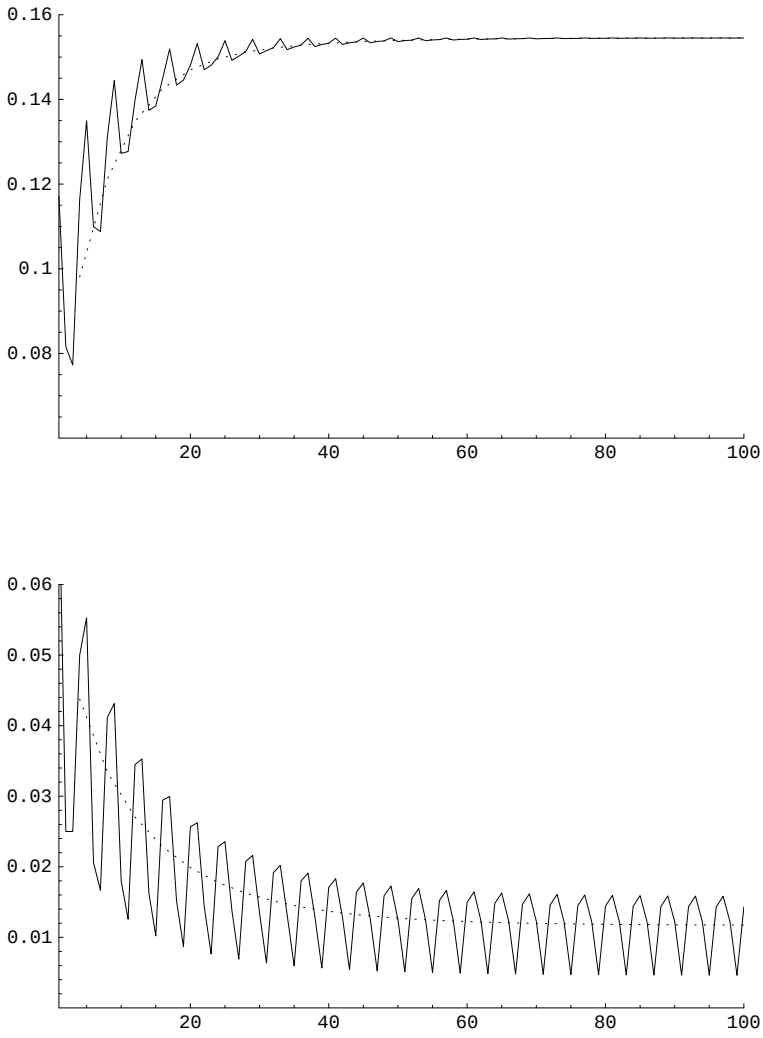


Figure 5.6: Arithmetic averages on *consistent* and *inconsistent input*, respectively. Full line joins I_1-projections to $\mathcal{S}_{\{E_1\}}, \mathcal{S}_{\{E_2\}}, \mathcal{S}_{\{E_3\}}$, and $\mathcal{S}_{\{E_4\}}$ computed in every cycle. Dotted line joins their *arithmetic averages*.

(**I1.AA** and **C1.AA**) and minimizes

$$\sum_{j=1}^{s} w_j \cdot I(\pi_{\mathcal{S}_{\{E_j\}}} Q \parallel Q),$$

which we have shown to be equivalent to I_1-*aggregate* (Lemma 5.1). (**I3.AA**). Note, that in the case of a *consistent input set* distribution minimization of I_1-*aggregate* requires the limit distribution to belong to $\mathcal{S}_\mathcal{E}$. However, having performed several computational experiments (see Figure 5.7) we conjecture that, generally, limit distributions of AA and regular IPFP are different (**C3.AA**).

5.9 Normalized geometric mean of I_2-projections

The *method of iterative geometric averages* (defined by formula 5.5 and abbreviated by GA) was proposed in [32]. If some distributions factorize with respect to $\mathcal{E}$ then their *geometric average* factorizes with respect to $\mathcal{E}$ as well. The property **G1.GA**) is an immediate consequence of this fact and Theorem 3.3. Thus comparing GA with AA we can see that an advantage of GA is the fact that if $Q_{(0)}$ factorize with respect to the set $\mathcal{E}$ then the iterated joint probability distributions $Q_{(i)}, i = 1, 2, \ldots$ factorize with respect to $\mathcal{E}$ as well.

[**I1.GA**, **I4.GA** and **C1.GA**] If GA is applied to an *inconsistent input set* then its convergence can not be generally guaranteed since during the computational process it may happen that for some $i = 1, 2, \ldots$ that $P_{E_j} \not\ll Q^{E_j}$, where $j = (i-1) \bmod s + 1$. Since the joint probability distribution is not defined the procedure fails. Otherwise, the method converges and the limit distribution equals to a Q minimizing I_2-*aggregate*. It was proved by F. Matúš (see [32, Consequence 2]) that the method minimizes

$$\sum_{j=1}^{s} w_j \cdot I(Q \parallel \pi_{\mathcal{S}_{\{E_j\}}} Q),$$

which we have shown to be equivalent to I_2-*aggregate* (Lemma 5.2).

The following theorem proves that if the *input set* is *consistent* and the initial probability distribution $Q_{(0)}$ factorizes with respect to $\mathcal{E}$ then the limit distributions of GA and IPFP are equivalent.

Theorem 5.2 (C3.GA) *If there exist a probability distribution $Q \in \mathcal{S}_\mathcal{E}, Q \ll Q_{(0)}$ and $Q_{(0)} \in \mathcal{R}_\mathcal{E}$ then sequence of probability distributions computed by GA converges to*

$$Q^* = \pi_{\bar{\mathcal{R}}_\mathcal{E}} Q = \pi_{\mathcal{S}_\mathcal{E}} Q_{(0)}.$$

Proof.
In the case of a *consistent input set* the minimum of $\sum_{j=1}^{s} I(Q_{(i)}^{E_j} \parallel P_{E_j})$ equals to zero.

$$I(P \parallel Q) \geq 0 \quad \text{and} \quad (\, I(P \parallel Q) = 0 \;\Leftrightarrow\; P = Q\,).$$

Therefore

$$\text{for } j = 1, 2, \ldots, s: \quad Q^{*E_j} = P_{E_j}$$

which can be rewritten as $Q^* \in \mathcal{S}_\mathcal{E}$. Since $Q^* \in \bar{\mathcal{R}}_\mathcal{E}$ it follows from Theorem 2.9 that

$$Q^* = \pi_{\bar{\mathcal{R}}_\mathcal{E}} Q.$$

It is a direct consequence of Theorem 3.5 that Q^* is equivalent to the limit distribution of IPFP, i.e.

$$Q^* = \pi_{\mathcal{S}_\mathcal{E}} Q_{(0)}.$$

□

Remark. Note that this proof suppose that $Q_{(0)} \in \mathcal{R}_\mathcal{E}$ and thus a proof of the property **C3.GA** for $Q_{(0)} \notin \mathcal{R}_\mathcal{E}$ is missing.

[**C2.GA** and **I2.GA**] Since the method is based on the geometric average of I_1-*projections* to $\mathcal{S}_{\{\mathcal{E}_i\}}, i = 1, 2, \ldots, s$ the respective limit distributions can not depend on the ordering of *input set*.

5.10 GEM-algorithm

Inspired by personal communication with G. Kleiter, we have proposed an application of *GEM-algorithm* (abbreviated by GEM) to the problem of *inconsistent input set* in [40]. The method is defined by formula 5.6. GEM constructs a joint probability distribution having its marginals "closed" to the marginals of a given *input set* generated by a *generating class* $\mathcal{E}$. Resulting probability distribution is required to factorize with respect to $\mathcal{F} = \{F_k, k = 1, 2, \ldots, t\}$. The application of results proved for the general *GEM-algorithm* is enabled by the fact that the distributions of *inconsistent input set* $\{P_{E_j}, j = 1, 2, \ldots, s\}$ can be manipulated as if they were estimated from *incomplete data* (see Section 5.4 for details). *GEM-algorithm* was defined by formula 5.6 and abbreviated by GEM. Since the algorithm definition (formula 5.6) may be hard to understand we will comment upon it.

