Integrating inconsistent data in a probabilistic model

Jiří Vomlel[*] [**]

[*] *Institute of Information Theory and Automation*
Pod vodárenskou věží 4,
182 08 Praha 8 - Libeň, Czech Republic
vomlel@utia.cas.cz

[**] *Laboratory for Intelligent Systems*
University of Economics,
Ekonomická 957,
148 01 Praha 4 - Kunratice, Czech Republic

ABSTRACT. *In this paper we discuss the process of building a joint probability distribution from an input set of low-dimensional probability distributions. Since the solution of the problem for a consistent input set of probability distributions is known we concentrate on a setup where the input probability distributions are inconsistent. In this case the iterative proportional fitting procedure (IPFP), which converges in the consistent case, tends to come to cycles. We propose a new algorithm that converges even in inconsistent case. The important property of the algorithm is that it can be efficiently implemented exploiting decomposability of considered distributions.*

KEYWORDS: *Probabilistic models, Iterative proportional fitting, Knowledge integration, Inconsistent data*

1. Introduction

In this paper we discuss the process of integrating contradicting data using the framework of classical probability. We define *knowledge integration* as a process of building a *joint probability distribution* Q from a set of low-dimensional probability distributions $\mathcal{P} = \{P_1, \ldots, P_k\}$, $k \in \mathbb{N}$ and from an *initial joint probability distribution* Q_0. To model a chosen domain we will use discrete random variables X_i indexed by natural numbers from $V = \{1, \ldots, n\} \subset \mathbb{N}$. Each low-dimensional probability distribution $P_j, j = 1, \ldots, k$ is defined on variables $\{X_\ell\}_{\ell \in E_j}, E_j \subseteq V$. *Knowledge integration* should yield a joint probability distribution $Q(X_1, \ldots, X_n)$ that would embody available knowledge about the chosen domain. The general problem of the construction of probability distributions with given marginals has a long history start-

ing with the papers by Hoeffding [HOE 40], Fréchet [FRE 51], Kellerer [KEL 64], and Vorobev [VOR 62].

If there exists a probability distribution $Q(X_1, \ldots, X_n)$ whose marginals equal to low-dimensional probability distributions $P_i \in \mathcal{P}, i = 1, \ldots, k$ then we say that probability distributions in set $\mathcal{P}$ are *consistent*. Otherwise we say that they are *inconsistent*.

EXAMPLE 1. — In Table 1 we give an example (borrowed from [JIR 95b]) of an input set of three distributions $\mathcal{P} = \{P_1(X_1, X_2), P_2(X_2, X_3), P_3(X_1, X_3)\}$.

Table 1. *Input set of distributions* $\{P_1, P_2, P_3\}$

$P_j, j = 1, 2$	$X_{j+1} = 0$	$X_{j+1} = 1$
$X_j = 0$	$\frac{1}{2} - \varepsilon$	ε
$X_j = 1$	ε	$\frac{1}{2} - \varepsilon$

P_3	$X_3 = 0$	$X_3 = 1$
$X_3 = 0$	ε	$\frac{1}{2} - \varepsilon$
$X_3 = 1$	$\frac{1}{2} - \varepsilon$	ε

If $\varepsilon = \frac{4}{20}$ then there exist probability distributions $Q(X_1, X_2, X_3)$ whose marginals equal to probability distributions P_1, P_2, and P_3. For example, the probability distribution given in Table 2 has its corresponding marginals equal to P_1, P_2, and P_3. Therefore the probability distributions in set $\mathcal{P}$ are consistent.

Table 2. *Joint probability distribution* $Q(X_1, X_2, X_3)$

	$X_2 = 0$		$X_2 = 1$	
Q	$X_3 = 0$	$X_3 = 1$	$X_3 = 0$	$X_3 = 1$
$X_1 = 0$	$\frac{3}{20}$	$\frac{3}{20}$	$\frac{1}{20}$	$\frac{3}{20}$
$X_1 = 1$	$\frac{3}{20}$	$\frac{1}{20}$	$\frac{3}{20}$	$\frac{3}{20}$

If $\varepsilon = \frac{3}{20}$ then there does not exist any distribution $Q(X_1, X_2, X_3)$ whose marginals equal to probability distributions P_1, P_2, and P_3, i.e., the probability distributions in set $\mathcal{P}$ are inconsistent.

We will use small letter x_i to denote a value of variable X_i. The set of all possible values of variable X_i will be denoted $\mathbb{X}_i$. We will use X to denote multidimensional random variable $(X_1, \ldots, X_n)$, x will denote a value of X and $\mathbb{X}$ the set of all values of X. For a set $A \subseteq V$ the symbol X^A denotes the multidimensional variable $\{X_i\}_{i \in A}$, x^A denotes a value of X^A, and Q^A denotes the marginal distribution of Q on variables X^A,i.e., $Q^A(x^A) = \sum_{x\prime : x\prime^A = x^A} Q(x\prime)$. Recall, that each low-dimensional probability distribution $P_j, j = 1, \ldots, k$ from $\mathcal{P}$ is defined on variables $X^{E_j}, E_j \subseteq V$.

DEFINITION 2. — *Let P, Q be two probability distributions. If $\{x \in \mathbb{X}, P(x) > 0\} \supseteq \{y \in \mathbb{X}, Q(y) > 0\}$ then P is dominated by Q (denoted as $P \ll Q$).*

REMARK 3. — In the sequel it will be convenient to have defined $0 \cdot \log 0 = 0$ and $0 \cdot \log \frac{0}{0} = 0$.

REMARK 4. — To simplify the notation we will omit the set $\mathbb{X}$ of values of X in the symbol for summation $\sum_{x \in \mathbb{X}}$ and write $\sum_x$ instead.

The rest of the paper will be organized as follows. We briefly introduce the information theoretical approach to knowledge integration in Section 2. In Section 3 we describe the iterative proportional fitting procedure (IPFP). In Section 4 we will define the I-aggregate - a criteria that measures the distance between the marginals of a joint distribution and the distributions of an input set. In Section 5 we propose a new algorithm that can be considered to be a generalization of IPFP for inconsistent data. We conclude the paper by a discussion of the properties of the proposed algorithm and by presenting two experiments. The proofs of the lemmas and the theorem are in the appendix.

2. The information theoretical approach to knowledge integration

The probability distribution given in Table 2 has its marginals equal to P_1, P_2, P_3 defined in Table 1 for $\varepsilon = \frac{4}{20}$. However, this is not the only distribution defined for three binary variables X_1, X_2, X_3 having its corresponding marginals equal to P_1, P_2, and P_3. There is a whole class of probability distributions that have this property. $\mathcal{S}_j$ will denote the set of probability distributions having their marginal equal to probability distribution P_j, i.e., $\mathcal{S}_j = \{Q : Q^{E_j} = P_j\}$. $\mathcal{S}$ or $\mathcal{S}_\mathcal{E}$ will be used to denote the set of distributions having their marginals equal to probability distributions from the input set $\mathcal{P}$, i.e., $\mathcal{S}_\mathcal{E} = \{Q : Q^{E_1} = P_1, \ldots, Q^{E_k} = P_k\}$. It means that $\mathcal{S}_\mathcal{E} = \cap_{j=1}^{k} \mathcal{S}_j$.

For practical reasons, only one representative of the whole class $\mathcal{S}_\mathcal{E}$ is often chosen. A criterion that is most often used is Shannon entropy - an information measure introduced by Shannon in [SHA 48].

