Convergence Results for Relational Bayesian Networks

Manfred Jaeger
Max-Planck-Institut für Informatik
Im Stadtwald, 66123 Saarbrücken
jaeger@mpi-sb.mpg.de

Abstract

Relational Bayesian networks are an extension of the method of probabilistic model construction by Bayesian networks. They define probability distributions on finite relational structures by conditioning the probability of a ground atom $r(a_1, \ldots, a_n)$ on first-order properties of $a_1, \ldots, a_n$ that have been established by previous random decisions. In this paper we investigate from a finite model theory perspective the convergence properties of the distributions defined in this manner. A subclass of relational Bayesian networks is identified that define distributions with convergence laws for first-order properties.

1 Introduction

Relational Bayesian networks are a framework for defining probability distributions on relational structures, and for effectively computing answers to probability-queries (Jaeger 1997, Jaeger 1998).

The original motivation for relational Bayesian networks derives from an artificial intelligence (AI) background, but theoretical investigations into this framework soon lead to questions about convergence of probabilities that also are interesting from a finite model theory perspective. In this paper we will analyze the convergence properties of distributions defined by relational Bayesian networks using techniques adapted from finite model theory. The results we obtain, in turn, contribute to finite model theory convergence results for a new class of distributions, as well as a new perspective on certain already known convergence laws.

1.1 Background in artificial intelligence

Reasoning with probabilistic information is one of the core problems in artificial intelligence. What is desired are formal representation and inference systems for probabilistic information that are expressive, have a well-defined (and understandable) semantics, and are computationally tractable. Favoring the last two of these factors in the inevitable tradeoff, Bayesian networks (Pearl 1988) have emerged as the method of choice for the specification of probability distributions in applications such as fault diagnosis and monitoring.

Bayesian networks can be used to define a probability distribution on a finite probability space consisting of the possible values for a set of finite-range random variables. Thus, if $X_1, \ldots, X_n$ are random variables, V_i is the finite range of possible values of X_i, this probability space is

$$V = V_1 \times \cdots \times V_n. \qquad (1)$$

In principle (and in many applications: in practice) we can restrict ourselves to the case where the X_i are propositional variables, i.e. $V_i = \{\mathit{true}, \mathit{false}\}$ for all i.

The problem with defining and manipulating a probability distribution on V lies in the fact that the size of V is exponential in n, and n, in useful applications, will become fairly large. Hence, maintaining a probability distribution on V explicitly by listing the probability of each individual member of V is infeasible. Bayesian networks make use of the fact that in distributions arising in practice, typically many conditional independencies hold between the variables X_i. A distribution for which this is the case can be represented in an economical way by factorizing it into a number of small dimensional conditional probability distributions. Such a compact representation formalism is complemented in Bayesian networks by inference algorithms that, likewise, attempt to utilize the factorized structure of the distribution for speedy calculations (this does not always work, though (Cooper 1990)). Inference here essentially means probability retrieval, i.e. computation of probabilities like $P(X_2 = \mathit{true}, X_{11} = \mathit{false})$.

Semantically, a distribution on V characterizes the attributes $X_1, \ldots, X_n$ of an individual object or event randomly sampled from some domain. For instance, in a fault diagnosis application, we may have variables $X_1 = \mathit{uneven\_printout}, X_2 = \mathit{toner\_low}, \ldots$. A probability distribution for these X_i then characterizes the attributes of a randomly observed event from the domain of printer failures. In a monitoring application there could be variables like $X_1 = \mathit{position}, X_2 = \mathit{speed}, \ldots$ (where the inherently continuous range of these variables usually would be discretized into finitely many intervals) describing attributes of one specimen from a domain of randomly moving objects.

The fact that a standard Bayesian network only defines a distribution over the fixed and simple space V makes it unsuitable for the specification of some more complex probabilistic models. Specifically, it is not possible to describe relations between several random objects, or events. In our fault diagnosis example, such a relation could be $r = \mathit{caused\_by}$ standing for the fact that one failure was caused by another. In the monitoring example, it might be important to also evaluate probabilities for relations like $r = \mathit{c\_o\_collision}$, standing for the fact that two of the objects are on a course of collision.

Building on previous work on "knowledge based model construction" (see e.g. (Wellman, Breese & Goldman 1992)), in (Jaeger 1997) *relational Bayesian networks* were introduced as a way to also model such relations between several random objects.

In distributions defined by relational Bayesian networks the probability of some proposition of interest, e.g. $\exists x\, \mathit{c\_o\_collision}(\mathit{object1}, x)$, where *object1* is some (observed) element from the domain D, will often depend on the size of D. The exact domain from which the observed random objects have been sampled, however, may be very large and hard to determine exactly. In these situations a natural approach is to use the limiting values for probabilities as the domain size increases as an approximation to the "correct" probability value we wish to infer (Bacchus, Grove, Halpern & Koller 1997). Thus, we are faced with the question: what is the limiting behavior of the probabilities $P(\phi(a))$ assigned by a relational Bayesian network to property ϕ of observed object $a \in D$ as a function of the size of D? Does a limit exist, and if so, what is it?