Within the *expectation step* a joint probability distribution $R_{(i-1)}$ is estimated using distributions from *input set* $\{P_{E_j}, j = 1, 2, \ldots, s\}$ and the joint probability distribution $Q_{(i-1)}$ from the previous step of the algorithm. For $l = 1, 2, \ldots, s$:

$$R_{(i)} \quad = \quad \sum_{k=1}^{s} w_k \cdot \pi_{\mathcal{S}_{\{E_k\}}} Q_{(i-1)}$$

In the *maximization step* marginals $R^{F_j}_{(i-1)}$, $j = 1, 2, \ldots t$ will be used. These marginals create a *consistent input set*. Since they were estimated from marginals constituting a given, possibly *inconsistent, input set* $\{P_{E_j}, j = 1, 2, \ldots, s\}$ we will call them *adapted marginals*. In the *maximization step* a probability distribution $Q_{(i)}$ factorizing with respect to *generating class* $\mathcal{F}$ is searched for. We define for $l = 1, 2, \ldots, t$:

$$\mathcal{S}'_{\{F_l\}} \quad = \quad \{P \in \mathcal{P} : P^{F_l} = R^{F_l}_{(i)}\}.$$

The *maximization step* is realized by t consecutive steps of IPFP with initial distribution $Q_{(i,0)}$ equal to $Q_{(i-1)}$ and *input set* consisting of *adapted marginals*. For $j = 1, 2, \ldots, t$:

$$Q_{(i,j)} \quad = \quad \pi_{\mathcal{S}'_{\{F_j\}}} Q_{(i,j-1)}.$$

The distribution that is passed to the next *expectation step* is $Q_{(i)} = Q_{(i,t)}$.
Remark. Note that if $\mathcal{F} = \{V\}$ then GEM is equivalent to AA.

It is a direct consequence of the properties of the general *GEM-algorithm*[14] that its instance defined by Definition 5.3 converges.

Theorem 5.3 (I1.GEM)
The GEM method defined by formula 5.6 converges to a probability distribution $Q^* \in \mathcal{P}$.

For a detail discussion on convergence of the *GEM-algorithm* see Wu [44].
[**C1.GEM**] Theorem 5.3 is a consequence of Theorem 5.3 since *consistent input set* is just a special case of *input set*.

Conjecture 5.4 (I3.GEM) *The GEM method defined by formula 5.6 converges to a probability distribution* $Q^* = Q$ *that minimizes* $\psi_1(Q)$ *for all* $Q \in \mathcal{P}$.

Note that for $\mathcal{F} = \{V\}$ the conjecture holds since GEM reduces to AA for which property I1.AA was proved to be valid.

Lemma 5.3 (G1.GEM)
If $Q_{(0)} \in \mathcal{R}_\mathcal{E}$ *and* $Q_{(i)}$ *is computed by GEM then for all* $i = 1, 2, \ldots : Q_{(i)} \in \mathcal{R}_\mathcal{E}$.

Proof. The *E-step* of the GEM algorithm adjust only the marginals that are fitted in the *M-step*. Let us use mathematical induction to prove the lemma. It is assumed that $Q_{(0)} \in \mathcal{R}_\mathcal{E}$. The induction step

$$Q_{(i-1)} \in \mathcal{R}_\mathcal{E} \quad \Rightarrow \quad Q_{(i)} \in \mathcal{R}_\mathcal{E}$$

follows from *inheritance of factorizability* for IPFP (Theorem 3.3) since the *M-step* consists of s consequent steps of IPFP. Therefore it follows that all distributions computed by GEM belong to $\mathcal{R}_\mathcal{E}$. □

The following conjecture is based on assertion of Theorem 5.3 and Conjectures 5.4 and 5.1.

Conjecture 5.5 *Let* $Q_{(0)} \in \mathcal{R}_\mathcal{E}$ *then GEM method defined by formula 5.6 converges to a probability distribution*

$$Q^* \in \pi_{\mathcal{T}_1} Q_{(0)}.$$

Conjecture 5.4 states that $Q^* \in \mathcal{T}_1$. Theorem 5.3 states that if $Q_{(0)} \in \mathcal{R}_\mathcal{E}$ then $Q^* \in \bar{\mathcal{R}}_\mathcal{E}$ as well. If Conjecture 5.1 holds then set $\mathcal{T}_1$ is an *additively constrained set* for which it holds

$$\pi_{\mathcal{T}_1} Q_{(0)} = \mathcal{T}_1 \cap \bar{\mathcal{R}}_\mathcal{E},$$

that is unique (Theorem 2.9).

Conjecture 5.6 (C3.GEM) *If input set* $\{P_1, \ldots, P_s\}$ *is strongly consistent, i.e.* $\mathcal{S}_\mathcal{E} \neq \emptyset$ *and there exist a probability distribution* $Q \in \mathcal{S}_\mathcal{E}$ *such that* $Q \ll Q_{(0)}$ *then the sequence of probability distributions computed by GEM converges to*

$$Q^* = \pi_{\bar{\mathcal{R}}_\mathcal{E}} Q = \pi_{\mathcal{S}_\mathcal{E}} Q_{(0)}.$$

The conjecture is a consequence of Conjecture5.5. In the case of a *consistent input set* the minimum of $\sum_{j=1}^{s} I(P_{E_j} \parallel Q_{(i)}^{E_j})$ equals to zero.

$$I(P \parallel Q) \geq 0 \quad \text{and} \quad (\ I(P \parallel Q) = 0 \ \Leftrightarrow \ P = Q)\,.$$

Therefore

$$\text{for } j = 1, 2, \ldots, s: \quad Q^{*E_j} = P_{E_j}$$

which can be rewritten as $Q^* \in \mathcal{S}_{\mathcal{E}}$. Since $Q^* \in \bar{\mathcal{R}}_{\mathcal{E}}$ it follows from Theorem 2.9 that

$$Q^* = \pi_{\bar{\mathcal{R}}_{\mathcal{E}}} Q.$$

It is a direct consequence of Theorem 3.5 that Q^* is equivalent to the limit distribution of IPFP, i.e.

$$Q^* = \pi_{\mathcal{S}_{\mathcal{E}}} Q_{(0)}.$$

[**C2.GEM**] would be a direct consequence of Conjecture 5.6 because Q^* is independent of the ordering of an *input set*.

[**I2.GEM**] would be a direct consequence of Conjecture 5.5 since Q^* would be independent of the ordering of an *input set*.

From Table 5.9 it can be seen that, in spite the fact that both AA (with $w_j = 1/4, j = 1, 2, 3, 4$) and GEM converge to the distribution having the same value of the I_1-aggregate, the limit distributions are different. Note that the limit distribution of the sequence created by the *GEM-algorithm* factorizes with respect to $\mathcal{E}$ while the limit distribution of the *method of iterative arithmetic averages* does not.

5.11 Convergence rate and computational complexity

In Tables 5.8 and 5.9 convergence rates of proposed procedures are compared. As a stopping criterion the *total variance* $|Q_{(4i-4)} - Q_{(4i)}| \leq 10^{-6}$ was used. In the column inscribed with i the number of performed cycles is given. Probability distributions were monitored for $x = \{X_1 = 1, X_2 = 1, X_3 = 2, X_4 = 1\}$. In the case of *consistent input set* values that differ less than 10^{-5} from I_1-projection to $\mathcal{S}_{\mathcal{E}}$ are highlighted, while in the case of *inconsistent input set* values of I_1-aggregate or I_2-aggregate that differ less than 10^{-5} from their optimal values are highlighted. As a reference the results achieved by optimization package LANCELOT minimizing I_1-aggregate and I_2-aggregate are used. For details see Appendix B.

Tables 5.8 and 5.9 and Figures 5.8 and 5.8 present examples of typical convergence character of the proposed methods. The results given in Table 5.9 may serve as arguments in favor of conjectures stating that

- [**I4.AA**] AA does not minimize I_2-*aggregate*,
- [**I3.GA**] GA does not minimize I_1-*aggregate*, and
- [**I4.GEM**] GEM does not minimize I_1-*aggregate*.