DEFINITION 5. — *Entropy*

$$H(P) \quad = \quad -\sum_{x \in \mathbb{X}} P(x) \log P(x)$$

The selection of the distribution that maximizes Shannon entropy can be justified by the fact that from all distributions in the class $\mathcal{S}_\mathcal{E}$ this distribution contains least additional information (see [HÁJ 92], Chapter 3.3). The principle of maximum entropy was first introduced as an inference procedure by Jaynes [JAY 57]. Shore and Johnson [SHO 80, JOH 83] prove there are four natural principles that determine that the function that should be maximized must be the Shannon entropy function. Also, Paris and Vencovská [PAR 90] showed that the maximum entropy inference process is the only inference process that satisfies natural principles of independence and consistency.

A generalization of the maximum entropy principle to cases when an estimate P of the unknown true distribution is known is the principle of minimum cross-entropy. Cross-entropy was introduced by Kullback and Leibler [KUL 51] as a divergence measure of two probability distributions. It is also known under several other names: Kullback-Leibler divergence, relative information, and I-divergence. We will refer to it as I-divergence.

DEFINITION 6. — *I-divergence*

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

REMARK 7. — *I-divergence* is not generally symmetric. $I(P \parallel Q) \geq 0$ 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 P.

In contrast to *I-divergence, total variance* satisfies the *measure properties*.

DEFINITION 8. — *Total variance*

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

I-divergence can be used to provide an upper bound on total variance. The following lemma is due to Kullback [KUL 66].

LEMMA 9. — *If P, Q are two probability distributions then the following inequality holds:*

$$|P - Q| \quad \leq \quad 2\sqrt{I(P \parallel Q)} \qquad [1]$$

I-divergence geometry, studied by Csiszár [CSI 75] and Čencov [ČEN 82] is an effective instrument enabling us to study iterative procedures based on the operations of I-projections to a set of probability distributions $\mathcal{Q}$. I-projection will be the core operation used in both iterative methods discussed in Section 3 and in Section 5.

DEFINITION 10. — *$Q^\star$ is an I-projection of a probability distribution P to a set of probability distributions $\mathcal{Q}$ if for all $Q \in \mathcal{Q}$ it holds that*

$$I(Q \parallel P) \quad \geq \quad I(Q^\star \parallel P) \ .$$

We will use $\pi(P, \mathcal{Q})$ to denote I-projection of P to $\mathcal{Q}$.

If the set $\mathcal{Q}$ is convex and compact[1] then there is a unique distribution $Q^\star \in \mathcal{Q}$ that maximizes $I(Q \parallel P)$. Minimization of I-divergence is strongly related to the

1. Note that $\mathcal{S}_\mathcal{E}$ is a convex and compact set.

principle of maximum entropy. We will later show (in Lemma 15) that under certain restrictions on probability distribution P the I-projection of P to $\mathcal{S}_\mathcal{E}$ corresponds to the distribution that has the maximal entropy from all distributions in $\mathcal{S}_\mathcal{E}$.

Let Q be a probability distribution such that $P_j \ll Q^{E_j}$. I-projection of Q to the set $\mathcal{S}_j$ can be computed [CSI 75] as

$$\pi(Q, \mathcal{S}_j)(x) \quad = \quad \begin{cases} 0, & \text{for } Q^{E_j}(x^{E_j}) = 0, \\ Q(x)\dfrac{P_j(x^{E_j})}{Q^{E_j}(x^{E_j})}, & \text{for } Q^{E_j}(x^{E_j}) > 0. \end{cases} \qquad [2]$$

3. Iterative proportional fitting procedure

In case of a consistent input set the knowledge integration task can be solved by means of the *Iterative Proportional Fitting Procedure* (IPFP), also known as the Iterative Proportional Scaling. IPFP dates back to 1940. It was proposed by W. E. Deming and F. F. Stephan [DEM 40] as a procedure for adjustment of frequencies in contingency tables.

In each step of the iterative proportional fitting procedure (IPFP) an I-projection of the distribution from previous step to a set $\mathcal{S}_j, j = 1, \ldots, k$ is computed. Projections are repeated until a convergence is reached[2]. A formal definition follows.

DEFINITION 11. — *Let $Q_{(0)}(X_1, \ldots, X_n)$ be an initial probability distribution such that for all its marginals $Q_{(0)}^{E_j}$ it holds that $P_j \ll Q_{(0)}^{E_j}$ for $j = 1, \ldots, k$. IPFP is the following algorithm:*

for $i = 0, 1, 2, \ldots$
  for $j = 1, \ldots, k$
    $Q_{(i\cdot k+j)} \;=\; \pi(Q_{(i\cdot k+j-1)}, \mathcal{S}_j)$,

REMARK 12. — If the input set contains k strictly positive distributions and the initial probability distribution is strictly positive then we can see (from formula [2]) that each step of IPFP corresponds to

$$Q_{(i\cdot k+j)} \quad = \quad Q_{(i\cdot k+j-1)} \cdot \frac{P_j^{E_j}}{Q_{(i\cdot k+j-1)}^{E_j}} \quad .$$

This makes IPFP an attractive procedure since it can be easily implemented.

The notion of *factorization* will play an important role in the characterization of the resulting distribution.

2. k consecutive steps $i \cdot k + 1, \ldots, (i + 1) \cdot k$ will be referred as a *cycle*.

DEFINITION 13. — *Probability distribution Q factorizes with respect to set $\mathcal{E} = \{E_1, \ldots, E_k\}$ if there exist nonnegative functions $\phi_{E_i} : \mathbb{X}^{E_i} \mapsto \mathbb{R}^+, i = 1, 2, \ldots, k$ (called potentials) such that for all $x \in \mathbb{X}$*

$$Q(x) \quad = \quad \prod_{E_i \in \mathcal{E}} \phi_{E_i}(x^{E_i}) \ .$$

The set of all distributions factorizing with respect to set $\mathcal{E}$ will be denoted $\mathcal{R}_\mathcal{E}$. Generally, it does not hold that if all probability distributions computed in a finite number of steps of IPFP factorize then also the limit distribution factorizes. Therefore we will use the closure of set $\mathcal{R}_\mathcal{E}$, defined as

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

LEMMA 14. — *If $\mathcal{S}_\mathcal{E} \neq \emptyset$ then $\mathcal{S}_\mathcal{E} \cap \overline{\mathcal{R}}_\mathcal{E}$ is unique, i.e., contains exactly one probability distribution.*

The following lemma can be used to characterize the limit probability distribution of a sequence of probability distributions computed by IPFP. It will be also used to characterize limit probability distributions of the new procedure proposed in Section 5.

LEMMA 15. — *Let $\mathcal{E} = \{E_1, \ldots, E_k\}$, $Q \in \overline{\mathcal{R}}_\mathcal{E}$, and $\mathcal{S}_\mathcal{E} = \{P : P^{E_1} = Q^{E_1}, \ldots, P^{E_k} = Q^{E_k}\}$. Then*

(a) $\forall R \in \mathcal{R}_\mathcal{E}, Q \ll R \implies Q = \pi(R, \mathcal{S}_\mathcal{E})$ and

(b) Q maximimizes entropy from all distributions in $\mathcal{S}_\mathcal{E}$.

If $Q_{(0)}$ factorizes with respect to $\mathcal{E}$ then all probability distributions computed within a finite number of iterations of IPFP factorize with respect to $\mathcal{E}$ as well. Consider $Q^\star$ that is the limit probability distribution of a sequence of probability distributions computed by IPFP. Since $Q^\star \in \overline{\mathcal{R}}_\mathcal{E}$ from Lemma 15 we know that $Q^\star$ maximizes entropy from all distributions having the marginals $\{P_1, \ldots, P_k\}$.

