Tensor rank-one decomposition of probability tables

Petr Savicky
Institute of Computer Science
Academy of Sciences of the Czech Republic
Pod vodárenskou věží 2
182 07 Prague, Czech Republic
http://www.cs.cas.cz/~savicky/

Jiří Vomlel
Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
Pod vodárenskou věží 4
182 08 Prague, Czech Republic
http://www.utia.cas.cz/vomlel

Abstract

We propose a new additive decomposition of probability tables - tensor rank-one decomposition. The basic idea is to decompose a probability table into a series of tables, such that the table that is the sum of the series is equal to the original table. Each table in the series has the same domain as the original table but can be expressed as a product of one-dimensional tables. Entries in tables are allowed to be any real number, i.e. they can be also negative numbers. The possibility of having negative numbers, in contrary to a multiplicative decomposition, opens new possibilities for a compact representation of probability tables. We show that tensor rank-one decomposition can be used to reduce the space and time requirements in probabilistic inference. We provide a closed form solution for minimal tensor rank-one decomposition for some special tables and propose a numerical algorithm that can be used in cases when the closed form solution is not known.

1 Introduction

A fundamental property of probabilistic graphical models that allows their application in domains with hundreds to thousands variables is the multiplicative factorization of the joint probability distribution. The multiplicative factorization is exploited in inference methods, e.g., in the junction tree propagation (Jensen et al., 1990). However, in some real applications the models may become intractable using the junction tree propagation and other exact inference methods because after the moralization and triangularization steps the graphical structure becomes too dense, the cliques consist of too many variables, and, consequently, the probability tables corresponding to the cliques are too large to be efficiently manipulated. In such case one usually turns to an approximative inference method.

Following the ideas presented in (Díez and Galán, 2003; Vomlel, 2002) we propose a new decomposition of probability tables that allows to use exact inference in some models where – without the suggested decomposition – the exact inference using the standard methods would be impossible. The basic idea is to decompose a probability table into a series of tables, such that the table that is the sum of the series is equal to the original table. Each table in the series has the same domain as the original table but can be expressed as a product of one-dimensional tables. Entries in tables are allowed to be any real number, i.e. they can be also negative numbers. The possibility of having negative numbers, in contrary to a multiplicative decomposition, opens new possibilities for compact representation of probability tables. To have the decomposition as compact as possible our goal is to find a shortest series.

It is convenient to formally specify the task using the tensor terminology[1]. Assume variables X_i, $i \in N \subset \mathbb{N}$ each variable X_i taking values (a value of X_i will be denoted x_i) from a finite set $\mathcal{X}_i$. Let for any $A \subseteq N$ the symbol x_A denotes a vector of the values $(x_i)_{i \in A}$, where for all $i \in A$: x_i is a value from $\mathcal{X}_i$.

Definition 1 Tensor
Let $A \subset N$. *Tensor* ψ over A is a mapping

$$\times_{i \in A} \mathcal{X}_i \quad \mapsto \quad \mathbb{R} \ .$$

The cardinality $|A|$ of the set A is called *tensor dimension.*

[1] An alternative would be to specify the task using operations with real-valued potentials (Jensen, 2001, Section 1.3.5), but we would need to introduce certain terms for potentials that are standard in the tensor terminology.

Note that every probability table can be looked upon as a tensor. Tensor ψ over A is an (unconditional) probability table if for every x^A it holds that $0 \leq \psi(x^A) \leq 1$ and $\sum_{x^A} \psi(x^A) = 1$. Tensor ψ is a conditional probability table (CPT) if for every x^A it holds that $0 \leq \psi(x^A) \leq 1$ and if there exists $B \subset A$ such that for every x_B it holds $\sum_{x_{A \setminus B}} \psi(x_B, x_{A \setminus B}) = 1$.

Next, we will recall the basic tensor notion. If $|A| = 1$ then tensor is a vector. If $|A| = 2$ then tensor is a matrix. The *outer product* $\psi \otimes \varphi$ of two tensors $\psi : \times_{i \in A} \mathcal{X}_i \mapsto \mathbb{R}$ and $\varphi : \times_{i \in B} \mathcal{X}_i \mapsto \mathbb{R}$, $A \cap B = \emptyset$ is a tensor $\xi : \times_{i \in A \cup B} \mathcal{X}_i \mapsto \mathbb{R}$ defined for all $x_{A \cup B}$ as

$$\xi(x_{A \cup B}) \quad = \quad \psi(x_A) \cdot \varphi(x_B) \ .$$

Now, let ψ and φ are defined on the same domain $\times_{i \in A} \mathcal{X}_i$. The sum $\psi + \varphi$ of two tensors is tensor $\xi : \times_{i \in A} \mathcal{X}_i \mapsto \mathbb{R}$ such that

$$\xi(x_A) \quad = \quad \psi(x_A) + \varphi(x_A) \ .$$

Definition 2 Tensor rank (Håstad, 1990)
Tensor of dimension $|A|$ has *rank* one if it is an outer product of $|A|$ vectors. *Rank* of tensor ψ is the minimal number of tensors of rank one that sum to ψ. *Rank* of tensor ψ will be denoted as $rank(\psi)$.

Note that standard matrix rank is a special case of tensor rank (for $|A| = 2$).

Now, we are ready to formalize the task of decomposition of a probability table into a shortest series of tables that are product of one-dimensional tables.