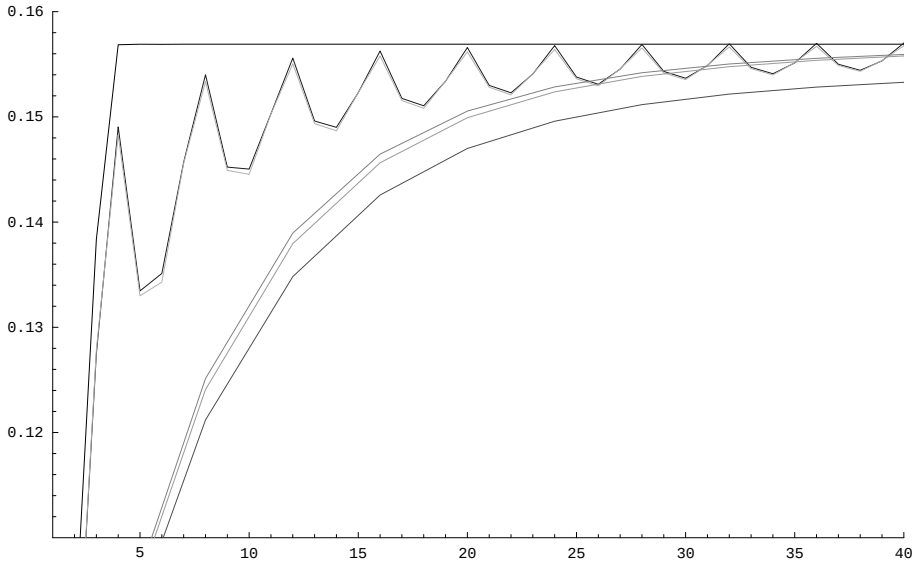


Figure 5.7: Comparison of methods applied to the *consistent input set*. The methods are (the uppermost first): IPFP (fast convergence), CC, LCC (oscillating and almost identical), GA, GEM (both monotonically increasing and almost identical) , AA (monotonically increasing but convergent to a different value).

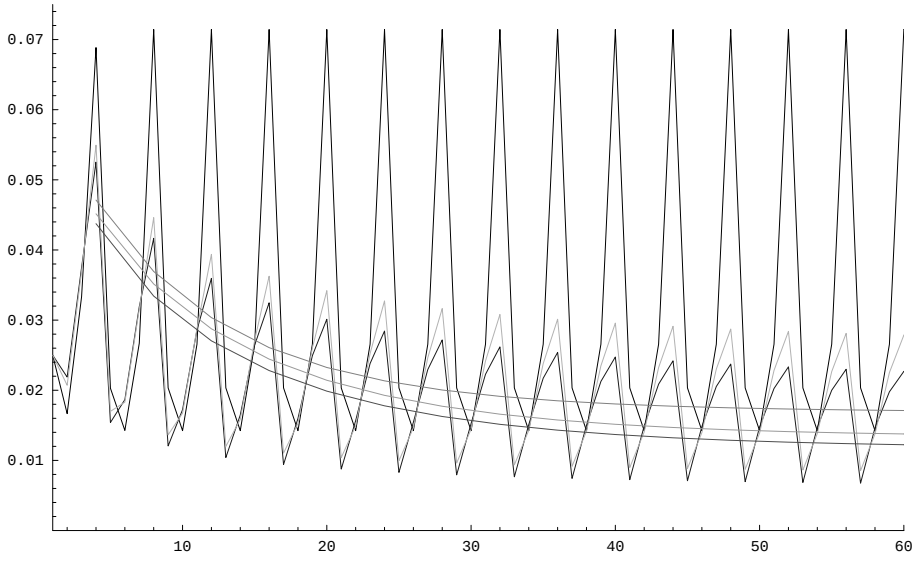


Figure 5.8: Comparison of methods applied to the *inconsistent input set*. The methods are (the uppermost first): IPFP, LCC, CC (all three are oscillating), GA, GEM, AA (all these three are monotonically decreasing each converges to a different value).

Algorithm	uniform $Q_{(0)}$		extremal $Q_{(0)}$	
	i	$Q_{(4i)}(x)$	i	$Q_{(4i)}(x)$
IPFP	4	**0.15691**	84	**0.17283**
CC	483	0.15688	5178	0.17188
LCC	283	0.15685	5131	0.17491
AA	40	0.15453	50	0.16141
GA	39	**0.15691**	358	**0.17283**
GEM	40	**0.15691**	5299	0.17284

Table 5.8: Comparisons of methods on *consistent input set*

Algorithm	uniform $Q_{(0)}$			
	i	$Q_{(4i)}(x)$	I_1-aggreg.	I_2-aggreg.
IPFP	5*	a cycle†		
CC	1694	0.013331	**0.484552**	0.586322
LCC	285	0.014706	0.507880	**0.560860**
AA	68	0.011678	**0.484547**	0.586708
GA	49	0.016847	0.507444	**0.560853**
GEM	82	0.013272	**0.484544**	0.586707
LANCELOT‡	14	0.000002	**0.484546**	0.586709
LANCELOT§	20	0.085970	0.507447	**0.560849**

Table 5.9: Comparison of methods on *inconsistent input set*

A serious difficulty which complicates practical usability of CC and LCC methods is their low convergence rate no matter if they are applied to *consistent* or *inconsistent input sets*. Convergence rates of AA, GA, and GEM are similar. Observe that they are slower than IPFP on a *consistent input set*. Speed of convergence depends on the type of an *generating class*. Similarly to IPFP applied to a *consistent input set*, the convergence rate of AA, GA, and GEM is slower if *generating class* is not *decomposable*. Furthermore, the convergence may further slow down if the *initial probability distribution* is extremal, i.e. far away from the uniform probability distribution (see Figure 5.9).

Figures 5.10 and 5.11 display function history of I_1-*aggregate* and I_2-*aggregate* during AA, GA, and GEM methods. Observe that for all three methods I_1-*aggregate* is monotonically decreasing, while I_2-*aggregate* is monotonically decreasing only for GA.

Computational complexity of the proposed methods is highly related with **G1** property.

*Number of cycles after which the procedure stabilizes in a limit cycle

†The values in the cycle were $\{0.020338, 0.014240, 0.026564, 0.071430\}$

‡Minimization using I_1-aggregate as the criterion

§Minimization using I_2-aggregate as the criterion

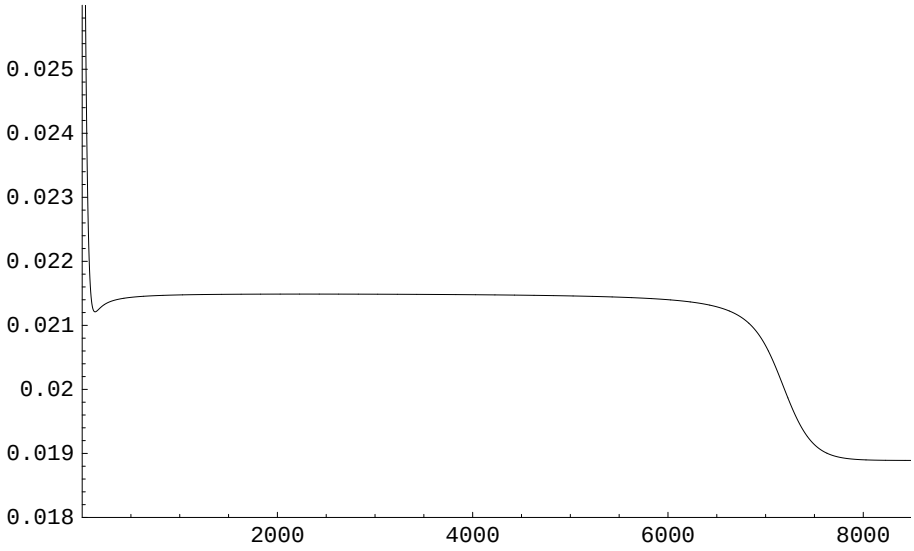


Figure 5.9: GEM algorithm applied to an extremal initial distribution .

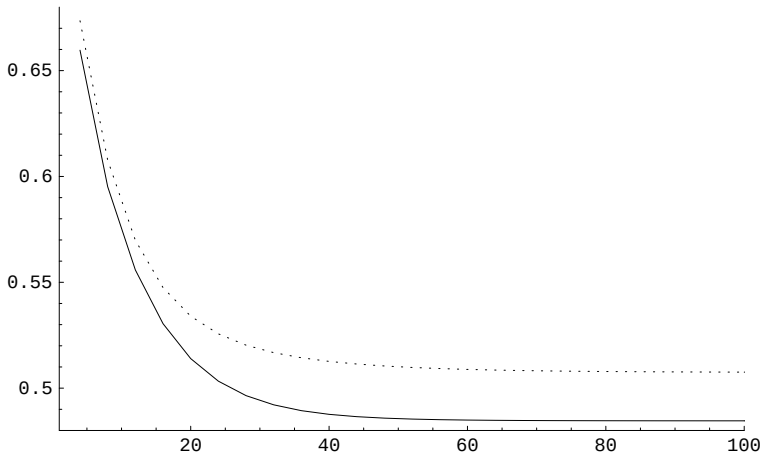


Figure 5.10: Comparison of function history of I_1-*aggregate* during processes of AA, GEM (almost identical - both are displayed by full line), and GA (dotted line) on the *inconsistent input set*.