4. Distance to a given input set

In case of an *inconsistent* input set of probability distributions, IPFP tends to come in cycles. Nevertheless, we wish to get a representative probability distribution even in the case when the input set is inconsistent. We endeavor to find an iterative procedure that converges 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.

Empirical observations of Osherson et al. [OSH 97] support the basic idea underlying the *knowledge integration* process. They argue that a consistent probability model that is fitted as closely as possible to the inconsistent judgments of an informant will often be closer to actual relative frequencies than the raw judgments of the informant.

Assume that to every probability distribution P_j from input set $\mathcal{P}$ a nonnegative weight w_j is assigned such that $\sum_{j=1}^{k} w_j = 1$. The weights can be understood as credibility weights of the corresponding probability distributions. Alternatively, the weights can be proportional to the number of data vectors used to estimate the values of the corresponding probability distribution. If we have no preferences among the distributions we set $w_j = \frac{1}{k}, j = 1, \ldots, k$. Recall that by Q^{E_j} we denote marginal distribution of a joint probability distribution Q defined on variables $X_i, i \in E_j$, where $E_j \subseteq V$ is the index set of variables of $P_j \in \mathcal{P}$.

In Definition 16 we define an I-aggregate, which will be used to measure the distance between the marginals of the resulting distribution and the distributions from the input set.

DEFINITION 16. — *Let $\mathcal{P} = \{P_1, \ldots, P_k\}$ be the input set of distributions, Q be a joint probability distribution, and w_j be nonnegative weights summing up to 1.*
I-aggregate is the functional

$$\psi(Q \parallel P_1, \ldots, P_k) \quad = \quad \sum_{j=1}^{k} w_j \cdot I(P_j \parallel Q^{E_j}) \ .$$

REMARK 17. — Since $I(P_j \parallel Q^{E_j}) \geq 0$ with equality iff $P_j = Q^{E_j}$ it follows that $\psi(Q \parallel P_1, \ldots, P_k) \geq 0$ with equality iff $j = 1, \ldots, k : P_j = Q^{E_j}$. It implies that in case of a consistent input set $\mathcal{P} = \{P_1, \ldots, P_k\}$ minimization of I-aggregate ensures that the resulting distribution Q has the required marginals, i.e., $Q \in \mathcal{S}$.

Generally, the I-aggregate need not be minimized by a unique probability distribution. If there exists a probability distribution Q that minimize the I-aggregate then all probability distributions having the same marginals $Q^E, E \in \mathcal{E}$ minimize the corresponding aggregate as well. Furthermore, also distributions having different marginals $Q^E, E \in \mathcal{E}$ can be minimizers of I-aggregate with respect to a given input set $\mathcal{P} = \{P_1, \ldots, P_k\}$. We will denote the set of all probability distribution minimizing the I-aggregate by $\mathcal{T}$. In case of a *consistent input set* $\mathcal{P}$ it follows from Remark 17 that $\mathcal{T} = \mathcal{S}$.

Using the principle of maximal entropy, we will prefer the distribution from set $\mathcal{T}$ that has the maximal entropy, i.e. $Q^\star = \arg\max_{Q \in \mathcal{T}} H(Q)$. In the case of a consistent input set $\mathcal{T} = \mathcal{S}$, therefore $Q^\star$ is the same as the limit distribution of IPFP.

EXAMPLE 18. — Assume input set $\mathcal{P} = \{P_1(X_1, X_2), P_2(X_2, X_3), P_3(X_1, X_3)\}$ from Example 1. Take $\varepsilon = \frac{3}{20}$. In this case there is no distribution having distributions from $\mathcal{P}$ as its marginals (i.e. no distribution Q having $\psi(Q \parallel P_1, P_2, P_3) = 0$). We search for a distribution minimizing the I-aggregate for $w_1 = w_2 = w_3 = \frac{1}{3}$). We provide such a distribution in Table 3.

Table 3. *Joint probability distribution* $Q(X_1, X_2, X_3)$

	$X_2 = 1$		$X_2 = 2$	
Q	$X_3 = 1$	$X_3 = 2$	$X_3 = 1$	$X_3 = 2$
$X_1 = 1$	$\frac{1}{6}$	$\frac{1}{6}$	0	$\frac{1}{6}$
$X_1 = 2$	$\frac{1}{6}$	0	$\frac{1}{6}$	$\frac{1}{6}$

5. Iterative minimization of the I-aggregate

Next we will describe a new procedure, which we call GEMA. The reason for its name is its similarity to the Generalized Expectation Maximization Algorithm [DEM 77].

Our goal is to find a probability distribution Q that minimize the I-aggregate

$$\psi(Q \parallel P_1, \ldots, P_k) \quad = \quad \sum_{j=1,\ldots,k} w_j \cdot I(P_j \parallel Q^{E_j}) \ .$$

The main idea of the proposed iterative algorithm is in order to minimize $\psi(Q \parallel P_1, \ldots, P_k)$ decrease the value of ψ in the next iteration[3], i.e.,

$$\psi(Q_{(i\cdot k)} \parallel P_1, \ldots, P_k) - \psi(Q_{((i+1)\cdot k)} \parallel P_1, \ldots, P_k) \quad \geq \quad 0 \ .$$

LEMMA 19. — *Assume an input set* $\{P_1, \ldots, P_k\}$. *For any two probability distributions* Q *and* R *such that for* $j = 1, \ldots, k$: $P_j \ll Q^{E_j} \ll R^{E_j}$ *it holds that*

$$\psi(R \parallel P_1, \ldots, P_k) - \psi(Q \parallel P_1, \ldots, P_k) \quad \geq \quad I(\widetilde{R} \parallel R) - I(\widetilde{R} \parallel Q) \ ,$$

where $\widetilde{R}$ *denotes the result of the application of operator* $\widetilde{\ }$ *on* R *defined as*

$$\widetilde{R} \quad = \quad \sum_{j=1}^{k} w_j \cdot \pi(R, \mathcal{S}_j) \ .$$

We can apply Lemma 19 with R being the probability distribution $Q_{(i\cdot k)}$ from the last iteration $i \cdot k$ and Q being the probability distribution $Q_{((i+1)\cdot k)}$ from the current iteration $(i+1) \cdot k$. If we take a $Q_{((i+1)\cdot k)}$ such that

$$I(\widetilde{Q}_{(i\cdot k)} \parallel Q_{(i\cdot k)}) - I(\widetilde{Q}_{(i\cdot k)} \parallel Q_{((i+1)\cdot k)}) \quad \geq \quad 0 \ . \qquad [3]$$

then we are sure that we do not increase the value of ψ in the next iteration $(i+1) \cdot k$.

3. We will later see that one iteration of the proposed algorithm consists of k steps. To have the notation ready we index the iterations with multiples of k.

For any finite integer n we will require $Q_{(n\cdot k)}$ to factorize[4] with respect to a class $\mathcal{E} = \{E_1, \ldots, E_k\}$. We will employ IPFP to meet this requirement. Next we will show that it suffices to perform one cycle of IPFP to satisfy condition [3].