1.2 Background in finite model theory

Formally, the question we have arrived at by pursuing a problem in reasoning under uncertainty for AI applications is one familiar from finite model theory. Here the subfield of the study of 0-1 laws is concerned with the question of the convergence of probabilities $P_n(\phi)$ $(n \to \infty)$, where P_n is a probability measure on the set of relational structures (for some fixed, finite vocabulary) on a domain of size n, and ϕ is a sentence in some logic. Of the two parameters that characterize this convergence problem – the sequence $(P_n)_n$ and the logic needed to express ϕ – the latter one has commanded the most interest. Following the seminal works of Fagin (1976) and Glebskiĭ et al. (1969), most efforts have gone into studying the convergence properties of $U_n(\phi)$, where U_n is the uniform distribution on structures of size n, and ϕ is a sentence in one or the other extension of first-order logic (e.g. (Blass, Gurevich & Kozen 1985, Kolaitis & M.Y.Vardi 1990, Kolaitis & M.Y.Vardi 1992)).

Works that focus on variations of $(P_n)_n$ mostly have considered random graphs only (i.e. relational structures with a single binary, symmetric and antireflexive relation). Here a complete picture has been obtained for the convergence of first-order and $\mathscr{L}^{\omega}_{\infty\omega}$- properties with respect to probability distributions determined by edge probabilities $n^{-\alpha}$ $(\alpha \geq 0)$ (Lynch 1997). A convergence result for a method of generating random graphs in which edges between different pairs of vertices are not stochastically independent is presented in (McColm 1996). Finally, the study of convergence of conditional probabilities $U_n(\phi \mid \psi)$ (as in (Oberschelp 1982) and (Bacchus et al. 1997)) fits our general description by taking P_n to be the (non-uniform) measures U_n^{ψ} defined as being uniform on models of ψ, and vanishing on models of $\neg\psi$.

On this background, we can view relational Bayesian networks as a new class of (sequences of) non-uniform probability measures $(P_n)_n$ on structures for arbitrary vocabularies. Study of convergence properties of these sequences, thus, is a natural extension of previous work in finite model theory.

2 Relational Bayesian networks

In this section the basic definitions for relational Bayesian networks, as introduced in (Jaeger 1997), are reviewed.

First, some general notational conventions need to be introduced. We use $r_1, r_2, \ldots$ to designate relation symbols. The arity of r is denoted $|r|$. Tuples of variables $v_1, v_2, \ldots$ or constants $a_1, a_2, \ldots$ are denoted by boldface letters $\boldsymbol{a}, \boldsymbol{v}, \ldots$. The length of a tuple $\boldsymbol{v}$ is denoted $|\boldsymbol{v}|$. We also view a tuple loosely as the set of its components, so that expressions like $v \in \boldsymbol{v}$ or $\boldsymbol{v} \subseteq \boldsymbol{w}$ make sense.

When $S = \{r_1, \ldots, r_k\}$ $(|r_i| \geq 1)$ is a finite relational vocabulary, we denote by $\mathrm{Mod}_n(S)$ the set of all S-structures over domain $n = \{0, \ldots, n-1\}$. Script characters $\mathscr{M}$ are used to denote individual S-structures.

We use delimiters $\}\!\!\}, \{\!\!\{$ to denote multisets, i.e. sets whose members q_i are assigned a multiplicity λ_i. Specific multisets can be represented in the form $\{\!\!\{q_1 : \lambda_1; \ldots ; q_m : \lambda_m\}\!\!\}$, or $\{\!\!\{q_i \mid i \in I\}\!\!\}$. Throughout this paper the only multisets we will encounter are finite multisets of probability values $q_i \in [0, 1]$.

Given a vocabulary S, it is now our aim to define a sequence of probability measures $(P_n)_{n\in\omega}$, such that P_n is a measure on $\text{Mod}_n(S)$. We follow the Bayesian network paradigm by defining measures on

$$\text{Mod}_n(S) = \text{Mod}_n(r_1) \times \ldots \times \text{Mod}_n(r_k)$$

in a piecemeal fashion: first the marginal distribution on $\text{Mod}_n(r_1)$ is defined. Then, for each $j = 2, \ldots , k$, and all possible interpretations $I(r_1), \ldots , I(r_{j-1})$, $I(r_j)$ of $r_1, \ldots , r_{j-1}, r_j$ in n we define

$$P_n(I(r_j) \mid I(r_1), \ldots , I(r_{j-1})). \qquad (2)$$

In the following, we abbreviate $(I(r_1), \ldots , I(r_{j-1})) \in \text{Mod}_n(r_1, \ldots , r_{j-1})$ by $\mathscr{M}_{j-1}$. Conditional probabilities (2) are defined by specifying for each $\boldsymbol{a} \in n^{|r_j|}$ the probability

$$P_n(\boldsymbol{a} \in I(r_j) \mid \mathscr{M}_{j-1}). \qquad (3)$$

Once (3) is defined, by equating (2) with

$$\prod_{\boldsymbol{a}\in I(r_j)} P_n(\boldsymbol{a} \in I(r_j) \mid \mathscr{M}_{j-1}) \cdot \prod_{\boldsymbol{a}\notin I(r_j)} (1 - P_n(\boldsymbol{a} \in I(r_j) \mid \mathscr{M}_{j-1})), \qquad (4)$$

a distribution on $\text{Mod}_n(S)$ is defined.