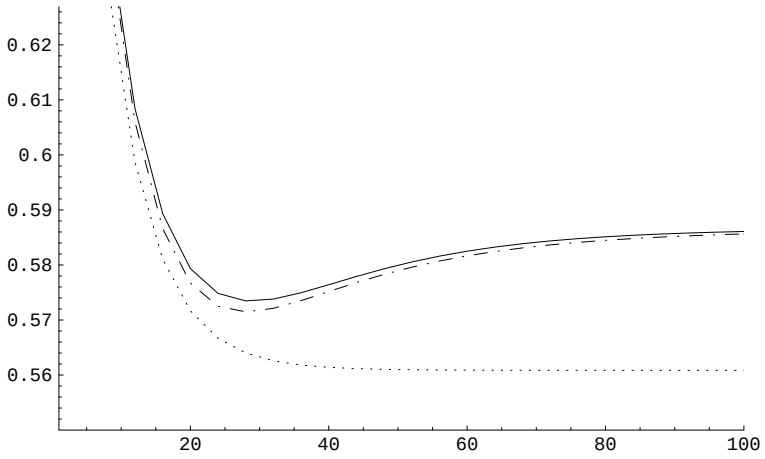


Figure 5.11: Comparison of function history of I_2-*aggregate* during processes of AA (full line), GA (dotted line) and GEM (dash and dot line) on the *inconsistent input set*.

We can read from the Table 5.2 that for IPFP, LCC, GA, and GEM it holds that

$$Q_{(0)} \in \mathcal{R}_{\mathcal{E}} \quad \Rightarrow \quad i = 1, 2, \ldots : \; Q_{(i)} \in \mathcal{R}_{\mathcal{E}},$$

while for CC and AA even if $Q_{(0)} \in \mathcal{R}_{\mathcal{E}}$ then the joint probability distributions of the next steps need not factorize with respect to $\mathcal{E}$. It can be utilized for an effective implementation of the procedure. Thus, if the initial joint probability distribution $Q_{(0)}$ belongs to $\mathcal{R}_{\mathcal{E}}$, all these four algorithms (IPFP, LCC, GA, and GEM) can be implemented so that the space requirements and computational complexity of one iterative step are proportional to the size $|\mathbb{X}^E|$ of the largest factor $\phi_E(x^E)$ of $Q_{(i)}(x) = \prod_{E\in\mathcal{E}} \phi_E(x^E)$.

For the IPFP algorithm a *space-saving implementation* was proposed by R. Jiroušek [22]. His implementation is based on assertion of Theorem 3.1. This computational scheme can be extended so that it can be exploited for LCC, GA and GEM methods as well. Three extensions to the Jiroušek's scheme are necessary:

- An effective way of performing **multiplication** of two distributions that *factorize* with respect to the same set $\mathcal{E}$. It is straightforward as it can be seen below.

$$P = \prod_{E\in\mathcal{E}} \phi_E(x^E) \;\text{ and }\; Q = \prod_{E\in\mathcal{E}} \xi_E(x^E) \quad \Leftrightarrow \quad P \cdot Q \;=\; \prod_{E\in\mathcal{E}} \Big(\phi_E(x^E) \cdot \xi_E(x^E)\Big)\,.$$

- In order to get a probability distribution the operation of **normalization** must be performed within LCC and GA algorithms. The normalization constant can be effectively computed by a subsequent marginalization. Suppose that a potential

$$f(x) = \prod_{E\in\mathcal{E}} \phi_E(x^E)$$

is to be normalized in order to get a probability distribution. Similarly to Theorem 3.1 we define $\mathcal{F} \geq \mathcal{E}$ to be *decomposable*. It can be ordered so that it satisfies *Running intersection property*. Suppose that $\mathcal{F} = \{F_1, \ldots, F_t\}$ is such an order, i.e.

$$\forall l = 2, \ldots, t \quad \exists k \quad (1 \leq k < l) \quad ((F_l \cap \bigcup_{m=1}^{l-1} F_m) \subset F_k),$$

$B_k = \left(\cup_{l=1}^{k-1} F_l\right) \cup \left(\cup_{l=k+1}^{t} F_l\right)$, and for $k = t, t-1, \ldots, 1$ corresponding class $\mathcal{E}_k$ is defined recursively as $\mathcal{E}_k = \{E \subseteq F_k, E \notin \mathcal{E}_{k+1}\}$. Then the normalization constant can be computed by

$$\begin{aligned}
c &= \sum_x f(x) \\
&= \sum_x \prod_{E \in \mathcal{E}} \phi_E(x^E) \\
&= \sum_{x^{F_t \setminus (E_t \cap B_t)}} \sum_{x^{F_{t-1} \setminus (E_{t-1} \cap B_{t-1})}} \cdots \sum_{x^{E_1}} \prod_{E \in \mathcal{E}} \phi_E(x^E) \\
&= \sum_{x^{F_t \setminus (E_t \cap B_t)}} \prod_{E_t \in \mathcal{E}_t} \phi_{E_t}(x^{E_t}) \cdot \sum_{x^{F_{t-1} \setminus (E_{t-1} \cap B_{t-1})}} \prod_{E_{t-1} \in \mathcal{E}_{t-1}} \phi_{E_{t-1}}(x^{E_{t-1}}) \cdot \ldots \\
&\quad \ldots \cdot \sum_{x^{E_1}} \prod_{E_1 \in \mathcal{E}_1} \phi_{E_1}(x^{E_1})
\end{aligned}$$

- In the *maximization step* of GEM, instead of joint probability distribution $R_{(i)}$ computed in the *expectation step*, only its marginals $R_{(i)}^{E_j}$ are used. Therefore the *expectation step* of GEM can be performed effectively using computational scheme of Lauritzen and Spiegelhalter [28].

Addition of two probability distribution factorizing with respect to the same class $\mathcal{E}$ does not generally result into a probability distribution factorizing with respect to $\mathcal{E}$. Therefore the computational scheme presented above can not be used for CC and AA methods. Consequently, the space required (and computational complexity) during one iterative step is proportional to the size $|\mathbb{X}|$ of a joint probability distribution, which implies that CC and AA are not applicable to *large* problems, i.e. if the joint probability distributions is defined on more than 40 dichotomic random variables.

We conclude the comparisons by presenting the opinion that the most favorable method from the presented methods applicable to *inconsistent knowledge integration* seems to be GEM since it best corresponds to the intention to design a method that is an extension of IPFP for *inconsistent input set*. Due to a possible *space-saving implementation* its computational complexity is proportional to to the size $|\mathbb{X}^E|$ of the largest factor. A weakness of the method is a relatively slow convergence for *generating classes* that are not *decomposable* and if an initial probability distribution is far away from the uniform probability distribution. In this case, some of *gradient optimization methods* would converge substantially faster. But, note that general *optimization methods* are not tractable if the joint probability distribution is *large*. Another disadvantage of classical optimization methods comparing with GEM is that two requirements should be applied gradually, e.g. first, marginals minimizing ψ_1 are searched for and than a distribution from $\mathcal{R}_{\mathcal{E}}$ having these marginals is computed. While during the GEM algorithm both requirements are refined together.

6 Conclusions

The basic goal of the thesis was the design and description of methods applicable to the integration of *knowledge* represented by a *set of probability distributions*. We have presented five methods that under some assumptions converge even if the *input set* is *inconsistent*. The theory of *additively constrained sets* and *multiplicatively constrained sets* has enabled us to describe and confirm the methods' properties. The fundamental advantage of proposed iterative methods is that, in comparison with conventional optimization methods, they do not require complicated calculations. This makes it possible to apply them to problems of high dimensionality. We have implemented all these methods and performed a number of computational experiments. A systematic summary of the achieved results is given. One of the proposed methods, an instance of the GEM-algorithm, can be considered to be the most appropriate method for the task of *inconsistent knowledge integration*.