LEMMA 20. — *Let IPFP be initiated with distribution $Q_{(i\cdot k)}$ and the input set consist of marginals of $\widetilde{Q}_{(i\cdot k)}$, i.e. $\{\widetilde{Q}^{E_1}_{(i\cdot k)}, \ldots, \widetilde{Q}^{E_k}_{(i\cdot k)}\}$. After one cycle of IPFP distribution $Q_{((i+1)\cdot k)}$ satisfies inequality [3] with equality iff for $j = 1, \ldots, k$: $Q^{E_j}_{(i\cdot k)} = \widetilde{Q}^{E_j}_{(i\cdot k)}$.*

Thus, each cycle of of the proposed algorithm, which we call GEMA, has two different phases.

During the *first phase* I-projections of the distribution from the previous cycle to sets $\mathcal{S}_j = \{Q : Q^{E_j} = P_j\}$ are computed. Subsequently, the weighted average of the I-projections is computed.

For the second phase we need to compute marginal distributions from the weighted average. Observe, that these distributions are necessarily consistent since they are marginals of a joint probability distribution.

The *second phase* consists of k consecutive steps of IPFP starting with the distribution from the previous iteration of GEMA and consequently fitting the marginal distributions computed in the first phase. A formal definition follows.

DEFINITION 21. — *Let $Q_{(0)}$ be an initial joint probability distribution satisfying the same conditions as the initial joint probability distribution in Definition 11. GEMA is the following algorithm:*

for $i = 0, 1, 2, \ldots$
  $\widetilde{Q}_{(i\cdot k)} = \sum_{j=1}^{k} w_j \cdot \pi(Q_{(i\cdot k)}, \mathcal{S}_j)$; // $\mathcal{S}_j = \{Q : Q^{E_j} = P_j\}$
  for $j = 1, \ldots, k$
    compute the marginal $\widetilde{Q}^{E_j}_{(i\cdot k)}$ *of* $\widetilde{Q}_{(i\cdot k)}$;
  for $j = 1, \ldots, k$
    $Q_{(i\cdot k+j)} = \pi(Q_{(i\cdot k+j-1)}, \mathcal{S}'_j)$; // $\mathcal{S}'_j = \{Q : Q^{E_j} = \widetilde{Q}^{E_j}_{(i\cdot k)}\}$

It is easy to compute the I-projections to sets $\mathcal{S}_j$, $\mathcal{S}'_j$ for $j = 1, \ldots, k$ using formula 2. Therefore, similarly as for IPFP, also GEMA can be easily implemented. We will use an example to show the computations that are performed during one iteration of GEMA.

4. In this moment we will not provide any justification for this requirement. Later we will see that this requirement is necessary to guarantee the required properties of the new algorithm.

EXAMPLE 22. — Assume input set $\mathcal{P} = \{P_1(X_1, X_2), P_2(X_2, X_3), P_3(X_1, X_3)\}$ and $w_1 = w_2 = w_3 = \frac{1}{3}$. In this case one iteration of GEMA consists of the following steps:

$$\begin{aligned}
\widetilde{Q}_{(i\cdot k)} &= \frac{1}{3} Q_{(i\cdot k)} \cdot \left(\frac{P_1}{Q^{E_1}_{(i\cdot k)}} + \frac{P_2}{Q^{E_2}_{(i\cdot k)}} + \frac{P_3}{Q^{E_3}_{(i\cdot k)}} \right) \\
Q_{((i\cdot k)+1)} &= Q_{(i\cdot k)} \cdot \frac{\widetilde{Q}^{E_1}_{(i\cdot k)}}{Q^{E_1}_{(i\cdot k)}} \\
Q_{((i\cdot k)+2)} &= Q_{((i\cdot k)+1)} \cdot \frac{\widetilde{Q}^{E_2}_{(i\cdot k)}}{Q^{E_2}_{((i\cdot k)+1)}} \\
Q_{((i\cdot k)+3)} &= Q_{((i\cdot k)+2)} \cdot \frac{\widetilde{Q}^{E_3}_{(i\cdot k)}}{Q^{E_3}_{((i\cdot k)+2)}}
\end{aligned}$$

In order to be able to apply previous results we need to show that GEMA does not introduce "unexpected" zero values into probability distributions.

LEMMA 23. — *Let $Q_{(0)}$ be an initial joint probability distribution satisfying the same conditions as the initial joint probability distribution in Definition 11. Further let $Q_{(i\cdot k)}, i = 0, 1, 2, \ldots$ be the distributions computed by GEMA with the input set $\{P_1, \ldots, P_k\}$. Then for $i = 0, 1, 2, \ldots$ and $j = 1, \ldots, k$ it holds:*

$$P_j \ll Q_{((i+1)\cdot k)} \ll Q_{(i\cdot k)} \ .$$

Now we are ready to prove the main result of this paper about the convergence of GEMA.

THEOREM 24. — *Let $Q_{(0)}$ be an initial joint probability distribution satisfying the same conditions as the initial joint probability distribution in Definition 11 and $\{Q_{(n)}\}_{n=0}^{\infty}$ be a sequence of probability distributions computed by GEMA. Then the sequence $\{\psi(Q_{(n)} \parallel P_1, \ldots, P_k)\}_{n=0}^{\infty}$ converges.*

We used Lemma 15 to characterize the limit probability distribution of a sequence of probability distributions computed by IPFP. We know that $Q^\star$ maximizes entropy from all distributions from $\mathcal{S}_\mathcal{E}$, i.e. from the set of all distributions having the marginals $\{P_1, \ldots, P_k\}$.

Similarly, we can characterize probability distributions computed by GEMA. If $Q_{(0)}$ factorizes with respect to $\mathcal{E}$ then probability distributions $Q_{(i\cdot k+n)}$ computed within a finite number of iterations of GEMA factorize with respect to set $\mathcal{E}$. Now, consider a $Q^\star \in \overline{\mathcal{R}}_\mathcal{E}$ that is the limit probability distribution of a sequence of probability distributions computed by GEMA. Then from Lemma 15 we know that $Q^\star$ maxi-

mizes entropy[5] from all distributions having their marginals equal to the marginals of $Q^\star$ on the sets that are elements of $\mathcal{E}$.

REMARK 25. — A simplified version of GEMA could consist of the first phase of GEMA only[6]. We could use the same line of reasoning to prove its convergence. However, typically the distributions computed by this procedure would not factorize with respect to $\mathcal{E}$ and, consequently, the limit distribution would be different.

We tested the behavior of IPFP and GEMA on *consistent* and *inconsistent* versions of input set $\mathcal{P} = \{P_1, P_2, P_3\}$ defined in Table 1. The plots in Figures 1 and 2 describe development of one probability value (vertical axis) of joint probability distribution $Q_{(i)}$ for one combination of values of random variables. Horizontal axis corresponds to iteration i . We can see that in the *consistent* case both methods converge to the same value. In the *inconsistent* case IPFP oscillates while GEMA converges. We have observed that, in some cases, the convergence of GEMA near the optimum is quite slow. For other experiments see [VOM 99].