Definition 3 Tensor rank-one decomposition
Assume a tensor ψ over A. A series of tensors $\{\varrho_b\}_{b=1}^r$ such that

- for $b = 1, \ldots, r$: $rank(\varrho_b) = 1$, i.e.,

  $$\varrho_b \quad = \quad \otimes_{i \in A} \ \varphi_{b,i} \ ,$$

  where $\varphi_{b,i}, i \in A$ are vectors and
- $\psi = \sum_{b=1}^r \varrho_b$

is called *tensor rank-one decomposition* of ψ.

Note that from the definition of tensor rank it follows that for $r \geq rank(\psi)$ such a series always exists. The decomposition is minimal if there is no shorter series satisfying two conditions of Definition 3.

Example 1 Let $\psi : \{0,1\} \times \{0,1\} \times \{0,1\} \mapsto \mathbb{R}$ be

$$\begin{pmatrix} (1,2)^T & (2,4)^T \\ (2,4)^T & (4,9)^T \end{pmatrix} \ .$$

This tensor has rank two since

$$\psi \quad = \quad (1,2) \otimes (1,2) \otimes (1,2) + (0,1) \otimes (0,1) \otimes (0,1)$$

and there are no three vectors whose outer product is equal to ψ. $\diamond$

The rest of the paper is organized as follows. In Section 2 we show using a simple example how tensor rank-one decomposition can be used to reduce the space and time requirements for the probability inference using the junction tree method. We compare sizes of the junction tree for the standard approach, the parent divorcing method, and the junction tree after tensor rank-one decomposition[2]. In Section 3 the main theoretical results are presented: the lower bound on the tensor rank for a class of tensors and minimal tensor rank-one decompositions for some special tensors – max, add, xor, and their noisy counterparts. In Section 4 we propose a numerical method that can be used to find a tensor rank-one decomposition. We also present results of experiments with the numerical method.

2 Tensor rank-one decomposition in probabilistic inference

We will use an example of a simple Bayesian network to show computational savings of the proposed decomposition. Assume a Bayesian network having structure given in Figure 1. Variables $X_1, \ldots, X_m$ are binary taking values 0 and 1. For simplicity, assume that $m = 2^d, d \in \mathbb{N}, 2 \leq d$. Further assume a variable $Y \stackrel{df}{=} X_{m+1}$, which is functionally dependent on $X_1, \ldots, X_m$ and the value y of Y is given by $y = \sum_{i=1}^m x_i$. This means that Y takes $m + 1$ values.

Using the standard junction tree construction (Jensen et al., 1990) we would need to marry all parents of Y. We need not perform triangulation since the graph is triangulated. The resulting junction tree consists of one clique containing all variables, i.e., $C_1 = \{X_1, \ldots, X_m, Y\}$.

Since the CPT $P(Y \mid X_1, \ldots, X_m)$ has a special form we can use the parent divorcing method (Olesen et al., 1989) and introduce a number of auxiliary variables, one auxiliary variable for a pair of parent variables. This is used hierarchically, i.e. we get a tree of auxiliary variables with node Y being the root of the tree. The resulting junction tree consists of $m - 1$ cliques.

[2]Several other methods were proposed to exploit a special structure of CPTs. For a review of these methods see, for example, Vomlel (2002). In this paper, due to the lack of space, we do comparisons with the parent divorcing method only.

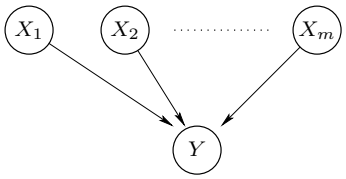

Figure 1: Bayesian network structure.

In Section 3 we will show that if the CPT corresponds to addition of m binary variables we can decompose this CPT to a series of $m+1$ tensors that are products of vectors

$$P(Y \mid X_1, \ldots, X_m) \quad = \quad \sum_{b=1}^{m+1} \otimes_{i=1}^{m+1} \varphi_{b,i} \ .$$

As suggested by Díez and Galán (2003) we can visualize an additive decomposition using one additional variable, which we will denote B. In case of addition of m binary variables variable B will have $m+1$ states. Instead of moralization we add variable B into the model and connect it with nodes corresponding to variables $Y, X_1, \ldots, X_m$. We get the structure given in Figure 2. It is not difficult to show (see Vomlel (2002)) that this model can be used to compute marginal probability distributions as in the original model. The resulting junction tree of this model is given in Figure 3.

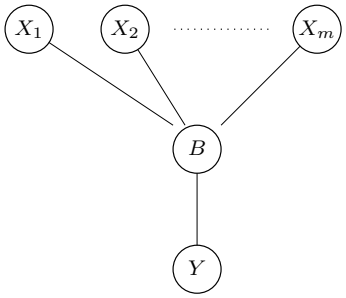

Figure 2: Bayesian network after the decomposition

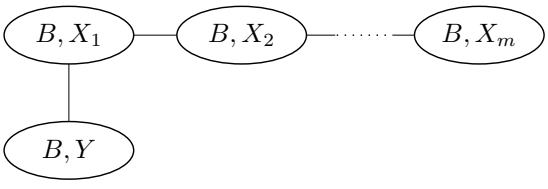

Figure 3: Junction tree for the model after the rank-one decomposition

After little algebra we get that the total clique size in the standard case is $(m+1)\cdot 2^m$, after parent divorcing it is $\frac{1}{3}m^3 + \frac{5}{2}m^2 + 2m\log m - \frac{11}{6}m - 1$, and after the tensor rank-one decomposition it is only $3m^2+4m+1$.