Relational Bayesian networks, now, essentially provide a formal framework for defining functional expressions, called *probability formulas*, that determine conditional probabilities of the form (3).

Centerpiece of probability formulas are *combination functions.* A combination function is any function that maps finite multisets over [0,1] into [0,1]. Important examples of combination functions are

noisy-or :	$\textit{n-o}\{\!\!\{q_i \mid i \in I\}\!\!\}$	$:= 1 - \prod_{i\in I}(1 - q_i)$
maximum :	$\textit{max}\{\!\!\{q_i \mid i \in I\}\!\!\}$	$:= \textit{max}\{q_i \mid i \in I\}$
mean :	$\textit{mean}\{\!\!\{q_i \mid i \in I\}\!\!\}$	$:= \frac{1}{\lvert I\rvert} \sum_{i\in I} q_i$.

In the following definition the term *equality constraint* is used to denote a boolean combination of equality expressions $v_i = v_j$.

Definition 2.1 The class of *probability formulas* over the relational vocabulary S is inductively defined as follows.

(i) (Constants) Each $q \in [0, 1]$ is a probability formula.

(ii) (Indicator functions) For each $r \in S$, and every $|r|$-tuple $\boldsymbol{v}$ of variables, $r(\boldsymbol{v})$ is a probability formula.

(iii) (Convex combinations) When F_1, F_2, F_3 are probability formulas, then so is $F_1F_2 + (1 - F_1)F_3$.

(iv) (Combination functions) When $F_1, \ldots , F_k$ are probability formulas, *comb* is any combination function, $\boldsymbol{v}$,$\boldsymbol{w}$ are tuples of variables, and $\epsilon(\boldsymbol{v}, \boldsymbol{w})$ is an equality constraint, then

$$\textit{comb}\{\!\!\{F_1, \ldots , F_k \mid \boldsymbol{w}; \epsilon(\boldsymbol{v}, \boldsymbol{w})\}\!\!\}$$

is a probability formula.

For a probability formula $F(\boldsymbol{v})$, a structure $\mathscr{M} \in \text{Mod}_n(S)$, and a tuple $\boldsymbol{a} \in n^{|\boldsymbol{v}|}$ it is straightforward to define a value $F(\boldsymbol{a})[\mathscr{M}] \in [0, 1]$ as the interpretation of $F(\boldsymbol{v})$ in $\mathscr{M}$ under the variable assignment $\boldsymbol{v} \mapsto \boldsymbol{a}$. A full inductive definition can be found in (Jaeger 1998). Here we only illustrate the general definition by an example.

Example 2.2 Let $S = \{r, s\}$, with unary r and binary s; let

$$F(v) \equiv \textit{n-o}\{\!\!\{0.3r(w)s(v, w), 0.6r(v) \mid w; w \neq v\}\!\!\}.$$

Let $\mathscr{M} \in \text{Mod}_{20}(S)$, $a = 0$. In order to determine $F(0)[\mathscr{M}]$, we first generate the list of elements $b < 20$ with $\mathscr{M} \models b \neq 0$, which yields the list $1, \ldots , 19$. For each $b \in \{1, \ldots , 19\}$ we then recursively evaluate $0.3r(0)s(0, w)[b][\mathscr{M}]$ and $0.6r(0)[b][\mathscr{M}]$. Assuming that $\mathscr{M} \models r(0)$, and $\mathscr{M} \models s(0, b) \wedge r(b)$ exactly for $b = 1, \ldots , 10$, we obtain the results 0.3 for $b = 1, \ldots , 10$ and 0 for $b = 11, \ldots , 19$ in the recursive evaluation of the first subformula, and the result 0.6 for each $b = 1, \ldots , 19$ in the second. The value of $F(0)[\mathscr{M}]$ then is computed as $\textit{n-o}\{\!\!\{0.3 : 10; 0 : 9; 0.6 : 19\}\!\!\} = 1 - 0.7^{10}0.4^{19}$.

This example illustrates the only non-obvious choice made in the definition of the semantics of $F(a)[\mathscr{M}]$: subformulas F_i inside probability formulas of the form $\textit{comb}\{\!\!\{F_1, \ldots , F_k \mid \boldsymbol{w}; \epsilon(\boldsymbol{v}, \boldsymbol{w})\}\!\!\}$ contribute for each $\boldsymbol{b}$ with $\epsilon(\boldsymbol{a}, \boldsymbol{b})$ a copy of $F(\boldsymbol{a}, \boldsymbol{b})[\mathscr{M}]$ to the multiset to which *comb* then is applied, no matter whether F_i actually contains all the variables $\boldsymbol{w}$ for which the values $\boldsymbol{b}$ are substituted. Specifically, with the combination function

count :

$$\textit{count}\{\!\!\{q_i \mid i \in I\}\!\!\} := (\textit{max}\{1, |\{i \in I \mid q_i > 0\}|\})^{-1}$$

that basically counts the number of nonzero elements in a multiset (and then inverts the result in order to satisfy the requirement to return values in [0,1]), and the probability formula

$$F \equiv \mathit{count}\{\!|1 \mid w; w = w|\!\}$$

we get $F[\mathscr{M}] = 1/n$ for every $\mathscr{M} \in \mathrm{Mod}_n(S)$.