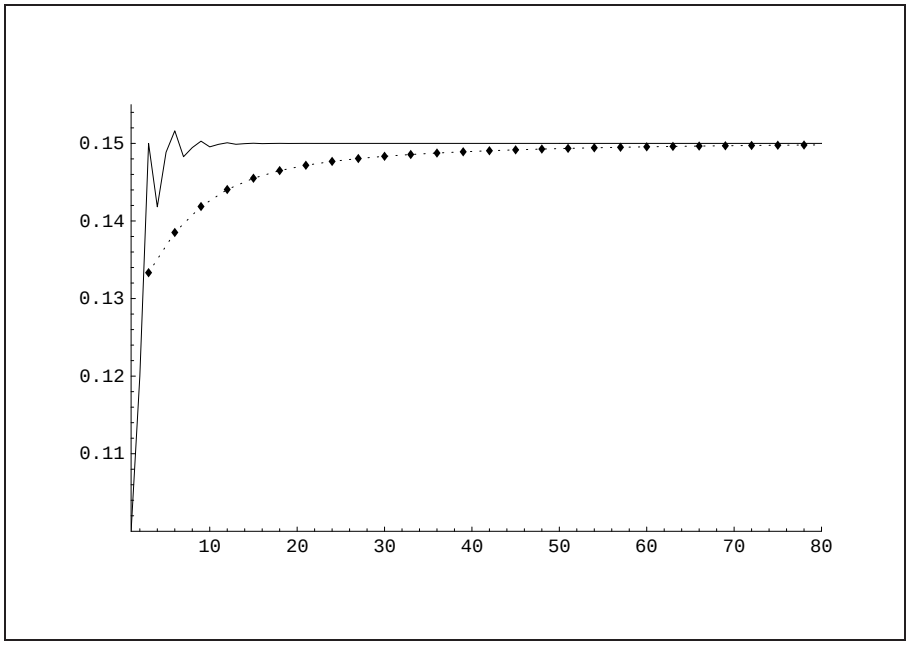


Figure 1. *IPFP (full line) and GEMA (dotted line) applied to the consistent input set* ($\varepsilon = \frac{4}{20}$)

The experiments suggests that the resulting distribution does not depend on the ordering of the input set $\{P_1, \ldots, P_k\}$. Also, the resulting distribution does not seem

5. This provides a justification for the requirement that for $n = 0, 1, 2, \ldots$ probability distributions $Q_{(n \cdot k)}$ should factorize with respect to $\mathcal{E}$.
6. Such a procedure was originally proposed by F. Matúš (personal communication).

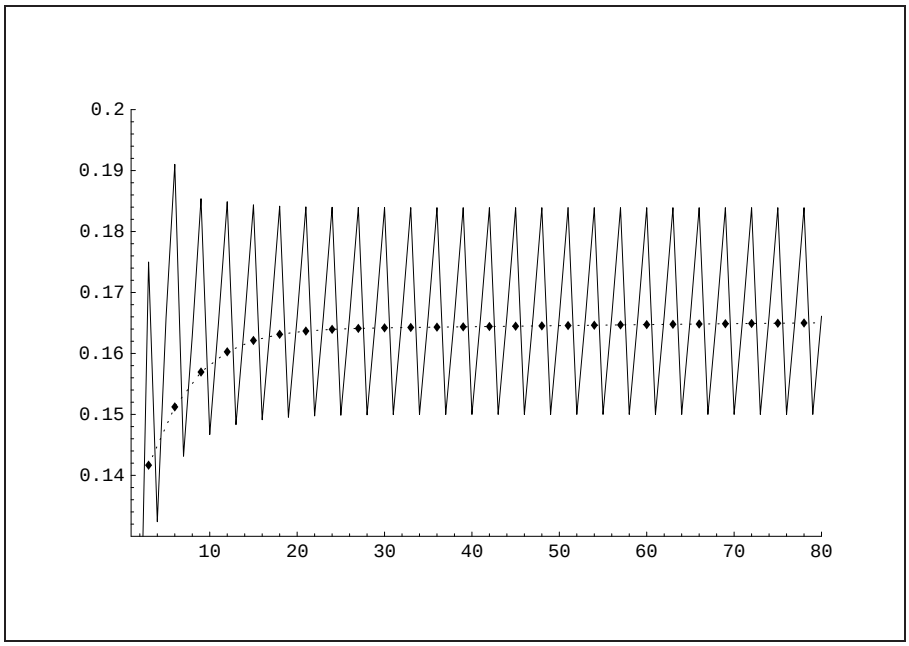


Figure 2. *IPFP (full line) and GEMA (dotted line) applied to the inconsistent input set ($\varepsilon = \frac{3}{20}$)*

to be influenced by the starting distributions provided it satisfies the conditions for the initial joint probability distribution given in Definition 11 and factorizes with respect to $\mathcal{E}$.

6. Efficient implementation of GEMA

In the framework of probabilistic models *factorization* is used to design computationally efficient representations of probability distributions.

DEFINITION 26. — *Set $\mathcal{E}$ of subsets of V is decomposable if either it has exactly one element or it is the union of two disjoint decomposable sets $\mathcal{E}_1$ and $\mathcal{E}_2$ such that there exist $E_1 \in \mathcal{E}_1$ and $E_2 \in \mathcal{E}_2$ so that:*

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

Observe that in GEMA we do not need to have $\widetilde{Q}_{(i \cdot k)}$ represented explicitly since we only need its marginals $\widetilde{Q}_{(i \cdot k)}^{E_n}, n = 1, \ldots, k$. We can construct a decomposable set $\mathcal{C} = \{C_1, \ldots, C_m\}$, $C_i \subseteq V, i = 1, \ldots, m$ such that it holds that for each $P_j(X^{E_j}) \in \mathcal{P}$ there exists $C_i \in \mathcal{C}$ such that $E_j \subseteq C_i$. Then, we can use the junction

tree propagation [JEN 90] with $C_i \in \mathcal{C}$ being the nodes of the junction tree to compute efficiently the marginals $\widetilde{Q}^{E_n}_{(i\cdot k)}, n = 1, \ldots, k$. Thus the first phase of GEMA can be implemented similarly as suggested by Lauritzen [LAU 95] for the EM-algorithm. The computational complexity of the update by one distribution P_j from the input set is proportional to $\sum_{C_i \in \mathcal{C}} |\mathbb{X}^{C_i}|$.

Since the second phase of GEMA is equivalent to one cycle of IPFP we can use the space-saving implementation of IPFP proposed by Jiroušek, see [JIR 91] or [JIR 95a]. The computational complexity of one step of the space-saving implementation of IPFP is also proportional to $\sum_{C_i \in \mathcal{C}} |\mathbb{X}^{C_i}|$.

Consequently, the computational complexity of one cycle of GEMA is proportional to $k \cdot \sum_{C_i \in \mathcal{C}} |\mathbb{X}^{C_i}|$, where k is the number of distributions in the input set $\mathcal{P}$.

7. Conclusions

The proposed algorithm - GEMA - satisfies the original intention to design a method that is an extension of IPFP for inconsistent input set. We have proved that it converges also when the input set of probability distributions is inconsistent. We have proposed to use the I-aggregate to measure how well the marginals of a probability distribution fits the input probability distributions. Although GEMA tends to minimize this criteria we did not provide any guarantee about its convergence to a global or local minima. The conditions that would guarantee such a convergence should be a topic of a future research.

An important property of GEMA is that it can be implemented efficiently. It means that it can be used for computations with probability distributions containing hundreds or thousands of variables provided the probability distributions have a compact factorized representation.

Acknowledgments

This paper benefited from personal communication with Radim Jiroušek, Gernot Kleiter, and Fero Matúš. The author was supported by the Grant Agency of the Czech Republic through grant nr. 201/02/1269.

8. References

[ČEN 82] ČENCOV N. N., *Statistical Decision Rules and Optimal Inference*, Translations of Mathematical Monographs, American Mathematical Society, Providence, 1982.

[CSI 75] CSISZÁR I., "*I*-Divergence Geometry of Probability Distributions and Minimization Problems", *Ann. Prob.*, vol. 3, num. 1, 1975, p. 146–158.

[DEM 40] DEMING W. E., STEPHAN F. F., "On a least square adjustment of a sampled frequency table when the expected marginal totals are known", *Ann. Math. Statist.*, vol. 11, 1940, p. 427–444.