In the case of *consistent input set iterative proportional fitting procedure* possess the best convergence rate of all presented iterative methods. Moreover, in the case of a *decomposable generating class*, there exist orderings of the *input set* that guarantee IPFP convergence within one cycle. The thesis extends the family of *decomposable generating classes* for which IPFP converges within one cycle regardless the ordering. On the other side, we have shown that there exist *decomposable generating classes* such that (in spite of Haberman's conjecture of 1974) IPFP need not converge even within two cycles.

The theory of *iterative methods* proposed for *probabilistic knowledge integration*, which we have presented, makes a self-contained and closed exposure. Iterative methods can, for instance, play the role of algorithms for learning parameters of probabilistic systems. At the present time, the learning algorithms receive great attention. *Expert* or *knowledge systems* with a probabilistic engine have been predicted to be applied in many domains in the near future.

6.1 The main original results

The first task was to prepare the theory that enable us to design and study iterative methods for integration of *knowledge* represented by a *set of probability distributions*. In Chapters 2 and 3 we presented self-contained theory of two basic types of probability distributions' sets: *additively constrained sets* and *multiplicatively constrained sets*. Afterwards, we applied

the theory to the study of iterative methods. Its application was based on the fact that all of the proposed methods iterate operations of *I-projections* to *additively constrained sets*. We have done original proofs of some theorems and lemmata of this part, let us make reference at least to Theorem 2.9. Some of the fundamental theorems of this part are rewritten from different papers [10, 11, 32, 33]. We had to adapt these theorems in order to give them in the form appropriate for the context. Some proofs were not available, thus we had to work them out afresh.

The second basic part (Chapter 4) was devoted to the study of *iterative proportional fitting procedure* in the case of *decomposable generating classes*. There are several original results in this section. We proved a sufficient condition for IPFP convergence within one cycle (Theorem 4.6), which extends the family of *decomposable generating classes* such that for any ordering of the sets in the class IPFP converges within one cycle. An important result is the counterexample to S. Haberman conjecture [17], that for all *decomposable generating classes* IPFP always converges within two *cycles*. It is known that orderings of the *input set* meeting *running intersection property* guarantee IPFP convergence within one cycle We proved a sufficient condition that extends the class of orderings that guarantee IPFP convergence within one cycle Theorem 4.7. We have proposed two conjectures. The first one offers a full characterization of chordal graphs corresponding to generating classes that guarantee IPFP convergence within one cycle. The second one is the analogy of the first one for the convergence within two cycles.

The third task that we solved (Chapter 5) was to find a method that converge to a joint probability distribution even if the *input set* is *inconsistent*. The basic requirements are that the marginals of the resulting distribution are close to the input distributions and the resulting distribution does not depend on the ordering of the *input set*. An application to the *consistent input sets* should result in the same distribution as *iterative proportional fitting procedure*. Results of empirical study of five *iterative methods* for the *probabilistic knowledge integration* are presented. We have proposed the instance of the GEM-algorithm, which we claim to be the most appropriate of all the methods described. Results proven for the general GEM-algorithm were exploited to prove convergence properties of the proposed method. This method meets all the demanded properties and since it results in the joint probability distribution that *factorizes* with respect to the *generating class* it can be considered to be an extension of *iterative proportional fitting procedure* to the case of of the not necessarily *consistent input set*.

6.2 Open questions

Open questions concern two areas. At first, the conjecture concerning full characterization of *decomposable generating classes* that guarantee IPFP convergence within one cycle (Conjecture 4.1) needs to be proven. Another open problem from the area of IPFP convergence is a necessary and sufficient condition of IPFP convergence within finite number of cycles (Conjecture 4.2).

Five methods for *inconsistent knowledge integration* were proposed in Chapter 5. Table 5.2 summarizes properties of these methods. One can see that several methods' properties are not proven. Some of the missing assertions need not require much effort to be proven but some proofs can be rather complicated. Furthermore, we strove for results that are not restricted to

strictly positive probability distribution. Some assertions concerning convergence properties of iterative methods were not achieved in this general case due to problems with probability distributions containing zeroes. Thus, convergence proofs of some proposed methods are still missing.

Bibliography

[1] N. I. M. Gould A. R. Conn and Ph. L. Toint. A globally convergent augmented lagrangian algorithm for optimization with general constraints and simple bounds. *SIAM J. Numer. Anal.*, 28:545–572, 1991.

[2] C. Beeri, R. Fagin, D. Maier, and M. Yannakakis. On the desirability of acyclic database schemes. *J. of the ACM*, 30:479–514, 1983.

[3] C. Berge. *Graphs and Hypergraphs.* North Holland, Amsterdam, 1976.

[4] Y. M. M. Bishop, S. E. Feinberg, and P. W. Holland. *Discrete Multivariate Analysis.* The MIT Press, Cambridge, MA, 1975.

[5] D.T. Brown. A note on aproximations to discrete probability distributions. *Infrom. and control*, 2:386–392, 1959.

[6] S. F. Buck. A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *J. Roy. Statist. Soc. Ser. B*, 22:302–306, 1960.

[7] N. N. Čencov. *Statistical Decision Rules and Optimal Inference.* Translations of Mathematical Monographs. American Mathematical Society, Providence, 1982.

[8] Y. Censor and S. A. Zenios. *Parallel Optimization: Theory, Algorithms, and Applications.* Oxford University Press, New York, 1997.

[9] T. T. Chen. Log-linear models for categorized data with misclassification and double sampling. *Journal of the American Statistical Association*, 74:481–488, 1979.

[10] I. Csiszár. *I*-divergence geometry of probability distributions and minimization problems. *Ann. Prob.*, 3(1):146–158, 1975.

[11] I. Csizsár and G. Tusnády. Information geometry and alternating minimization procedures. *Statistics & Decisions*, Supplement Issue No. 1:205–237, 1984.

[12] A. D'Atri and M. Moscarini. On hypergraph acyclicity and graph chordality. *Information Processing Letters*, 29(5):271–274, 1988.

[13] W. E. Deming and F. F. Stephan. On a least square adjustment of a sampled frequency table when the expected marginal totals are known. *Ann. Math. Statist.*, 11:427–444, 1940.

[14] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 39:1–38, 1977.

[15] R. Fagin. Degrees of acyclicity for hypergraphs and relational database schemes. *Journal of the Association for Computing Machinery*, 30(3):514–550, 1983.

[16] S. E. Fienberg. An iterative procedure for estimation in contigency tables. *Ann. Math. Statist.*, 41(3):907–917, 1970.

[17] S. Haberman. *The Analysis of Frequency Data.* The University of Chicago Press, Chicago and London, 1974.

[18] S. J. Haberman. Log-linear models for frequency tables derived by indirect observation: maximum likelihood equations. *Annals of Statistics*, 2(5):911–924, 1974.

[19] S. J. Haberman. Product models for frequency tables involving indirect observation. *Annals of Statistics*, 5(6):1124–1147, 1977.

[20] P. Hájek, T. Havránek, and R. Jiroušek. *Uncertain Information Processing in Expert Systems.* CRC Press, Boca Raton, Ann Arbor, London, Tokyo, 1992.

[21] C.T. Ireland and S. Kullback. Contigency tables with given marginals. *Biometrika*, 55(1):179–188, 1968.

[22] R. Jiroušek. Solution of the marginal problem and decomposable distributions. *Kybernetika*, 27(5):403–412, 1991.

[23] R. Jiroušek. *Probabilistic Methods in Artificial Intelligence (DrSc Thesis).* Institute of Information Theory and Automation, Prague, 1993.

[24] R. Jiroušek and J. Vomlel. Inconsistent knowledge integration in a probabilistic model. In *Proceedings of Workshop "Mathematical Models for handling partial knowledge in A.I."*, New York, 1995. Plenum Publ. Corp.

[25] H. G. Kellerer. Verteilungsfunktionen mit gegeben Marginalverteilungen. *Z. Wahrsch. Verw. Gebiete*, 3:247–270, 1964.

[26] S. Kullback. An information-theoretic derivation of certain limit relations for a stationary Markov chain. *J. SIAM Control*, 4(3):454–459, 1966.

[27] S. Kullback and M. A. Khairat. A note on minimum discrimination information. *Ann. Math. Statist.*, 37:279–280, 1966.