The following lemma, whose easy proof is contained in (Jaeger 1997), shows that in a probability formula we can not only, via indicator functions, access the truth value in $\mathscr{M}$ of atomic sentences, but, via suitable probability formulas, also the truth value of arbitrary first-order formulas.

Lemma 2.3 Let $\phi(\boldsymbol{v})$ be a first-order formula over S. Then there exists a probability formula $F_\phi(\boldsymbol{v})$ over S that uses *max* as the only combination function, s.t. for every $\mathscr{M} \in \mathrm{Mod}_n(S)$, and every $\boldsymbol{a} \in n^{|\boldsymbol{v}|}$: $F_\phi(\boldsymbol{a})[\mathscr{M}] = 1$ iff $\mathscr{M} \models \phi(\boldsymbol{a})$, and $F_\phi(\boldsymbol{a})[\mathscr{M}] = 0$ else.

We can now summarize:

Definition 2.4 Let $S = \{r_1, \dots, r_k\}$. A relational Bayesian network N is a set of probability formulas $F_{r_1}, \dots, F_{r_k}$, where $F_{r_j}(\boldsymbol{v})$ is a probability formula over $\{r_1, \dots, r_{j-1}\}$ with $|\boldsymbol{v}| = |r_j|$. For each $n \in \omega$ the network N defines a distribution P_n^N on $\mathrm{Mod}_n(S)$ by equating (2) with (4), and defining

$$P_n^N(\boldsymbol{a} \in I(r_j) \mid \mathscr{M}_{j-1}) := F_{r_j}(\boldsymbol{a})[\mathscr{M}_{j-1}]. \qquad (5)$$

Example 2.5 (Random graphs) Let r_1, r_2 be binary relation symbols. Define

$F_{r_1}(v_1, v_2) :\equiv \frac{1}{\sqrt{2}}$ and
$F_{r_2}(v_1, v_2) :\equiv \mathit{max}\{\!|r_1(v_1, v_2), r_1(v_2, v_1) \mid \emptyset; v_1 \neq v_2|\!\}$.

Then the network N consisting of F_{r_1} and F_{r_2} defines distributions P_n^N on $\mathrm{Mod}_n(r_1, r_2)$ whose marginal distribution on $\mathrm{Mod}_n(r_2)$ is just the uniform measure on symmetric, antireflexive relations r_2, i.e. random graphs with edge probability $\frac{1}{2}$. Note that the same distributions on $\mathrm{Mod}_n(r_2)$ could not be obtained using a single probability formula F_{r_2}, because then, according to (4), $r_2(a_1, a_2)$ and $r_2(a_2, a_1)$ $(a_1, a_2 \in n)$ would be independent, and symmetry could not be enforced.

Extending the method indicated by the foregoing example, one can show that it is possible to define for every parametric class (cf. (Oberschelp 1982, Ebbinghaus & Flum 1995)) of S-structures a relational Bayesian network N for some $S' \supseteq S$, such that P_n^N marginalized on S is the uniform distribution on the given parametric class.

Example 2.6 (Sparse random graphs) For $\alpha \geq 0$ define

$$\mathit{count}^\alpha: \quad \mathit{count}^\alpha\{\!|q_i \mid i \in I|\!\} := (\mathit{count}\{\!|q_i \mid i \in I|\!\})^\alpha$$

with *count* as in example 2.2. With r_1, r_2, F_{r_2} as in example 2.5, and $F_{r_1} :\equiv \mathit{count}^{\alpha/2}\{\!|1 \mid w; w = w|\!\}$ we then obtain that P_n^N marginalized on $\mathrm{Mod}_n(r_2)$ describes random graphs with edge probability $n^{-\alpha}$.

Example 2.7 (Discrete approximation of Gilbert random graphs) Gilbert random graphs for the unit circle and threshold $\delta > 0$ are generated by the following procedure (cf.(McColm 1996)): on the unit circle in the plane n points are scattered at random (w.r.t. the uniform distribution). Two points are joined by an edge e iff their distance on the circle is $< \delta$. The distributions Q_n induced by this process on $\mathrm{Mod}_n(e)$ can not be modeled exactly by relational Bayesian networks. However, we can obtain an approximation to Q_n as follows. We imagine the unit circle to be partitioned into m segments of equal length. Let $s_1, \dots, s_m$ be unary relation symbols with the intended meaning: $s_i(v)$ iff point v lies in the i-th segment of the circle. The random distribution of points on the unit circle then can be approximated by assigning each point to exactly one of the relations s_i via the probability formulas

$$F_{s_i}(v) :\equiv (1 - s_1(v)) \cdots (1 - s_{i-1}(v)) \frac{1}{m - i + 1}$$

$(i = 1, \dots, m)$. Also, we need a formula $F_e(v_1, v_2)$ that models the (deterministic) connection of points v_1, v_2, depending on whether they are assigned to sufficiently closely neighbored segments (in a sense that approximates the δ-threshold). This can easily be done in a lengthy, but structurally simple (combination function free), probability formula. The resulting measure P_n^N on $\mathrm{Mod}_n(s_1, \dots, s_m, e)$, marginalized to $\mathrm{Mod}_n(e)$, then approximates the measure Q_n.