[DEM 77] DEMPSTER A. P., LAIRD N. M., RUBIN D. B., "Maximum Likelihood from Incomplete Data via the EM Algorithm", *J. Roy. Statist. Soc. Ser. B*, vol. 39, 1977, p. 1–38.

[FRE 51] FRECHET M., "Sur les tableaux de corrélation dont les marges sont données", *Ann. Univ. Lyon Sect. A, Series 3*, vol. 14, 1951, p. 53-77.

[HÁJ 92] HÁJEK P., HAVRÁNEK T., JIROUŠEK R., *Uncertain Information Processing in Expert Systems*, CRC Press, Boca Raton, Ann Arbor, London, Tokyo, 1992.

[HOE 40] HOEFFDING W., "Masstabinvariante Korrelationstheorie", *Schriften des mathematischen Instituts und des Instituts für angewandte Mathematik der Universität Berlin*, vol. 5, 1940, p. 179-233.

[JAY 57] JAYNES E. T., "Information theory and statistical mechanics I", *Physics Reviews*, vol. 106, 1957, p. 620–630.

[JEN 90] JENSEN F. V., LAURITZEN S. L., OLESEN K. G., "Bayesian updating in causal probabilistic networks by local computations", *Computational Statistics Quarterly*, vol. 4, 1990, p. 269-282.

[JIR 91] JIROUŠEK R., "Solution of the Marginal Problem and Decomposable Distributions", *Kybernetika*, vol. 27, num. 5, 1991, p. 403–412.

[JIR 95a] JIROUŠEK R., PŘEUČIL S., "On the effective implementation of the iterative proportional fitting procedure", *Computational Statistics & Data Analysis*, vol. 19, 1995, p. 177-189.

[JIR 95b] JIROUŠEK R., VOMLEL J., "Inconsistent knowledge integration in a probabilistic model", *Proc. of Workshop "Mathematical Models for handling partial knowledge in A.I."*, New York, 1995, Plenum Publ. Corp.

[JOH 83] JOHNSON R. W., SHORE J. E., "Comments on and corrections to 'Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy", *IEEE Transactions on Information Theory*, vol. 29, num. 6, 1983, p. 942–943.

[KEL 64] KELLERER H. G., "Verteilungsfunktionen mit gegeben Marginalverteilungen", *Z. Wahrsch. Verw. Gebiete*, vol. 3, 1964, p. 247–270, In German.

[KUL 51] KULLBACK S., LEIBLER R. A., "On Information and Sufficiency", *Annals of Mathematical Statistics*, vol. 22, 1951, p. 79-86.

[KUL 66] KULLBACK S., "An information-theoretic derivation of certain limit relations for a stationary Markov chain", *J. SIAM Control*, vol. 4, num. 3, 1966, p. 454–459.

[LAU 95] LAURITZEN S. L., "The EM algorithm for graphical association models with missing data", *Computational Statistics & Data Analysis*, vol. 19, 1995, p. 191-201.

[LAU 96] LAURITZEN S. L., *Graphical Models*, Clarendon Press, Oxford, 1996.

[OSH 97] OSHERSON D., SHAFIR E., KRANTZ D. H., SMITH E. E., "Probability Bootstrapping: Improving prediction by Fitting Extensional Models to Knowledgeable but Incoherent Probability Judgements", *Organizational Behavior and Human Decision Processes*, vol. 69, num. 1, 1997, p. 1–8.

[PAR 90] PARIS J. B., VENCOVSKÁ A., "A Note on the Inevitability of Maximum Entropy", *International Journal of Approximate Reasoning*, vol. 4, num. 3, 1990, p. 183–223.

[SHA 48] SHANNON C. E., "A Mathematical Theory of Communication", *The Bell System Technical Journal*, vol. 27, 1948, p. 379–423,623–656.

[SHO 80] SHORE J. E., JOHNSON R. W., "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy", *IEEE Transactions on Information Theory*, vol. 26, num. 1, 1980, p. 26-37.

[VOM 99] VOMLEL J., "Methods of Probabilistic Knowledge Integration", PhD thesis, Czech Technical University, Faculty of Electrical Engineeging, 1999, `http://www.utia.cas.cz/vomlel/prace.ps`.

[VOR 62] VOROBEV N. N., "Consistent families of measures and their extensions", *Theory of Probability and Applications*, vol. 7, 1962, p. 147–163.

A. Proofs

LEMMA 14. — *If $\mathcal{S}_\mathcal{E} \neq \emptyset$ then $\mathcal{S}_\mathcal{E} \cap \overline{\mathcal{R}}_\mathcal{E}$ is unique, i.e., contains exactly one probability distribution.*

PROOF. — The idea of the proof is the same as of the proof of Lemma 3.14 in [LAU 96]. Let Q be a probability distribution from $\mathcal{S}_\mathcal{E} \cap \overline{\mathcal{R}}_\mathcal{E}$. Assume an arbitrary $R \in \mathcal{S}_\mathcal{E} \cap \overline{\mathcal{R}}_\mathcal{E}$. We can write[7]:

$$R = \lim_{n\to\infty} \prod_{i=1}^{k} \psi_{E_i}^{(n)} \quad \text{and} \quad Q = \lim_{n\to\infty} \prod_{i=1}^{k} \phi_{E_i}^{(n)}$$

By exploiting the above factorizations we get:

$$\begin{aligned}
\sum_x R(x)\log Q(x) &= \sum_x R(x) \sum_{i=1}^{k} \log \lim_{n\to\infty} \phi_{E_i}^{(n)}(x^{E_i}) \\
&= \lim_{n\to\infty} \sum_{i=1}^{k} \sum_{x^{E_i}} R^{E_i}(x^{E_i}) \log \phi_{E_i}^{(n)}(x^{E_i}) \\
&= \lim_{n\to\infty} \sum_{i=1}^{k} \sum_{x^{E_i}} Q^{E_i}(x^{E_i}) \log \phi_{E_i}^{(n)}(x^{E_i}) \\
&= \sum_x Q(x) \sum_{i=1}^{k} \log \lim_{n\to\infty} \phi_{E_i}^{(n)}(x^{E_i}) \\
&= \sum_x Q(x)\log Q(x) \qquad [4]
\end{aligned}$$

7. The superscript (n) serves as an index of the sequences of potentials $\{\psi_{E_i}^{(n)}\}_{n=1}^{\infty}$ and $\{\phi_{E_i}^{(n)}\}_{n=1}^{\infty}$.

Similarly, we can get:

$$\sum_x Q(x) \log R(x) \quad = \quad \sum_x R(x) \log R(x) \qquad [5]$$

We will use the following well-known inequality that holds for arbitrary probability distributions Q_1, Q_2

$$\sum_x Q_1(x) \log Q_2(x) \quad \leq \quad \sum_x Q_1(x) \log Q_1(x) \ , \qquad [6]$$

and equality holds iff $Q_1 = Q_2$. When [4] and [5] are combined with [6] with R and Q substituted for Q_1 and Q_2 we get

$$\begin{aligned} \sum_x R(x) \log R(x) \ &= \ \sum_x Q(x) \log R(x) \quad \leq \quad \sum_x Q(x) \log Q(x) \\ \sum_x Q(x) \log Q(x) \ &= \ \sum_x R(x) \log Q(x) \quad \leq \quad \sum_x R(x) \log R(x) \ , \end{aligned}$$

which holds only if $R = Q$. Thus we have shown that $\mathcal{S}_\mathcal{E} \cap \overline{\mathcal{R}}_\mathcal{E} \ = \ \{Q\}$. ■

LEMMA 15. — *Let* $\mathcal{E} = \{E_1, \ldots, E_k\}$, $Q \in \overline{\mathcal{R}}_\mathcal{E}$, *and* $\mathcal{S}_\mathcal{E} = \{P : P^{E_1} = Q^{E_1}, \ldots, P^{E_k} = Q^{E_k}\}$. *Then*

(a) $\forall R \in \mathcal{R}_\mathcal{E}, Q \ll R \implies Q = \pi(R, \mathcal{S}_\mathcal{E})$ *and*

(b) Q *maximimizes entropy from all distributions in* $\mathcal{S}_\mathcal{E}$.