In Figure 4 we compare dependence of the total size of junction trees on the number of parent nodes[3] m of node Y.

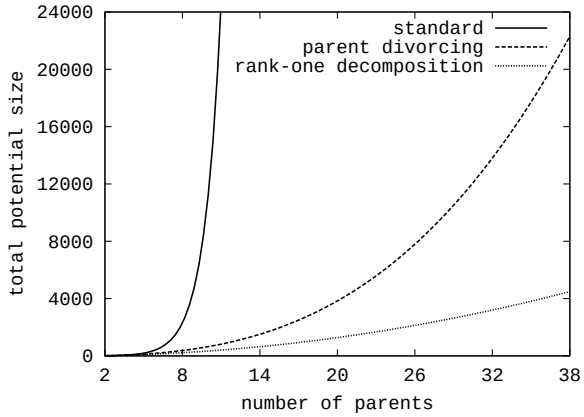

Figure 4: Comparison of the total size of junction tree.

It should be noted that the tensor rank-one decomposition can be applied to any probabilistic model. The savings depends on the graphical structure of the probabilistic model. In fact, by avoiding the moralization of the parents, we give the triangulation algorithm more freedom for the construction of a triangulated graph so that the resulting tables corresponding to the cliques of the triangulated graph can be smaller.

Another possible application of tensor rank-one decomposition could be the compression of probability tables when they become too large to be handled efficiently. In this case we would approximate tensor ψ with another tensor ψ' having sufficiently low rank r'. This could lead to an approximative propagation scheme similar to the penilless propagation (Cano et al., 2000). In the penilless propagation the tables are represented by probability trees, while in our case they would be represented by a series of rank-one tensors, each represented by a set of vectors.

3 Minimal tensor rank-one decomposition

In this section we will present the main theoretical results. It was proven by Håstad (1990) that the computation of tensor rank is an NP-hard problem, therefore determining the minimal rank-one decomposition is also an NP-hard problem. However, we will provide closed-form solution for minimal rank-one decomposition of some special tensors that play an important

[3] It may seem unrealistic to have a node with more than ten parents in an real world application, but it can easily happen, especially, when we need to introduce logical constraints into the model.

role, since they correspond to CPTs that are often used, when creating a Bayesian network model.

The class of tensors of our special interest are tensors ψ_f that represent a functional dependence of one variable $Y \stackrel{df}{=} X_{m+1}$ on variables $X_1, \ldots, X_m$. Let $\mathcal{X} = \mathcal{X}_1 \times \ldots \mathcal{X}_m$ and $x = (x_1, \ldots, x_m) \in \mathcal{X}$. Further, let

$$I(expr) \quad = \quad \begin{cases} 1 & \text{if } expr \text{ is true} \\ 0 & \text{otherwise.} \end{cases}$$

Then for a function $f : \mathcal{X} \mapsto \mathcal{Y}$ the tensor is defined for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$ as $\psi_f(x, y) = I(y = f(x))$.

Let $r = rank(\psi_f)$ and $\xi_b \stackrel{df}{=} \varphi_{m+1,b}$ for all b. Then a minimal tensor rank-one decomposition of ψ_f is

$$\psi_f \quad = \quad \sum_{b=1}^{r} \xi_b \otimes (\otimes_{i=1}^{m} \varphi_{i,b}) \quad . \tag{1}$$

First, we will provide a lower bound on the rank of tensors from this class. This bound will be latter used to prove minimality of certain rank-one decompositions.

Lemma 1 *Let $\psi_f : \mathcal{X} \times \mathcal{Y} \mapsto \{0, 1\}$ be a tensor representing a functional dependence f. Then $rank(\psi_f) \geq |\mathcal{Y}|$.*

Proof For a minimal tensor rank-one decomposition of ψ_f it holds for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$ that

$$\psi_f(x, y) \quad = \quad \sum_{b=1}^{r} \xi_b(y) \cdot \prod_{i=1}^{m} \varphi_{i,b}(x_i) \quad , \tag{2}$$

where $r = rank(\psi_f)$. Consider the matrices[4]

$$\mathbf{W} \quad = \quad \{\psi_f(x, y)\}_{x \in \mathcal{X}}^{y \in \mathcal{Y}} , \quad \mathbf{U} \quad = \quad \{\xi_b(y)\}_{b \in \{1,\ldots,r\}}^{y \in \mathcal{Y}} ,$$

$$\mathbf{V} \quad = \quad \left\{ \prod_{i=1}^{m} \varphi_{i,b}(x_i) \right\}_{x \in \mathcal{X}}^{b \in \{1,\ldots,r\}} .$$

Equation (2) can be rewritten as

$$\mathbf{W} = \mathbf{U}\mathbf{V} \quad .$$

Each column of W contains exactly one nonzero entry, since f is a function. Moreover, each row of W contains at least one nonzero entry, since each $y \in \mathcal{Y}$ is in the range of f. Hence, no row is a linear combination of other rows. Therefore there are $|\mathcal{Y}|$ independent rows in $\mathbf{W}$ and $rank(\mathbf{W}) = |\mathcal{Y}|$. Clearly, $rank(\mathbf{W}) \leq rank(\mathbf{U}) \leq r$. Altogether, $|\mathcal{Y}| \leq r$. □

[4]The upper index labels the rows and the lower index labels the columns.