3 Convergence

In practical applications of relational Bayesian networks we will mostly be interested in probabilities of a form like $P_n^N(r_2(\boldsymbol{a}) \wedge r_{11}(\boldsymbol{b}))$, where $\boldsymbol{a}$ and $\boldsymbol{b}$ represent tuples of observed random objects or events, which are sampled from a domain that, without loss of generality, we can identify with some $n \in \omega$. Thus, the convergence problem of greatest interest to us is that regarding the convergence of $P_n^N(\psi(\boldsymbol{a}))$, where ψ is a quantifier free first-order sentence with constants

$\boldsymbol{a} \subseteq \omega$. These constants are interpreted deterministically, i.e. they are not subject to random interpretation in the domain: the constant $a \in \omega$ is interpreted by a in structures over a domain $n > a$. Sentences that contain a we do not interpret over domains $n \leq a$.

It follows from lemma 2.3 that the convergence problem for quantifier free $\psi(\boldsymbol{a})$ already is equivalent to the convergence problem for arbitrary first-order $\phi(\boldsymbol{a})$: for a given network N over S, and an S-formula $\phi(\boldsymbol{v})$ we can choose a new relation symbol r_ϕ with $|r_\phi| = |\boldsymbol{v}|$, and define a network N' over $S \cup \{r_\phi\}$ by adding to N the probability formula F_ϕ for r_ϕ as given by lemma 2.3. Then

$$P_n^{N'}(\forall \boldsymbol{v}(\phi(\boldsymbol{v}) \leftrightarrow r_\phi(\boldsymbol{v}))) = 1,$$

and hence for all $\boldsymbol{a}$

$$P_n^N(\phi(\boldsymbol{a})) = P_n^{N'}(\phi(\boldsymbol{a})) = P_n^{N'}(r_\phi(\boldsymbol{a})).$$

3.1 Proportional extension axioms

Probabilities of the form $P_n^N(r(\boldsymbol{a}))$ are essentially given by the expected size of the interpretation of r in a random structure $\mathscr{M}_n$. We therefore now introduce the necessary tools to investigate the behavior of the size of interpretations, and, more generally, of types, in random structures.

We use the following standard terminology: An S-*type* in the variables $\boldsymbol{v} = v_0, \ldots, v_n$ is a maximally consistent conjunction $\tau(\boldsymbol{v})$ of S-literals in the variables $\boldsymbol{v}$. A type $\tau(\boldsymbol{v})$ is called *proper* if it contains all the formulas $v_i \neq v_j$ $(v_i, v_j \in \boldsymbol{v}, i \neq j)$. A type $\sigma(\boldsymbol{v}, w)$ *extends* the type $\tau(\boldsymbol{v})$, written $\tau \subset \sigma$, if every conjunct of τ is a conjunct of σ.

To express quantitative properties of random structures, we use a part of an extension of first-order logic that was developed by Bacchus (1990) and Grove et al. (1992) for reasoning about statistical information.

Definition 3.1 Let $\sigma(\boldsymbol{v}, \boldsymbol{w})$ be a proper type. The expression $|\sigma(\boldsymbol{v}, \boldsymbol{w})|_{\boldsymbol{w}}$ is called a *proportion term.* For $p \in [0,1]$ and $s \in (0,1)$ the expression $|\sigma(\boldsymbol{v}, \boldsymbol{w})|_{\boldsymbol{w}} = (1 \pm s)p$ is called a *proportion equation.* We define satisfaction of proportion equations by $\mathscr{M} \in \mathrm{Mod}_n(S)$ and variable assignment $\boldsymbol{v} \mapsto \boldsymbol{a}$:

$$\mathscr{M}[\boldsymbol{v}/\boldsymbol{a}] \models \ |\sigma(\boldsymbol{v}, \boldsymbol{w})|_{\boldsymbol{w}} = (1 \pm s)p$$

iff

$$\frac{|\{\boldsymbol{b} \mid \mathscr{M}[\boldsymbol{v}/\boldsymbol{a}, \boldsymbol{w}/\boldsymbol{b}] \models \sigma(\boldsymbol{v}, \boldsymbol{w})\}|}{n^{|\boldsymbol{w}|}} \in [(1-s)p, (1+s)p]. \tag{6}$$

Thus, $\mathscr{M}[\boldsymbol{v}/\boldsymbol{a}] \models |\sigma(\boldsymbol{v}, \boldsymbol{w})|_{\boldsymbol{w}} = (1 \pm s)p$ if the proportion of tuples $\boldsymbol{b}$ that satisfy $\sigma(\boldsymbol{a}, \boldsymbol{b})$, among all $|\boldsymbol{w}|$-tuples, is (approximately) p. We also allow the degenerate case $\boldsymbol{w} = \emptyset$. The term on the left hand side of (6) then is interpreted as 1 if $\mathscr{M}[\boldsymbol{v}/\boldsymbol{a}] \models \sigma(\boldsymbol{v})$, and else as 0.