PROOF. — First, observe that $Q \in \mathcal{S}_\mathcal{E} \cap \overline{\mathcal{R}}_\mathcal{E}$. From Lemma 14 we know that there is no other distribution in $\mathcal{S}_\mathcal{E} \cap \overline{\mathcal{R}}_\mathcal{E}$ than Q. It suffices to show that $\forall R \in \mathcal{R}_\mathcal{E}, Q \ll R$:

$$\pi(R, \mathcal{S}_\mathcal{E}) \quad \in \quad \mathcal{S}_\mathcal{E} \cap \overline{\mathcal{R}}_\mathcal{E} \ = \ \{Q\} \ .$$

This can be shown using properties of IPFP proven by Csiszár in [CSI 75]. We can start IPFP from arbitrary $R \in \mathcal{R}_\mathcal{E}, Q \ll R$ and iterate fitting the distributions from the input set $\{Q^{E_1}, \ldots, Q^{E_k}\}$. We have that $\mathcal{S}_\mathcal{E} \neq \emptyset$. In such a case IPFP converges to a $Q' \in \mathcal{S}_\mathcal{E} \cap \overline{\mathcal{R}}_\mathcal{E}$. Since this intersection contains only one element we have $Q' = Q$. Csiszár also proved that IPFP converges to $\pi(R, \mathcal{S}_\mathcal{E})$. It follows that $Q = \pi(R, \mathcal{S}_\mathcal{E})$.

Now, assume R to be the uniform probability distribution, i.e., $R(x^V) = c$, where c is a constant such that $\sum_{x^V} R(x^V) = 1$. Observe that for any $\mathcal{E}$ it holds that $R \in \overline{\mathcal{R}}_\mathcal{E}$. Assertion (a) implies that

$$\begin{aligned} Q \quad &= \quad \arg \min_{S \in \mathcal{S}_\mathcal{E}} I(S \parallel R) \ = \ \arg \min_{S \in \mathcal{S}_\mathcal{E}} \sum_{x^V} S(x^V) \log \frac{S(x^V)}{c} \\ &= \quad \arg \max_{S \in \mathcal{S}_\mathcal{E}} - \sum_{x^V} S(x^V) \log S(x^V) \ = \ \arg \max_{S \in \mathcal{S}_\mathcal{E}} H(S) \ . \end{aligned}$$

■

LEMMA 19. — *Assume an input set $\{P_1, \ldots, P_k\}$. For any two probability distributions Q and R such that for $j = 1, \ldots, k$: $P_j \ll Q^{E_j} \ll R^{E_j}$ it holds that*

$$\psi(R \parallel P_1, \ldots, P_k) - \psi(Q \parallel P_1, \ldots, P_k) \quad \geq \quad I(\widetilde{R} \parallel R) - I(\widetilde{R} \parallel Q) \ ,$$

where $\widetilde{R}$ denotes the result of the application of operator $\widetilde{\ }$ on R defined as

$$\widetilde{R} \quad = \quad \sum_{j=1}^{k} w_j \cdot \pi(R, \mathcal{S}_j) \ .$$

PROOF. — Since for $j = 1, \ldots, k$: $P_j \ll Q^{E_j} \ll R^{E_j}$ it holds that

$$R^{E_j}(x^{E_j}) = 0 \implies Q^{E_j}(x^{E_j}) = 0 \implies P_j(x^{E_j}) = 0 \ .$$

Also

$$R^{E_j}(x^{E_j}) = 0 \implies \forall x'^V, x'^{E_j} = x^{E_j} : R(x'^V) = 0$$

and similarly for Q. This together with the convention $0 \log \frac{0}{0} = 0$ allows to perform the sum in the following expressions only over $x^{E_j} : R^{E_j}(x^{E_j}) > 0$. In order to avoid an extensive notation we will omit this condition in the following.

$$\begin{aligned}
&\psi(R \parallel P_1, \ldots, P_k) - \psi(Q \parallel P_1, \ldots, P_k) \\
&= \quad + \sum_{j=1}^{k} w_j \cdot \sum_{x^{E_j}} P_j(x^{E_j}) \cdot \log \frac{P_j(x^{E_j})}{R^{E_j}(x^{E_j})} \\
&\qquad - \sum_{j=1}^{k} w_j \cdot \sum_{x^{E_j}} P_j(x^{E_j}) \cdot \log \frac{P_j(x^{E_j})}{Q^{E_j}(x^{E_j})} \\
&= \quad \sum_{j=1}^{k} w_j \cdot \sum_{x^{E_j}} P_j(x^{E_j}) \cdot \log \frac{Q^{E_j}(x^{E_j})}{R^{E_j}(x^{E_j})} \\
&= \quad \sum_{j=1}^{k} w_j \cdot \sum_{x^{E_j}} P_j(x^{E_j}) \cdot \log \sum_{x^{V \setminus E_j}} \left(\frac{Q(x^{E_j}, x^{V \setminus E_j})}{R^{E_j}(x^{E_j})} \cdot \frac{R(x^{V \setminus E_j} \mid x^{E_j})}{R(x^{V \setminus E_j} \mid x^{E_j})} \right)
\end{aligned}$$

Now we apply Jensen's inequality

$$\log \sum_i \lambda_i x_i \quad \geq \quad \sum_i \lambda_i \log x_i \ \text{ if } \textstyle\sum_i \lambda_i = 1$$

to obtain

$$\begin{aligned}
&\psi(R \parallel P_1, \ldots, P_k) - \psi(Q \parallel P_1, \ldots, P_k) \\
&\geq \sum_{j=1}^{k} w_j \cdot \sum_{x^{E_j}} P_j(x^{E_j}) \cdot \sum_{x^{V \setminus E_j}} R(x^{V \setminus E_j} \mid x^{E_j}) \cdot \log \frac{Q(x^{V \setminus E_j}, x^{E_j})}{R(x^{E_j}, x^{V \setminus E_j})} \\
&\geq \sum_{x^V} \left(\sum_{j=1}^{k} w_j \cdot P_j(x^{E_j}) \cdot \frac{R(x^V)}{R^{E_j}(x^{E_j})} \right) \cdot \log \frac{Q(x^V)}{R(x^V)} \\
&\geq \sum_{x^V} \widetilde{R}(x^V) \cdot \log \frac{Q(x^V)}{R(x^V)} \ , \\
&\geq \sum_{x^V} \widetilde{R}(x^V) \cdot \log \frac{\widetilde{R}(x^V)}{R(x^V)} + \sum_{x^V} \widetilde{R}(x^V) \cdot \log \frac{Q(x^V)}{\widetilde{R}(x^V)} \\
&\geq I(\widetilde{R} \parallel R) - I(\widetilde{R} \parallel Q) \ .
\end{aligned}$$