3.1 Maximum and minimum

Additive decomposition of max was originally proposed by Díez and Galán (2003). This result is included in this section for completeness and we add a result about its optimality. The proofs are constructive, i.e., they provide a minimal tensor rank-one decomposition for *max* and *min*.

Let us assume that $\mathcal{X}_i = [a_i, b_i]$ is an interval of integers for each $i = 1, \ldots, m$. Clearly, the range $\mathcal{Y} = [y_{min}, y_{max}]$ of max on $\mathcal{X}_1 \times \ldots \times \mathcal{X}_m$ is $[\max_{i=1}^m a_i, \max_{i=1}^m b_i]$ and the range $\mathcal{Y} = [y_{min}, y_{max}]$ of min is $[\min_{i=1}^m a_i, \min_{i=1}^m b_i]$.

Theorem 1 *If $f(x) = \max\{x_1, \ldots, x_m\}$, $x_i \in [a_i, b_i]$ for $i = 1, \ldots, m$, $y_{min} = \max_{i=1}^m a_i$, $y_{max} = \max_{i=1}^m b_i$, and $\mathcal{Y} = [y_{min}, y_{max}]$ then $rank(\psi_f) = |\mathcal{Y}|$.*

Proof Let for $i \in \{1, \ldots, m\}$, $x_i \in \mathcal{X}_i$, and $b \in \mathcal{Y}$

$$\varphi_{i,b}(x_i) \quad = \quad \begin{cases} 1 & x_i \leq b \\ 0 & \text{otherwise} \end{cases}$$

and for $y \in \mathcal{Y}$, $b \in \mathcal{Y}$

$$\xi_b(y) \quad = \quad \begin{cases} +1 & b = y \\ -1 & b = y - 1 \\ 0 & \text{otherwise.} \end{cases}$$

Let $\omega(x, y) = I(\max(x_1, \ldots, x_m) \leq y)$. Clearly,

$$\omega(x, y) = \prod_{i=1}^{m} \varphi_{i,y}(x_i)$$

and

$$\psi_f(x, y) \quad = \quad \omega(x, y) - \omega(x, y - 1),$$

where the last term is considered to be zero, if $y = y_{min}$. Altogether,

$$\psi_f(x, y) = \prod_{i=1}^{m} \varphi_{i,y}(x_i) - \prod_{i=1}^{m} \varphi_{i,y-1}(x_i) \quad .$$

Since the product $\xi_b(y) \cdot \prod_{i=1}^m \varphi_{i,b}(x_i)$ is nonzero only for $b = y$ and $b = y - 1$, we have

$$\begin{aligned} \psi_f(x, y) \quad &= \quad \xi_y(y) \cdot \prod_{i=1}^{m} \varphi_{i,y}(x_i) \\ &\quad\quad + \xi_{y-1}(y) \cdot \prod_{i=1}^{m} \varphi_{i,y-1}(x_i) \\ &= \quad \sum_{b \in \mathcal{Y}} \xi_b(y) \cdot \prod_{i=1}^{m} \varphi_{i,b}(x_i) \, . \end{aligned}$$

Taking $b' = b + y_{min} - 1$ we get the required decomposition

$$\psi_f \quad = \quad \sum_{b'=1}^{|\mathcal{Y}|} \xi_{b'} \otimes (\otimes_{i=1}^{m} \varphi_{i,b'}) \quad .$$

By Lemma 1, this is a minimal tensor rank-one decomposition of ψ_f. □

Theorem 2 *If* $f(x) = \min\{x_1, \ldots, x_m\}$, $x_i \in [a_i, b_i]$ *for* $i = 1, \ldots, m$, $y_{min} = \min_{i=1}^{m} a_i$, $y_{max} = \min_{i=1}^{m} b_i$, *and* $\mathcal{Y} = [y_{min}, y_{max}]$ *then* $rank(\psi_f) = |\mathcal{Y}|$.

Proof Let for $i \in \{1, \ldots, m\}$, $x_i \in \mathcal{X}_i$, and $b \in \mathcal{Y}$

$$\varphi_{i,b}(x_i) \quad = \quad \begin{cases} 1 & x_i \geq b \\ 0 & \text{otherwise} \end{cases}$$

and for $y \in \mathcal{Y}$, $b \in \mathcal{Y}$

$$\xi_b(y) \quad = \quad \begin{cases} +1 & b = y \\ -1 & b = y + 1 \\ 0 & \text{otherwise.} \end{cases}$$

and follow an analogous argument as in the proof of Theorem 1 to obtain tensor rank-one decomposition of ψ_f. Again, by Lemma 1, this is a minimal tensor rank-one decomposition. □

Remark If for $i \in \{1, \ldots, m\}$ $\mathcal{X}_i = \{0, 1\}$, then the functions $\max\{x_1, \ldots, x_m\}$ and $\min\{x_1, \ldots, x_m\}$ correspond to logical disjunction $x_1 \vee \ldots \vee x_m$ and logical conjunction $x_1 \wedge \ldots \wedge x_m$, respectively. In Example 2 we illustrate how this can be generalized to Boolean expressions consisting of negations and disjunctions.