The semantics of proportion equations is extended canonically to first-order formulas containing proportion equations as atomic formulas. In particular those of the following form.

Definition 3.2 Let $\tau(\boldsymbol{v}) \subseteq \sigma(\boldsymbol{v}, \boldsymbol{w})$ be proper types. Let $p \in [0,1], s \in (0,1)$. The formula

$$\begin{aligned} &pea(\tau, \sigma, p, s) := \\ &\qquad \forall \boldsymbol{v}(\tau(\boldsymbol{v}) \rightarrow |\sigma(\boldsymbol{v}, \boldsymbol{w})|_{\boldsymbol{w}} = (1 \pm s)p) \end{aligned} \tag{7}$$

is called a *proportional extension axiom.* The pair $(\tau(\boldsymbol{v}), \sigma(\boldsymbol{v}, \boldsymbol{w}))$ alone is called a *proportional extension axiom schema.*

Our plan, now, is to identify a certain subclass of relational Bayesian networks N, so that the following holds for $(P_n^N)_n$: for every pair $\tau(\boldsymbol{v}) \subseteq \sigma(\boldsymbol{v}, \boldsymbol{w})$ of proper types there exists $p \in [0,1]$, so that for every $s \in (0,1)$

$$P_n^N(pea(\tau, \sigma, p, s)) \rightarrow 1 \quad (n \rightarrow \infty). \tag{8}$$

3.2 Convergence results

We use an inductive approach to prove that (8) holds whenever N only contains probability formulas that satisfy certain conditions. In every induction step we assume that (8) holds for $\{r_1, \ldots, r_{i-1}\}$-types τ, σ. Then admissible probability formulas $F_{r_i}(\boldsymbol{v})$ are limited to those for which $F_{r_i}(\boldsymbol{a}_n)[\mathscr{M}_n]$ converges (sufficiently fast), given that $(\mathscr{M}_n)_n$ is a sequence that satisfies certain proportional extension axioms. From the convergence of $F_{r_i}(\boldsymbol{a}_n)[\mathscr{M}_n]$ we then derive the new extension axioms for $\{r_1, \ldots, r_i\}$-types τ, σ.

We will need the following standard concepts relating to rates of convergence: when $(a_n)_{n\in\omega}$ and $(b_n)_{n\in\omega}$ are sequences of real numbers, we write $a_n = o(b_n)$ iff $lim_{n\rightarrow\infty} \mid a_n/b_n \mid = 0$. Of particular interest to us is the case where $a_n = p + o(c^n)$ for some p and $c \in (0,1)$, i.e. when $(a_n)_n$ converges exponentially fast to p. Throughout we assume c in the expression $o(c^n)$ to be instantiated by some $c \in (0,1)$ that is appropriately chosen in the given context. In other words, the statement $a_n = p + o(c^n)$, without further qualifications, means "there exists $c \in (0,1)$ s.t. $a_n = p + o(c^n)$", or more loosely, "(a_n) converges exponentially fast to p".

The class of relational Bayesian networks for which we will obtain convergence results is characterized by a restriction on the combination functions that may be used in these networks. The following definitions characterize these admissible combination functions.

Definition 3.3 Let $k \in \omega$, and for $j = 1, \ldots, k$ let $p_j, q_j \in [0,1]$; $l_j \in \omega \setminus 0$, $m_j \in \omega$. A sequence $A_n = \{\!|p_{n,i} \mid i \in I_n|\!\}$ $(n \in \omega)$ of finite multisets is called *convergence testing (with parameters $(p_j, q_j, l_j, m_j)_j$)* iff there exist $e \in (0,1)$, and a sequence $s_n \to 0$ such that for each n

(i) For all $i \in I_n$: $p_{n,i} \in \cup_{j=1}^{k} [p_j - e^n, p_j + e^n]$.

(ii) For $j = 1, \ldots, k : |\{i \mid p_{n,i} \in [p_j - e^n, p_j + e^n]\}| \in [(1-s_n)q_j n^{l_j} + m_j, (1+s_n)q_j n^{l_j} + m_j]$.

Roughly speaking, $(A_n)_n$ is convergence testing, if the A_n contain polynomially growing numbers of elements that are clustered in intervals around the p_j of exponentially decreasing length.

Definition 3.4 A combination function *comb* is called *exponentially convergent* iff for every set of parameters $(p_j, q_j, l_j, m_j)_{j=1,\ldots,k}$ there exists $r \in [0,1]$, so that for every sequence $(A_n)_n$ of multisets that is convergence testing with parameters $(p_j, q_j, l_j, m_j)_j$, we have that

$$comb A_n = r + o(c^n). \tag{9}$$

Lemma 3.5 *maximum* and *noisy-or* are exponentially convergent.

Proof: The statement for *maximum* is obvious. To prove the statement for *noisy-or*, let $(A_n)_n$ be a convergent testing sequence of multisets. It is sufficient to consider the case where $k = 1$, i.e. $(A_n)_n$ has parameters (p, q, l, m), because after proving the lemma for this case, the general result follows from the fact that exponential convergence is preserved under products of k factors.

If $q = 0$, then the number of elements in A_n is constant, and the result follows again from the fact that a product of a constant number of exponentially convergent factors is exponentially convergent.