[28] S. L. Lauritzen. The EM algorithm for graphical association models with missing data. Technical report, Aalborg University, 1991.

[29] S. L. Lauritzen. *Graphical Models.* Clarendon Press, Oxford, 1996.

[30] S. L. Lauritzen, T. P. Speed, and K. Vijayan. Decomposable graphs and hypergraphs. *J. Austral. Math. Soc. (Series A)*, 36:12–29, 1984.

[31] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data.* John Willey & Sons, New York, 1987.

[32] F. Matúš. On alternating minimization of I-divergence and averages of I-projections. *(unpublished manuscript)*, 1997.

[33] F. Matúš. On I-divergence geometry of affine and log-affine sets. *(unpublished manuscript)*, 1999.

[34] G. J. McLachlan and T. Krishnan. *The EM algorithm and extensions.* Wiley series in probability and statistics. John Wiley & Sons, Inc., 1997.

[35] D. B. Rubin. Inference with missing data. *Biometrika*, 63:581–592, 1976.

[36] D. B. Rubin. *Multiple Imputation for Nonresponse in Surveys.* Wiley, New York, 1987.

[37] L. Rüschendorf. Convergence of the iterative proportional fitting procedure. *Ann. Statist.*, 23(4):1160–1174, 1995.

[38] R. Sundberg. Maximum likelihood theory for incomplete data from an exponential family. *Scandinavian Journal of Statistics*, 1:49–58, 1974.

[39] I. Vajda. *Theory of Statistical Inference and Information.* Kluwer Academic Publishers, Dordrecht, The Netherlands, 1989.

[40] J. Vomlel. Statistical methods for probabilistic model parameter estimation from incomplete data and their application to the marginal problem. In *Proceedings of Workshop on Uncertainty Processing*, Prague, Czech Republic, 1997. University of Economics.

[41] N.N. Vorobe'v. Consistent families of measures and their extensions. *Theor. Probab. Appl.*, 7:147–163, 1962.

[42] S. S. Wilks. *Mathematical Statistics.* John Wiley & Sons, New York, London, 1962.

[43] S. Wolfram. *The Mathematica Book, Fourth Edition.* Cambridge University Press, 1998.

[44] C. F. Jeff Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.

[45] K. Yosida. *Functional Analysis.* Springer-Verlag, Berlin, Heidelberg, New York, 1980.

[46] W. I. Zangwill. *Nonlinear Programming: A Unified Approach.* Prentice Hall, Englewood Cliffs, 1969.

A Mathematica Source Code

The reason we have chosen to perform all computations in *Mathematica* is based on *Mathematica*'s ability to set an acquired accuracy of results. Thus, without much additional effort we were able to switch between computations in rational arithmetic and computations in real numbers with the accuracy of results being under the control.

During the computation the program keeps track of which digits in a result can be affected by unknown digits in the input. In the result not affected digits are displayed only. If the number of displayed digits was too low to make any conclusion about where does a method converge, we simply increased the precision of program and performed the computation with the increased value again. Since *Mathematica* always overestimates a potential effect of unknown digits we can be sure that the displayed numbers are correct. Sometimes, this fact, e.g. in case of GEM algorithm on extremal initial distribution, requires a very high precision to be set.

In the appendix, source code of two programs written in *Mathematica* can be found. First program was used to reject Haberman's conjecture. The computations were performed in rational arithmetics. Since numerators and denominators of the resulting fractions are enormously huge integers, the values were displayed as rounded real numbers at the very end of algorithm.

Second program was used to make experiments with *iterative methods*. The parameters that have impact on the accuracy of results are `ND` which gives the initial number of digits used to store numbers and `$MaxExtraPrecision` which is an global option of *Mathematica* defining how many additional digits it uses for its internal calculations.

Both files are available via world wide web. The addresses are given at the beginning of the files. They can be simply used to verify results of the thesis. Having the files downloaded and saved to the computer they can be directly open in *Mathematica* and just by performing menu item `Kernel -> Evaluation -> Evaluate Notebook` the results can be achieved. In order to be able to get results for different methods on different input set and with different initial distributions the switches are given in the second file. The required combination can be chosen by putting out the comments' symbols `(* ... *)` around the corresponding options,

while the other should be put into the comments part conversely.

It should be noted that some of computations may take dozens of minutes. However, it would be satisfactory for many real problems to use fixed-precision computations, which would speed up the process substantially.

A.1 Program used to reject Haberman's conjecture

```
(* This program is available at:                          *)
(*  http://www.utia.cas.cz/user_data/vomlel/motylek.nb *)
(*  http://lisp.vse.cz/~vomlel/motylek.nb              *)

(* Definition of probability distributions' values    *)

P1 = {{{5/100,8/100},{5/100,11/100}},{{13/100,17/100},{31/100,10/100}}};
P2 = {{{3/100,13/100},{7/100,17/100}},{{3/100,10/100},{17/100,30/100}}};
P3 = {{{7/100,5/100},{11/100,7/100}},{{23/100,8/100},{27/100,12/100}}};
P4 = {{{12/100,4/100},{11/100,13/100}},{{8/100,5/100},{13/100,34/100}}};
P5 = {{{2/100,4/100},{11/100,12/100}},{{10/100,14/100},{20/100,27/100}}};
P6 = {{{3/100,7/100},{11/100,19/100}},{{5/100,15/100},{25/100,15/100}}};

(* Definition of computational steps of IPFP based on marginalizations,
  divisions, and multiplications of lists *)

Marg1=Function[x,
 Inner[Times,Inner[Times,Inner[Times,Inner[Times,Inner[Times,
 Transpose[x,{8,1,7,2,3,4,5,6}],{1,1}],{1,1}],{1,1}],{1,1}],{1,1}]];
Cond1=Function[x,
 Transpose[Outer[Divide,x,Marg1[x]],{1,2,3,4,5,6,7,8,2,4,5}]];
IPFP1=Function[x,
 Transpose[Outer[Times,Cond1[x],P1],{1,2,3,4,5,6,7,8,2,4,5}]];

Marg2=Function[x,
 Inner[Times,Inner[Times,Inner[Times,Inner[Times,Inner[Times,
 x,{1,1}],{1,1}],{1,1}],{1,1}],{1,1}]];
Cond2=Function[x,
 Transpose[Outer[Divide,x,Marg2[x]],{1,2,3,4,5,6,7,8,1,2,3}]];
IPFP2=Function[x,
 Transpose[Outer[Times,Cond2[x],P2],{1,2,3,4,5,6,7,8,1,2,3}]];

Marg3=Function[x,
 Inner[Times,Inner[Times,Inner[Times,Inner[Times,Inner[Times,
 Transpose[x,{8,7,1,2,6,3,5,4}],{1,1}],{1,1}],{1,1}],{1,1}],{1,1}]];
Cond3=Function[x,
 Transpose[Outer[Divide,x,Marg3[x]],{1,2,3,4,5,6,7,8,3,4,6}]];
IPFP3=Function[x,
```

```
 Transpose[Outer[Times,Cond3[x],P3],{1,2,3,4,5,6,7,8,3,4,6}]];

Marg4=Function[x,
 Inner[Times,Inner[Times,Inner[Times,Inner[Times,Inner[Times,
 Transpose[x,{1,2,8,7,6,5,3,4}],{1,1}],{1,1}],{1,1}],{1,1}],{1,1}]];
Cond4=Function[x,
 Transpose[Outer[Divide,x,Marg4[x]],{1,2,3,4,5,6,7,8,1,2,7}]];
IPFP4=Function[x,
 Transpose[Outer[Times,Cond4[x],P4],{1,2,3,4,5,6,7,8,1,2,7}]];

Marg5=Function[x,
 Inner[Times,Inner[Times,Inner[Times,Inner[Times,Inner[Times,
 Transpose[x,{8,1,2,3,4,5,6,7}],{1,1}],{1,1}],{1,1}],{1,1}],{1,1}]];
Cond5=Function[x,
 Transpose[Outer[Divide,x,Marg5[x]],{1,2,3,4,5,6,7,8,2,3,4}]];
IPFP5=Function[x,
 Transpose[Outer[Times,Cond5[x],P5],{1,2,3,4,5,6,7,8,2,3,4}]];

Marg6=Function[x,
 Inner[Times,Inner[Times,Inner[Times,Inner[Times,Inner[Times,
 Transpose[x,{1,8,2,7,6,5,4,3}],{1,1}],{1,1}],{1,1}],{1,1}],{1,1}]];
Cond6=Function[x,
 Transpose[Outer[Divide,x,Marg6[x]],{1,2,3,4,5,6,7,8,1,3,8}]];
IPFP6=Function[x,
 Transpose[Outer[Times,Cond6[x],P6],{1,2,3,4,5,6,7,8,1,3,8}]];

(* 2*6 = 12 iterative steps of IPFP *)

Q0  = Array[1/256&,{2,2,2,2,2,2,2,2}];
Q1  = IPFP1[Q0];  Unprotect[Q0];  Q0=.;
Q2  = IPFP2[Q1];  Unprotect[Q1];  Q1=.;
Q3  = IPFP3[Q2];  Unprotect[Q2];  Q2=.;
Q4  = IPFP4[Q3];  Unprotect[Q3];  Q3=.;
Q5  = IPFP5[Q4];  Unprotect[Q4];  Q4=.;
Q6  = IPFP6[Q5];  Unprotect[Q5];  Q5=.;
Q7  = IPFP1[Q6];  Unprotect[Q6];  Q6=.;
Q8  = IPFP2[Q7];  Unprotect[Q7];  Q7=.;
Q9  = IPFP3[Q8];  Unprotect[Q8];  Q8=.;
Q10 = IPFP4[Q9];  Unprotect[Q9];  Q9=.;
Q11 = IPFP5[Q10]; Unprotect[Q10]; Q10=.;
Q12 = IPFP6[Q11]; Unprotect[Q11]; Q11=.;
```

```
(* Total variances between marginals of 12th iteration    *)
(* probability distribution and the input set distributions *)

N[Apply[Plus,Flatten[Abs[
 Transpose[Outer[Subtract,Marg1[Q12],P1],{1,2,3,1,2,3}]]]],20]
N[Apply[Plus,Flatten[Abs[
 Transpose[Outer[Subtract,Marg2[Q12],P2],{1,2,3,1,2,3}]]]],20]
N[Apply[Plus,Flatten[Abs[
 Transpose[Outer[Subtract,Marg3[Q12],P3],{1,2,3,1,2,3}]]]],20]
N[Apply[Plus,Flatten[Abs[
 Transpose[Outer[Subtract,Marg4[Q12],P4],{1,2,3,1,2,3}]]]],20]
N[Apply[Plus,Flatten[Abs[
 Transpose[Outer[Subtract,Marg5[Q12],P5],{1,2,3,1,2,3}]]]],20]
N[Apply[Plus,Flatten[Abs[
 Transpose[Outer[Subtract,Marg6[Q12],P6],{1,2,3,1,2,3}]]]],20]

(* Results *)
\!\(2.549584300946522044805794746326 6`20*^-12\)
\!\(1.053348474506592115241443438469 8`20*^-10\)
\!\(1.801831983110316014511344305046 3`20*^-13\)
\!\(1.053348474506592115241443438469 8`20*^-10\)
\!\(2.549584300946522044805794746326 6`20*^-12\)
0
```