Example 2 In order to achieve minimal tensor rank-one decomposition of

$$\psi(x_1, x_2, y) = I(y = x_1 \vee \neg x_2)$$

with variable B having two states 0 and 1, it is sufficient to use functions:

$$\begin{aligned} \varphi_{1,b}(x_1) &= I(x_1 \leq b) \\ \varphi_{2,b}(x_2) &= I(\neg x_2 \leq b) \\ \xi_b(y) &= \begin{cases} +1 & y = b \\ -1 & y = 1,\ b = 0 \\ 0 & y = 0,\ b = 1 \end{cases} \end{aligned}$$

◇

3.2 Addition

In this section, we assume an integer r_i for each $i = 1, \ldots, m$ and assume that $\mathcal{X}_i$ is the interval of integers $[0, r_i]$. This assumption is made for simplicity and without loose of generality. If $\mathcal{X}_i$ are intervals of integers, which do not start at zero, it is possible to transform the variables by subtracting the left boundaries of the intervals to obtain variables satisfying our assumption. Moreover, let $f : \mathbb{N}^m \to \mathbb{N}$ be a function, such that $f(x) = f_0(\sum_{i=1}^{m} x_i)$ where $f_0 : \mathbb{N} \to \mathbb{N}$. Let $\mathcal{A}$ be the interval of integers $[0, \sum_{i=1}^{m} r_i]$. Clearly, $\mathcal{A}$ is the range of $\sum_{i=1}^{m} x_i$.

Theorem 3 *Let* f_0, f *and* $\mathcal{A}$ *be as above. Then* $rank(\psi_f) \leq |\mathcal{A}|$. *Moreover, if* f_0 *is the identity function, then* $rank(\psi_f) = |\mathcal{A}|$.

Proof Let $\alpha_1, \ldots, \alpha_{|\mathcal{A}|}$ be any pairwise distinct real numbers. Let $\varphi_b(x_i) = \alpha_b^{x_i}$ for $i = 1, \ldots, m$, where x_i is an exponent and $\alpha^0 = 1$ for every α. To prove the first assertion of the theorem it is sufficient to show that

$$I(y = f_0(\sum_{i=1}^{m} x_i)) = \sum_{b=1}^{|\mathcal{A}|} \xi_b(y) \cdot \alpha_b^{\sum_{i=1}^{m} x_i} \tag{3}$$

for all combinations of the values of x and y. Substituting $t = \sum_{i=1}^{m} x_i$, we obtain that formula (1) is satisfied for all combinations of the values of x and y, if and only if for all $t, y \in \mathcal{A}$ we have

$$I(y = f_0(t)) = \sum_{b=1}^{|\mathcal{A}|} \xi_b(y) \cdot \alpha_b^t \,. \tag{4}$$

For a fixed y, we can consider the equations (4) for all $t \in \mathcal{A}$ as a system of $|\mathcal{A}|$ linear equations with variables $\xi_b(y)$, $b \in \mathcal{A}$, whose matrix is

$$\begin{pmatrix} \alpha_1^0 & \alpha_2^0 & \ldots & \alpha_{|\mathcal{A}|}^0 \\ \alpha_1^1 & \alpha_2^1 & \ldots & \alpha_{|\mathcal{A}|}^1 \\ \ldots & & & \\ \alpha_1^{|\mathcal{A}|} & \alpha_2^{|\mathcal{A}|} & \ldots & \alpha_{|\mathcal{A}|}^{|\mathcal{A}|} \end{pmatrix} \quad . \tag{5}$$

This matrix is non-singular, since the corresponding Vandermonde determinant is non-zero. The solutions of (4) for each y separately determine the function $\xi_b(y)$, for which (4) and, hence, (3) is satisfied.

If f_0 is the identity, then the range of f is the whole $\mathcal{A}$. It follows from Lemma 1 that $rank(\psi_f) \geq |\mathcal{A}|$ and therefore the above decomposition is minimal. □

Example 3 Let $\mathcal{X}_i = \{0, 1\}$ for $i = 1, 2$, $f(x_1, x_2) = x_1 + x_2$ and $\mathcal{Y} = \{0, 1, 2\}$. We have

$$\begin{aligned} \psi_f(x_1, x_2, y) &= I(y = x_1 + x_2) \\ &= \begin{pmatrix} (1,0,0)^T & (0,1,0)^T \\ (0,1,0)^T & (0,0,1)^T \end{pmatrix} \end{aligned}$$

As in the proof of Theorem 3, we assume $\varphi_{i,b}(x_i) = \alpha_b^{x_i}$ for $i = 1, 2$ and distinct α_b, $b = 0, 1, 2$. For simplicity

of notation, let us assume $\alpha_0 = \alpha$, $\alpha_1 = \beta$ and $\alpha_2 = \gamma$. Let us substitute these $\varphi_{i,b}(x_i)$ into (1) and rewrite it using tensor product as follows.

$$\begin{aligned}\psi_f(x_1, x_2, y) &= (\alpha^0, \alpha^1) \otimes (\alpha^0, \alpha^1) \otimes (u_0, u_1, u_2) \\ &+ (\beta^0, \beta^1) \otimes (\beta^0, \beta^1) \otimes (v_0, v_1, v_2) \\ &+ (\gamma^0, \gamma^1) \otimes (\gamma^0, \gamma^1) \otimes (w_0, w_1, w_2)\end{aligned}$$