Assume, then, that $q > 0$. If $p > 0$ we get

$$n\text{-}o\, A_n \geq 1 - (1 - p + e^n)^{(1-s_n)qn^l} = 1 - o(c^n),$$

because $(1 - p + e^n)^{(1-s_n)q} < d$ for some $d < 1$ and all sufficiently large n.

When $p = 0$, we have

$$n\text{-}o\, A_n \leq 1 - (1 - e^n)^{(1+s_n)qn^l}.$$

To conclude the proof, it is therefore sufficient to show that for $e \in (0,1)$, $l \in \omega \setminus 0$

$$(1 - e^n)^{n^l} = 1 - o(c^n). \tag{10}$$

To prove (10), we write its left hand side as

$$\sum_{i=0}^{n^l} (-1)^i \binom{n^l}{i} e^{in} = 1 - n^l e^n + \ldots + (-1)^{n^l} e^{nn^l}. \tag{11}$$

Since

$$\frac{\binom{n^l}{i+1} e^{(i+1)n}}{\binom{n^l}{i} e^{in}} = \frac{n^l - i}{i+1} e^n < 1$$

for $i = 1, \ldots, n^l$ when n is sufficiently large, we find that the absolute values of the terms in (11) are monotonically decreasing in i, for all n greater than some n_0. Hence, for all such n

$$1 \geq (1 - e^n)^{n^l} \geq 1 - n^l n^l e^n = 1 - o(c^n).$$

□

The other combination functions we have met – *mean* and *count* – clearly are not exponentially convergent.

The following theorem constitutes the first important part of our convergence results. It links the convergence of $F(\boldsymbol{a}_n)[\mathscr{M}_n]$ to exponential convergence of combination functions and the validity in $\mathscr{M}_n$ of proportional extension axioms. The theorem also contains an assertion about computability. Here and in the following we tacitly assume that statements about computability are qualified by the obvious conditions, like that probability formulas do not contain any nonrecursive reals; the values of combination functions are computable for arguments provided in a suitable finite representation, and limits r in (9) are computable from the parameters (p_j, q_j, l_j, m_j).

Theorem 3.6 Let $F(\boldsymbol{v})$ be a probability formula that only contains exponentially convergent combination functions. There exists a finite set

$$\mathsf{Ax} := \{(\tau_i(\boldsymbol{v}_i), \sigma_i(\boldsymbol{v}_i, \boldsymbol{w}_i)) \mid 1 \leq i \leq M\}$$

of proportional extension axiom schemas such that for all sequences of sets

$$\mathsf{Ax}(\boldsymbol{q}, s_n) := \{pea(\tau_i, \sigma_i, q_i, s_n) \mid 1 \leq i \leq M\}\ (n \in \omega),$$

of proportional extension axioms obtained by instantiating the schemas in Ax with parameters $q_i \in [0,1]$,

$s_n \in (0,1)$ with $s_n \to 0$, and for every S-type $\tau(\boldsymbol{v})$, there exists $p \in [0,1]$, such that the following holds: whenever for each $n \in \omega$ we have $\mathscr{M}_n \in \mathrm{Mod}_n(S)$ and $\boldsymbol{a}_n \in n^{|\boldsymbol{v}|}$ with

$$\mathscr{M}_n \models \mathsf{Ax}(\boldsymbol{q}, s_n) \wedge \tau(\boldsymbol{a}_n),$$

then

$$F(\boldsymbol{a}_n)[\mathscr{M}_n] = p + o(c^n). \tag{12}$$

Given F we can effectively determine Ax. Given Ax and $\boldsymbol{q}$ we can compute p.

Proof: The idea of the proof can be summarized very briefly: We simply enforce with the axioms in $\mathsf{Ax}(\boldsymbol{q}, s_n)$ that the combination functions contained in F are applied to convergence testing sequences of multisets. By the assumption that the combination functions are exponentially convergent, this forces the convergence of $F(\boldsymbol{a}_n)[\mathscr{M}_n]$.

For a formal proof, we proceed by induction on the structure of F. The case $F \equiv q$ is trivial. Suppose, next, that $F(\boldsymbol{v}) \equiv r(\boldsymbol{v})$. Let $\mathscr{M}_n \in \mathrm{Mod}_n(S)$, and $\boldsymbol{a}_n \in n^{|\boldsymbol{v}|}$ with $\mathscr{M}_n \models \tau(\boldsymbol{a}_n)$. If $\tau(\boldsymbol{a}_n) \models r(\boldsymbol{a}_n)$ then $F(\boldsymbol{a}_n)[\mathscr{M}_n] = 1$; if $\tau(\boldsymbol{a}_n) \models \neg r(\boldsymbol{a}_n)$ then $F(\boldsymbol{a}_n)[\mathscr{M}_n] = 0$. Hence, the theorem holds with $\mathsf{Ax} = \emptyset$, and the appropriate $p \in \{0,1\}$.

In the case $F \equiv F_1F_2 + (1-F_1)F_3$ the result follows from the fact that exponential convergence is preserved under sums and products.