A.2 Implementation of proposed methods

```
(* This program is available at:                                  *)
(*  http://www.utia.cas.cz/user_data/vomlel/ipfp-modif.nb        *)
(*  http://lisp.vse.cz/~vomlel/ipfp-modif.nb                     *)
<<Graphics`MultipleListPlot`;

(* Control options *)

ND= 500;(* Starting precision, i.e. number of digits *)
$MaxExtraPrecision=300; (* Number of extra digits to be stored
                           during the computation *)
(*$MinPrecision = ND; *)(* Forced minimal computational precision *)
(*$MaxPrecision = ND; *)(* Forced maximal computational precision *)
NC=5000;                (* Number of cycles *)
i=3;(* Index of the value that will be monitored in list L *)
TOL = 1/1000000;
(* Absolute difference in total variance between distributions
   at the end of two consequent cycles used to stop computation *)

Consistent="Yes";
(* Consistent = "No";*)
(* Algorithm = "IPFP"; *)
(* Algorithm = "CC";*)
(* Algorithm = "LCC";*)
(* Algorithm = "AA";*)
(* Algorithm = "GA"; *)
Algorithm = "GEM";
(* StartPoint = "Uniform";*)
StartPoint = "Extremal";

(* Function that defines coeficients for CC and LCC algorithms *)
alpha=Function[x,1/(1+x)];

(* Definition of probability distributions' values as structured lists *)
Switch [Consistent,
    "Yes",{P1 = {{89/288,43/144},{7/36,19/96}};
           P2 ={{17/96,47/144},{43/288,25/72}};
           P3 ={{31/144,1/9},{67/144,5/24}};
           P4 ={{15/32,5/36},{61/288,13/72}}},
    "No", {P1 ={{1/10,3/10},{2/10,4/10}};
           P2 ={{7/10,1/10},{1/10,1/10}};
   P3 ={{2/10,1/10},{2/10,5/10}};
           P4 ={{3/10,3/10},{3/10,1/10}}}];
```

```
(* Definition of computational steps of IPFP based on marginalizations,
  divisions, and multiplications of lists *)
Marg1= Function[x,Inner[Times,Inner[Times,x,{1,1}],{1,1}]];
Cond1= Function[x,
        Transpose[Outer[Divide,x,Marg1[x]],{1,2,3,4,1,2}]];
IPFP1= Function[{x,y},
        Transpose[Outer[Times,Cond1[x],y],{1,2,3,4,1,2}]];
Marg2= Function[x,
        Inner[Times,Inner[Times,Transpose[x,{3,1,2,4}],{1,1}],{1,1}]];
Cond2= Function[x,
        Transpose[Outer[Divide,x,Marg2[x]],{1,2,3,4,2,3}]];
IPFP2= Function[{x,y},
        Transpose[Outer[Times,Cond2[x],y],{1,2,3,4,2,3}]];
Marg3= Function[x,
        Inner[Times,Inner[Times,Transpose[x,{3,4,1,2}],{1,1}],{1,1}]];
Cond3= Function[x,
        Transpose[Outer[Divide,x,Marg3[x]],{1,2,3,4,3,4}]];
IPFP3= Function[{x,y},
        Transpose[Outer[Times,Cond3[x],y],{1,2,3,4,3,4}]];
Marg4= Function[x,
        Inner[Times,Inner[Times,Transpose[x,{1,3,4,2}],{1,1}],{1,1}]];
Cond4= Function[x,
        Transpose[Outer[Divide,x,Marg4[x]],{1,2,3,4,1,4}]];
IPFP4= Function[{x,y},
        Transpose[Outer[Times,Cond4[x],y],{1,2,3,4,1,4}]];

(* I_1 aggregate *)
I1A=Function[{Q0,P1,P2,P3,P4},
     Apply[Plus,Flatten[Transpose[Outer[Times,P1,Log[Transpose[
      Outer[Divide,P1, Marg1[Q0]],{1,2,1,2}]]],{1,2,1,2}]]]+
     Apply[Plus,Flatten[Transpose[Outer[Times,P2,Log[Transpose[
      Outer[Divide,P2, Marg2[Q0]],{1,2,1,2}]]],{1,2,1,2}]]]+
     Apply[Plus,Flatten[Transpose[Outer[Times,P3,Log[Transpose[
      Outer[Divide,P3, Marg3[Q0]],{1,2,1,2}]]],{1,2,1,2}]]]+
     Apply[Plus,Flatten[Transpose[Outer[Times,P4,Log[Transpose[
      Outer[Divide,P4, Marg4[Q0]],{1,2,1,2}]]],{1,2,1,2}]]]];

(* I_2 aggregate *)
I2A=Function[{Q0,P1,P2,P3,P4},
     Apply[Plus,Flatten[Transpose[Outer[Times, Marg1[Q0],Log[Transpose[
      Outer[Divide, Marg1[Q0],P1],{1,2,1,2}]]],{1,2,1,2}]]]+
     Apply[Plus,Flatten[Transpose[Outer[Times,Marg2[Q0],Log[Transpose[
      Outer[Divide,Marg2[Q0],P2 ],{1,2,1,2}]]],{1,2,1,2}]]]+
     Apply[Plus,Flatten[Transpose[Outer[Times, Marg3[Q0],Log[Transpose[
      Outer[Divide, Marg3[Q0],P3],{1,2,1,2}]]],{1,2,1,2}]]]+
```

```
     Apply[Plus,Flatten[Transpose[Outer[Times, Marg4[Q0],Log[Transpose[
      Outer[Divide, Marg4[Q0],P4],{1,2,1,2}]]],{1,2,1,2}]]]];

(* Total variance between Q1 and Q2 *)
VAR=Function[{Q1,Q2}, Apply[Plus,Flatten[Abs[Transpose[
     Outer[Subtract,Q1,Q2],{1,2,3,4,1,2,3,4}]]]]];

(* Definition of a starting probability distribution *)
Switch [StartPoint,
        "Uniform",  Q0= N[Array[1/2^4&,{2,2,2,2}],ND] ,
        "Extremal", Q0= N[
                    {{{{1/100000,1/100000},{1/100000,1/100000}},
                      {{1/100000,1/100000},{1/100000,1/100000}}},
                      {{{1/100000,1/100000},{1/100000,1/100000}},
                      {{1/100000,1/100000},{1/100000,99985/100000}}}},ND]
       ];

(* Initialization of variables *)
L={};(* List monitoring values of Q[i] for every step j *)
LA={};(* List monitoring values of Q[i] for every cycle *)
j=1;
k=1;
differ=1;

(* The core  *)
While[k<=NC && differ > TOL,
 QP=Q0;
 Q1=IPFP1[QP,P1];QT=Flatten[Q1];L=Append[L,{j,QT[[i]]}];j=j+1;
 Switch[Algorithm,
"IPFP",QP=Q1,
"CC", {a=alpha[k];QP=(1-a)*QP+ a*Q1},
"LCC",{a=alpha[k];QP=QP^(1-a)*Q1^a;QP=(1/Apply[Plus,Flatten[QP]])*QP}];
 Q2=IPFP2[QP,P2];QT=Flatten[Q2];L=Append[L,{j,QT[[i]]}];j=j+1;
 Switch[Algorithm,
"IPFP",QP=Q2,
"CC", {a=alpha[k];QP=(1-a)*QP+ a*Q2},
"LCC",{a=alpha[k];QP=QP^(1-a)*Q2^a;QP=(1/Apply[Plus,Flatten[QP]])*QP}];
 Q3=IPFP3[QP,P3];QT=Flatten[Q3];L=Append[L,{j,QT[[i]]}];j=j+1;
 Switch[Algorithm,
"IPFP",QP=Q3,
"CC", {a=alpha[k];QP=(1-a)*QP+ a*Q3},
"LCC",{a=alpha[k];QP=QP^(1-a)*Q3^a;QP=(1/Apply[Plus,Flatten[QP]])*QP}];
 Q4=IPFP4[QP,P4];QT=Flatten[Q4];L=Append[L,{j,QT[[i]]}];j=j+1;
 Switch[Algorithm,
  "IPFP",QN=Q4,
```

```
  "CC", {a=alpha[k];QN=(1-a)*QP+ a*Q4},
  "LCC",{a=alpha[k];QP=QP^(1-a)*Q4^a; QN=(1/Apply[Plus,Flatten[QP]])*QP},
  "AA", QN=(1/4)*(Q1+Q2+Q3+Q4),
  "GA", {QP=Transpose[Outer[Times,
              Transpose[Outer[Times,Q1^(1/4),Q2^(1/4)],{1,2,3,4,1,2,3,4}],
              Transpose[Outer[Times,Q3^(1/4),Q4^(1/4)],{1,2,3,4,1,2,3,4}]],
              {1,2,3,4,1,2,3,4}];
         QN=(1/Apply[Plus,Flatten[QP]])*QP},
  "GEM",{QP= (1/4)*(Q1+Q2+Q3+Q4);
         Q1=IPFP1[Q0,Marg1[QP]];
         Q2=IPFP2[Q1,Marg2[QP]];
         Q3=IPFP3[Q2,Marg3[QP]];
         QN=IPFP4[Q3,Marg4[QP]]}
 ];
 QT=Flatten[QN];LA=Append[LA,{j-1,QT[[i]]}];
 differ = VAR[Q0,QN];Unprotect[Q0];Q0=.;
 Q0=QN;
 Unprotect[QN];QN=.; Unprotect[Q1];Q1=.; Unprotect[Q2];Q2=.;
 Unprotect[Q3];Q3=.; Unprotect[QP];QP=.; Unprotect[QT];QT=.;
 k=k+1;
];

(* Resulting total variance between Q(k-1) and Q(k-2) used to stop  *)
N[differ,6]

(* Number of cycles *)
k-1

(* Resulting probability distribution *)
N[Q0,5]

(* I_1 aggregate of resulting probability distribution *)
N[I1A[Q0,P1,P2,P3,P4],6]

(* I_2 aggregate of resulting probability distribution *)
N[I2A[Q0,P1,P2,P3,P4],6]

MultipleListPlot[L,LA,AxesOrigin -> {1,0.00},PlotJoined -> True,
                 Background ->RGBColor[1, 1, 1], SymbolShape->None,
                 ImageSize->{400,300}, PlotRange ->{{1,(j-1)},{0.00,0.20}}]
```