For each $y = 0, 1, 2$ we require

$$\begin{pmatrix} I(y=0) & I(y=1) \\ I(y=1) & I(y=2) \end{pmatrix} = u_y \cdot \begin{pmatrix} \alpha^0 & \alpha^1 \\ \alpha^1 & \alpha^2 \end{pmatrix} + v_y \cdot \begin{pmatrix} \beta^0 & \beta^1 \\ \beta^1 & \beta^2 \end{pmatrix} + w_y \cdot \begin{pmatrix} \gamma^0 & \gamma^1 \\ \gamma^1 & \gamma^2 \end{pmatrix} ,$$

which defines a system of three linear equations with three variables u_y, v_y, w_y

$$\begin{pmatrix} I(y=0) \\ I(y=1) \\ I(y=2) \end{pmatrix} = \begin{pmatrix} \alpha^0 & \beta^0 & \gamma^0 \\ \alpha^1 & \beta^1 & \gamma^1 \\ \alpha^2 & \beta^2 & \gamma^2 \end{pmatrix} \cdot \begin{pmatrix} u_y \\ v_y \\ w_y \end{pmatrix} .$$

If α, β, and γ are pairwise distinct real numbers then the corresponding Vandermonde determinant is non-zero and a solution exists. The solution for $\alpha = 1, \beta = 2, \gamma = 3$ is

$$\begin{aligned}\psi_f(x_1, x_2, y) &= (1,1) \otimes (1,1) \otimes (3, -\frac{5}{2}, \frac{1}{2}) \\ &+ (1,2) \otimes (1,2) \otimes (-3, 4, -1) \\ &+ (1,3) \otimes (1,3) \otimes (1, -\frac{3}{2}, \frac{1}{2}) .\end{aligned}$$

◇

3.3 Generalized addition

In this section, we present a tensor rank-one decomposition of ψ_f, where f is defined as $f(x) = f_0(\sum_{i=1}^m f_i(x_i))$. Let $\mathcal{A}$ be the set of all possible values of $\sum_{i=1}^m f_i(x_i)$. The rank of ψ_f depends on the nature of functions f_i, more exactly, on the range of the values of $\sum_{i=1}^m f_i(x_i)$. The decomposition is useful, if this range is substantially smaller than $|\mathcal{X}_1| \cdot \ldots \cdot |\mathcal{X}_m|$.

Theorem 4 *If $f(x) = f_0(\sum_{i=1}^m f_i(x_i))$, where f_i are integer valued functions, then $rank(\psi_f) \leq |\mathcal{A}|$.*

Proof Without loose of generality, we may assume that $f_i(x_i) \geq 0$ for $i = 1, \ldots, m$ and that zero is in the range of f_i. If not, this may be achieved by using $f_i(x_i) - \min_z f_i(z_i)$ instead of f_i and modifying f_0 so that f does not change.

Let $\alpha_1, \ldots, \alpha_{|\mathcal{A}|}$ be positive pairwise distinct real or complex numbers. Let $\varphi(x_i, b) = \alpha_b^{f_i(x_i)}$ for $i = 1, \ldots, m$, where $\alpha^0 = 1$ for every α. To prove the assertion of the theorem it is sufficient to show that

$$I(y = f_0(\sum_{i=1}^m f_i(x_i))) = \sum_{b=1}^{|\mathcal{A}|} \xi_b(y) \cdot \alpha_b^{\sum_{i=1}^m f_i(x_i)} .$$

for all combinations of the values x and y. Substituting $t = \sum_{i=1}^m f_i(x_i)$, we obtain that formula (1) is satisfied for all combinations of the values of x and y, if and only if for all $t \in \mathcal{A}$ and $y \in \mathcal{Y}$, we have

$$I(y = f_0(t)) = \sum_{b=1}^{|\mathcal{B}|} \xi_b(y) \cdot \alpha_b^t \ . \tag{6}$$

For a fixed y, we can consider the equations (6) for all $t \in \mathcal{A}$ as a system of $|\mathcal{A}|$ linear equations with variables $\xi_b(y)$, $b = 1, \ldots, |\mathcal{A}|$, whose matrix is (5) exactly as in the proof of Theorem 3. □

3.4 Exclusive-or (parity) function

Let $\oplus$ denote the addition modulo two, which is also known as the exclusive-or operation[5]. By the parity or exclusive-or function, we will understand the function $x_1 \oplus \ldots \oplus x_m$.

Theorem 5 *Let $\mathcal{X}_i = \mathcal{Y} = \{0, 1\}$ for $i = 1, \ldots, m$ and $f(x) = x_1 \oplus \ldots \oplus x_m$. Then $rank(\psi_f) = 2$.*

Proof The exclusive-or function may easily be expressed as a product, if the values $\{0, 1\}$ are replaced by $\{1, -1\}$ using substitution $0 \mapsto 1$, $1 \mapsto -1$. An odd number of ones in the 0/1 representation is equivalent to a negative product of the corresponding values in the $1/-1$ representation. Expressing the required transformations in the form of a linear transformation, we obtain

$$x_1 \oplus \ldots \oplus x_m = \frac{1}{2} \cdot (1 - (1 - 2x_1) \cdot \ldots \cdot (1 - 2x_m)).$$

Since $\psi_f(x, y) = I(y \oplus x_1 \oplus \ldots \oplus x_m = 0) = y \oplus x_1 \oplus \ldots \oplus x_m \oplus 1$, we have

$$\psi_f(x, y) = \frac{1}{2} \cdot (1 + (1 - 2y) \cdot (1 - 2x_1) \cdot \ldots \cdot (1 - 2x_m)).$$