■

LEMMA 20. — *Let IPFP be initiated with distribution $Q_{(i \cdot k)}$ and the input set consist of marginals of $\widetilde{Q}_{(i \cdot k)}$, i.e. $\{\widetilde{Q}^{E_1}_{(i \cdot k)}, \ldots, \widetilde{Q}^{E_k}_{(i \cdot k)}\}$. After one cycle of IPFP distribution $Q_{((i+1) \cdot k)}$ satisfies inequality [3] with equality iff for $j = 1, \ldots, k$: $Q^{E_j}_{(i \cdot k)} = \widetilde{Q}^{E_j}_{(i \cdot k)}$.*

PROOF. — In each step of IPFP we compute I-projection to a set $\mathcal{S}$, where $\mathcal{S}$ is a set of all distributions having a given marginal. We can use an analog of Pythagoras' theorem in I-divergence geometry [CSI 75]. For $P \in \mathcal{S}$ and any Q that has defined $\pi(Q, \mathcal{S})$ it holds that

$$I(P \parallel Q) \quad = \quad I(P \parallel \pi(Q, \mathcal{S})) + I(\pi(Q, \mathcal{S}) \parallel Q)$$

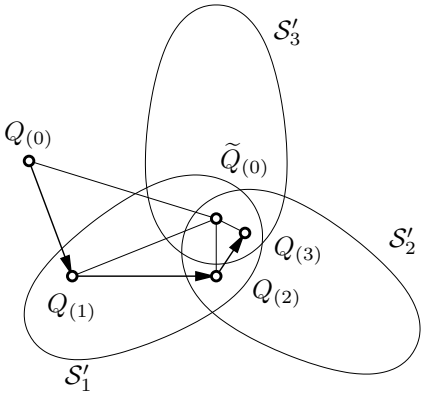
See Figure 3 for an example of the theorem applied to the first cycle of IPFP when there are three distributions in the input set.

In the general case, the analog of "Pythagoras' theorem" implies that

$$\begin{aligned}
I(\widetilde{Q}_{(i \cdot k)} \parallel Q_{(i \cdot k)}) - I(\widetilde{Q}_{(i \cdot k)} \parallel Q_{(i \cdot k+1)}) &= I(Q_{(i \cdot k+1)} \parallel Q_{(i \cdot k)}) \\
I(\widetilde{Q}_{(i \cdot k)} \parallel Q_{(i \cdot k+1)}) - I(\widetilde{Q}_{(i \cdot k)} \parallel Q_{(i \cdot k+2)}) &= I(Q_{(i \cdot k+2)} \parallel Q_{(i \cdot k+1)}) \\
&\ldots \\
I(\widetilde{Q}_{(i \cdot k)} \parallel Q_{(i \cdot k+k-1)}) - I(\widetilde{Q}_{(i \cdot k)} \parallel Q_{(i \cdot k+k)}) &= I(Q_{(i \cdot k+k)} \parallel Q_{(i \cdot k+k-1)})
\end{aligned}$$

The sum all equations gives

$$\begin{aligned}
&I(\widetilde{Q}_{(i \cdot k)} \parallel Q_{(i \cdot k)}) - I(\widetilde{Q}_{(i \cdot k)} \parallel Q_{((i+1) \cdot k)}) \\
&\quad = I(Q_{(i \cdot k+1)} \parallel Q_{(i \cdot k)}) + \ldots + I(Q_{((i+1) \cdot k)} \parallel Q_{((i+1) \cdot k-1)}) \ . \qquad [7]
\end{aligned}$$



$$I(\widetilde{Q}_{(0)} \parallel Q_{(0)}) = I(\widetilde{Q}_{(0)} \parallel Q_{(1)}) + I(Q_{(1)} \parallel Q_{(0)})$$
$$I(\widetilde{Q}_{(0)} \parallel Q_{(1)}) = I(\widetilde{Q}_{(0)} \parallel Q_{(2)}) + I(Q_{(2)} \parallel Q_{(1)})$$
$$I(\widetilde{Q}_{(0)} \parallel Q_{(2)}) = I(\widetilde{Q}_{(0)} \parallel Q_{(3)}) + I(Q_{(3)} \parallel Q_{(2)})$$

Figure 3. *The "Pythagoras' theorem" applied to the first cycle of IPFP with three distributions in the input set*

Since I-divergence is nonnegative it implies that

$$I(\widetilde{Q}_{(i\cdot k)} \parallel Q_{(i\cdot k)}) - I(\widetilde{Q}_{(i\cdot k)} \parallel Q_{((i+1)\cdot k)}) \quad \geq \quad 0 \ ,$$

where the left hand side equals zero if and only if $Q_{(i\cdot k)}$ has all its marginals $Q_{(i\cdot k)}^{E_j}$ for $j = 1, \ldots, k$ equal to the corresponding marginals of $\widetilde{Q}_{(i\cdot k)}$. Therefore one cycle of IPFP satisfies condition [3]. ■

LEMMA 23. — *Let $Q_{(0)}$ be an initial joint probability distribution satisfying the same conditions as the initial joint probability distribution in Definition 11. Further let $Q_{(i\cdot k)}, i = 0, 1, 2, \ldots$ be the distributions computed by GEMA with the input set $\{P_1, \ldots, P_k\}$. Then for $i = 0, 1, 2, \ldots$ and $j = 1, \ldots, k$ it holds:*

$$P_j \ll Q_{((i+1)\cdot k)} \ll Q_{(i\cdot k)} \ .$$

PROOF. — Since $P_j \ll Q_{(0)}$ it suffices to show for $j = 1, \ldots, k$ that

$$P_j \ll Q_{(i\cdot k)} \quad \Longrightarrow \quad P_j \ll Q_{((i+1)\cdot k)} \ \& \ Q_{((i+1)\cdot k)} \ll Q_{(i\cdot k)}$$

We will show it in the same order as GEMA computes these distributions:

$$\pi(Q_{(i\cdot k)}, \mathcal{S}_j)(x) = 0 \quad \Longleftrightarrow \quad Q_{(i\cdot k)}(x) = 0 \ \vee \ P_j(x^{E_j}) = 0$$
$$\widetilde{Q}_{(i\cdot k)}(x) = 0 \quad \Longleftrightarrow \quad j = 1, \ldots, k : \pi(Q_{(i\cdot k)}, \mathcal{S}_j)(x) = 0$$

and for $j = 1, \ldots, k$ it holds that

$$Q_{((i\cdot k)+j)}(x) = 0 \quad \Longleftrightarrow \quad Q_{((i\cdot k)+j-1)}(x) = 0 \ \vee \ P_j(x^{E_j}) = 0 \ .$$

■

THEOREM 15. — *Let $Q_{(0)}$ be an initial joint probability distribution satisfying the same conditions as the initial joint probability distribution in Definition 11 and $\{Q_{(n)}\}_{n=0}^{\infty}$ be a sequence of probability distributions computed by GEMA. Then the sequence $\{\psi(Q_{(n)} \parallel P_1, \ldots, P_k)\}_{n=0}^{\infty}$ converges.*

PROOF. — From Lemmas 19, 20, and 23 we have that $\psi(Q_{(i\cdot k)} \parallel P_1, \ldots, P_k)$ for $Q_{(i\cdot k)}$ computed by GEMA is a non-increasing function of i. It is bounded from below by $\psi(Q_{(i\cdot k)} \parallel P_1, \ldots, P_k) \geq 0$ (see Remark 17). Therefore the sequence of values of $\psi(Q_{(i\cdot k)} \parallel P_1, \ldots, P_k)$ of $Q_{(i\cdot k)}$ computed by GEMA converges. ■