B Source code for optimization in AMPL

In order to be able to compare proposed methods with standard optimization algorithms we have decided to use AMPL, *A Modeling Language for Mathematical Programming*. AMPL is a comprehensive and powerful algebraic modeling language for linear and nonlinear optimization problems, in discrete or continuous variables, developed at Bell Laboratories. To solve a problem its optimization model is formulated and then the model can be solved by any appropriate solver.

There are many solvers for which interfaces to AMPL have been constructed. Since the problem we wanted to solve is constrained and nonlinear we decided to use LANCELOT. LANCELOT is a globally convergent augmented Lagrangian algorithm for optimization with general constraints and simple bounds [1]. There is a friendly and generally accessible web interface to the program, available at: `http://204.178.31.5:8001/ampl.pl`. Its use is simple. To get the distributions we refer to in our thesis the reader should select (uncomment) the criteria in the following file, paste it into the *Model and data* window, select a solver (in our case `lancelot`), and send the input to the server. We have used the program to minimize the I_1 and I_2 aggregates.

```
# This file is available at:
#  http://www.utia.cas.cz/user_data/vomlel/optim-ampl.txt
#  http://lisp.vse.cz/~vomlel/optim-ampl.txt
set P;
set R;
set Marg;
set MVal {Marg,R} within P;

param l {j in P};
```

```
param u {j in P};
param m {Marg,R};

var X {j in P} <= 1;
var MX {k in Marg, v in R} = sum {j in MVal[k,v]} X[j];
var I1X = sum {k in Marg, v in R}
          (m[k,v]*log(m[k,v]/(sum {j in MVal[k,v]} X[j])));
var I2X = sum {k in Marg, v in R} ((sum {j in MVal[k,v]}
            X[j])*log((sum {j in MVal[k,v]} X[j])/m[k,v]));

minimize sum_i_1_agg:  I1X;
#minimize sum_i_2_agg:  I2X;

subject to prob_values {j in P}:  0.1e-20 <= X[j] <= 1;
subject to prob_distr:   sum {j in P} X[j] = 1;

data;  ############ DATA STARTS HERE ############
set P := 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16;
set R := 1 2 3 4;
set Marg := 1 2 3 4;
set MVal[1,1]:= 1  5  9 13;
set MVal[1,2]:= 2  6 10 14;
set MVal[1,3]:= 3  7 11 15;
set MVal[1,4]:= 4  8 12 16;
set MVal[2,1]:= 1  3  5  7;
set MVal[2,2]:= 2  4  6  8;
set MVal[2,3]:= 9 11 13 15;
set MVal[2,4]:=10 12 14 16;
set MVal[3,1]:= 1  2  3  4;
set MVal[3,2]:= 5  6  7 8;
set MVal[3,3]:= 9 10 11 12;
set MVal[3,4]:=13 14 15 16;
set MVal[4,1]:= 1  2  9 10;
set MVal[4,2]:= 3  4 11 12;
set MVal[4,3]:= 5  6 13 14;
set MVal[4,4]:= 7  8 15 16;

param m: 1   2   3   4 :=
#each line stands for one marginal distribution
    1   0.1 0.3 0.2 0.4
    2   0.7 0.1 0.1 0.1
    3   0.2 0.1 0.2 0.5
    4   0.3 0.3 0.3 0.1;

let {j in P} X[j] := 0.0625;
```

```
solve;
display _varname, _var;
```

Index