Hence, ψ_f may be expressed as a sum of two functions, the first of which is the constant $\frac{1}{2}$ and the second is $(\frac{1}{2} - y) \cdot (1 - 2x_1) \cdot \ldots \cdot (1 - 2x_m)$. It is now easy to express ψ_f in the form of (1), if we use tensors defined

[5] Parity is often used in coders and decoders. We conjecture tensor rank-one decomposition may substantially speed up exact inference in probabilistic graphical models used to model decoders for noisy channels.

as follows. Let for $i \in \{1,\ldots,m\}$, $x_i \in \{0,1\}$, and $b \in \{1,2\}$

$$\varphi_{i,b}(x_i) \quad = \quad \begin{cases} 1 & b=1 \\ 1-2x_i & b=2 \end{cases}$$

and

$$\xi_b(y) \quad = \quad \begin{cases} \frac{1}{2} & b=1 \\ \frac{1}{2}-y & b=2\,. \end{cases}$$

It follows from Lemma 1 that this defines a minimal tensor rank-one decomposition of exclusive-or. □

3.5 Noisy functional dependence

For every $i = 1,\ldots,m$ we define a dummy variable X'_i taking values x'_i from set $\mathcal{X}'_i = \mathcal{X}_i$. The noisy functional dependence of Y on $X = (X_1,\ldots,X_m)$ is defined by

$$\psi(x,y) \quad = \quad \sum_{x'} \psi_f(x',y) \cdot \prod_{i=1}^{m} \varkappa_i(x_i,x'_i) \;, \qquad (7)$$

where ψ_f is tensor that represent a functional dependence $y = f(x')$ and for $i = 1,\ldots,m$ tensors $\varkappa_i$ represent the noise for variable X_i. Note that models like noisy-or, noisy-and, etc., fit the above definition. Actually, the definition covers the whole class of models known as models of independence of causal influence (ICI) (Heckerman, 1993).

Theorem 6 *Let tensor ψ represent the noisy functional dependence f defined by formula 7. Then $rank(\psi) \leq rank(\psi_f)$.*

Proof Let $r = rank(\psi_f)$. Then

$$\psi_f(x',y) \quad = \quad \sum_{b=1}^{r} \xi_b(y) \cdot \prod_{i=1}^{m} \varphi_{b,i}(x'_i) \;.$$

Substituting this to formula 7 we get

$$\begin{aligned} \psi(x,y) \quad &= \quad \sum_{x'} \sum_{b=1}^{r} \xi_b(y) \cdot \prod_{i=1}^{m} (\varphi_{b,i}(x'_i) \cdot \varkappa_i(x_i,x'_i)) \\ &= \quad \sum_{b=1}^{r} \xi_b(y) \prod_{i=1}^{m} \sum_{x'_i} (\varphi_{b,i}(x'_i) \cdot \varkappa_i(x_i,x'_i)) \\ &= \quad \sum_{b=1}^{r} \xi_b(y) \cdot \prod_{i=1}^{m} \varphi'_{b,i}(x_i) \;, \end{aligned}$$

where $\varphi'_{b,i}(x_i) = \sum_{x'_i} (\varphi_{b,i}(x'_i) \cdot \varkappa_i(x_i,x'_i))$. The last equation proves that $rank(\psi) \leq rank(\psi_f)$. □

4 Numerical method for tensor rank-one decomposition

Definition 4 Tensor rank-one approximation Assume a tensor ψ and an integer $s \geq 1$. A *tensor rank-one approximation* of length s is a series $\{\varrho_b\}_{b=1}^{s}$ of rank-one tensors ϱ_b that is a tensor rank-one decomposition of a tensor $\hat{\psi}$ with $rank(\hat{\psi}) = s$.
If $\hat{\psi}$ minimizes $\sum_x (\psi(x) - \hat{\psi}(x))^2$ we say that it is a *best tensor rank-one approximation of length s.*

Note that if $s = rank(\psi)$ then the minimal value of $\sum_x (\psi(x) - \hat{\psi}(x))^2$ is zero and a best tensor rank-one approximation of length s is also a minimal tensor rank-one decomposition of ψ. Therefore, we can search numerically for a minimal tensor rank-one decomposition by solving the task from Definition 4 starting with $s = 1$ and then incrementing s by one until $\sum_x (\psi(x) - \hat{\psi}(x))^2$ is sufficiently close to zero.

We performed tests with several gradient methods. The best performance was achieved with Polak-Ribière conjugate gradient method that used the Newton method in one dimension. We performed experiments for tensors corresponding to the exclusive-or and maximum functions of three binary variables. For these functions we know the tensor rank is two therefore we could verify whether for $s = 2$ the algorithm found a tensor rank-one decomposition of these tensors.

The initial values for the algorithm were random numbers from interval $[-0.5,+0.5]$. In most cases the algorithm converged to vectors that were tensor rank-one decomposition. However, sometimes we needed to restart the algorithm from another starting values since it got stuck in a local minima. Figures 5 and 6 illustrate the convergence using three sample runs. The displayed value is one value of $\hat{\psi}$ as it changes with the progress of the algorithm.

5 Related work

Higher-dimensional tensors are studied in multilinear algebra (De Lathauwer and De Moor, 1996). The problem of tensor rank-one decomposition is also known as *canonical decomposition* (CANDECOMP) or *parallel factors* (PARAFAC). A typical task is to find tensor of rank one that is a best approximation of a tensor ψ. This task is usually solved using an alternating least square algorithm (ALS) that is a higher-order generalization of the power method for matrices (De Lathauwer et al., 2000).

An example of application of tensor rank-one decomposition is image sequence compression (Wang and Ahuja, 2004). In this paper the authors use a greedy method for tensor rank-one decomposition. They start

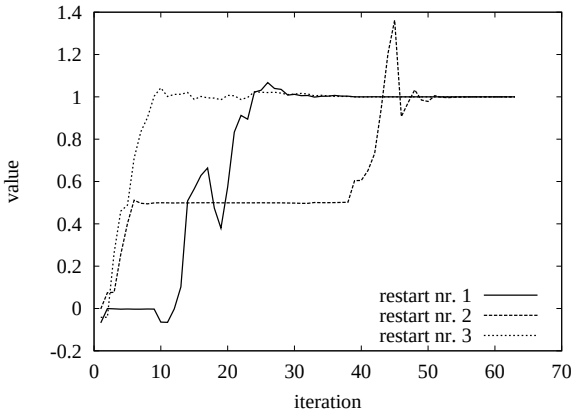

Figure 5: Development of one value of ψ' in case of decomposition of xor.

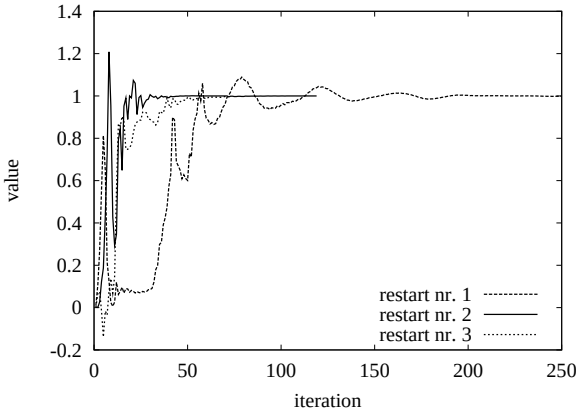

Figure 6: Development of one value of ψ' in case of decomposition of max.

with the original tensor ψ. Than they find a tensor ψ' of rank one that is a best approximation of tensor ψ and compute residuum $\psi - \psi'$. For this residuum again a rank-one tensor that is a best approximation of the residuum is found and the process is repeated until a stopping condition is satisfied. They tested this algorithm using two video sequences and report much higher quality images with the same compression ratio as Principle Component Analysis. We have tested the greedy approach for tensor rank-one decomposition of max and xor function. We observed that, in contrary to the numerical method proposed in Section 4, the greedy approach is not suitable for tensor rank-one decomposition of these tensors.

Acknowledgments

The authors were supported by the Ministry of Education of the Czech Republic under the project nr. 1M0021620808 Institute of Theoretical Computer Science (P. Savicky) and nr. 1M0572 Data, Algorithms, and Decision-Making (J. Vomlel). J. Vomlel was also supported by the Grant Agency of the Czech Republic under the grant project nr. 201/04/0393.

References

Cano A, Moral S, and Salmerón A. Penniless propagation in join trees. *International Journal of Intelligent Systems*, 15:pp. 1027–1059 (2000).

De Lathauwer L and De Moor B. From matrix to tensor: multilinear algebra and signal processing. In *4th Int. Conf. on Mathematics in Signal Processing, Part I*, IMA Conf. Series, pp. 1–11. Warwick (1996). Keynote paper.

De Lathauwer L, De Moor B, and Vandewalle J. On the best rank-1 and rank-$(R_1, R_2, \ldots, R_N)$ approximation of higher-order tensors. *SIAM Journal on Matrix Analysis and Applications*, 21(4):pp. 1324–1342 (2000).

Díez FJ and Galán SF. An efficient factorization for the noisy MAX. *International Journal of Intelligent Systems*, 18:pp. 165–177 (2003).

Heckerman D. Causal independence for knowledge acquisition and inference. In D Heckerman and A Mamdani, eds., *Proc. of the Ninth Conf. on Uncertainty in AI*, pp. 122–127 (1993).

Håstad J. Tensor rank is NP-complete. *Journal of Algorithms*, 11:pp. 644–654 (1990).

Jensen FV. *Bayesian networks and decision graphs*. Statistics for Engineering and Infromation Science. Springer Verlag, New York, Berlin, Heidelberg (2001).

Jensen FV, Lauritzen SL, and Olesen KG. Bayesian updating in recursive graphical models by local computation. *Computational Statistics Quarterly*, 4:pp. 269–282 (1990).

Olesen KG, Kjærulff U, Jensen F, Jensen FV, Falck B, Andreassen S, and Andersen SK. A MUNIN network for the median nerve — a case study on loops. *Applied Artificial Intelligence*, 3:pp. 384–403 (1989). Special issue: Towards Causal AI Models in Practice.

Vomlel J. Exploiting functional dependence in Bayesian network inference. In *Proceedings of the Eightteenth Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 528–535. Morgan Kaufmann Publishers (2002).

Wang H and Ahuja N. Compact representation of multidimensional data using tensor rank-one decomposition. In *Proceedings od International Conference on Pattern Recognition (ICPR)* (2004).