We now turn to the case

$$F \equiv \mathit{comb}\{\!\{F_1(\boldsymbol{v},\boldsymbol{w}), \ldots, F_L(\boldsymbol{v},\boldsymbol{w}) \mid \boldsymbol{w}; \epsilon(\boldsymbol{v},\boldsymbol{w})\}\!\}$$

with *comb* an exponentially convergent combination function. For $j = 1, \ldots, L$ let

$$\mathsf{Ax}_j = \{(\tau_{ji}, \sigma_{ji}) \mid 1 \leq i \leq M_j\}$$

be the sets of proportional extension axiom schemas as given by the induction hypothesis that the theorem holds for the F_j. We now proceed as follows: we define a new set Ax of proportional extension axiom schemas, such that for all instantiations $\mathsf{Ax}(\boldsymbol{q}, s_n)$ of Ax, and all $(\mathscr{M}_n)_n$, $(\boldsymbol{a}_n)_n$ as in the theorem, we have that

$$(A_n(\mathscr{M}_n, \boldsymbol{a}_n))_{n\in\omega} \tag{13}$$

with

$$A_n(\mathscr{M}_n, \boldsymbol{a}_n) := \{\!\{F_1(\boldsymbol{a}_n, \boldsymbol{b})[\mathscr{M}_n], \ldots, F_l(\boldsymbol{a}_n, \boldsymbol{b})[\mathscr{M}_n] \mid \boldsymbol{b}, \mathscr{M}_n \models \epsilon(\boldsymbol{a}_n, \boldsymbol{b})\}\!\}$$

is a convergence testing sequence of multisets. Then (12) immediately follows from (9). The new set Ax will consist of the union of the Ax_j and some new schemas $\mathsf{Ax}_{\mathrm{new}}$. The old Ax_j will ensure condition (i) of definition 3.3, whereas $\mathsf{Ax}_{\mathrm{new}}$ will provide for condition (ii).

Let $\tau(\boldsymbol{v})$ be some given S-type. We determine a set $\mathsf{Ax}_{\mathrm{new}}(\tau)$ such that (13) is a convergence testing sequence of multisets when

$$\mathscr{M}_n \models \wedge_j \mathsf{Ax}_j(\boldsymbol{q}_j, s_n) \wedge \mathsf{Ax}_{\mathrm{new}}(\tau)(\boldsymbol{q}, s_n) \wedge \tau(\boldsymbol{a}_n)$$

for some $\boldsymbol{q}_j, \boldsymbol{q}, s_n$. The set $\mathsf{Ax}_{\mathrm{new}}$ then will just be the union of the $\mathsf{Ax}_{\mathrm{new}}(\tau)$ over all S-types τ for $\boldsymbol{v}$.

First, consider the case that $\tau(\boldsymbol{v}) \models \neg\epsilon(\boldsymbol{v},\boldsymbol{w})$. In that case $A_n(\mathscr{M}_n, \boldsymbol{a}_n) = \emptyset$ whenever $\mathscr{M}_n \models \tau(\boldsymbol{a}_n)$. We can define $\mathsf{Ax}_{\mathrm{new}}(\tau) = \emptyset$, and obtain (12) with $p = \mathit{comb}\,\emptyset$.

Assume, then, that $\tau(\boldsymbol{v}) \not\models \neg\epsilon(\boldsymbol{v},\boldsymbol{w})$. Let $\sigma_1(\boldsymbol{v},\boldsymbol{w}), \ldots, \sigma_K(\boldsymbol{v},\boldsymbol{w})$ be the S-types for $(\boldsymbol{v},\boldsymbol{w})$ that extend $\tau(\boldsymbol{v})$ and are consistent with $\epsilon(\boldsymbol{v},\boldsymbol{w})$. By induction hypothesis, for each $1 \leq h \leq K$, each $1 \leq j \leq L$, and each sequence $(\mathscr{M}_n, (\boldsymbol{a}_n, \boldsymbol{b}_n))_n$ with

$$\mathscr{M}_n \models \mathsf{Ax}_j(\boldsymbol{q}_j, s_n) \wedge \sigma_h(\boldsymbol{a}_n, \boldsymbol{b}_n)$$

we have that

$$F_j(\boldsymbol{a}_n, \boldsymbol{b}_n)[\mathscr{M}_n] = p_{jh} + o(c^n) \tag{14}$$

for some $p_{jh} \in [0,1]$. Thus, we see that according to the induction hypothesis the $A_n(\mathscr{M}_n, \boldsymbol{a}_n)$, satisfy condition (i) of definition 3.3. In order to make sure that the $A_n(\mathscr{M}_n, \boldsymbol{a}_n)$ also satisfy condition (ii), we enforce via instantiations of suitable proportional extension axiom schemas that

$$|\{\boldsymbol{b}_n \mid \mathscr{M}_n \models \sigma_h(\boldsymbol{a}_n\boldsymbol{b}_n)\}| \in [(1-s_n)qn^l + m, (1+s_n)qn^l + m] \tag{15}$$

for some parameters l, m, q, s_n. This, of course, is essentially what we can stipulate directly by a proportional extension axiom. There is a small technical problem that we need to deal with, though: proportional extension axioms (axiom schemas) only were defined for proper types. Here we can assume neither for τ nor for σ_h that it is a proper type.

Particularly, there is the degenerate case to be accounted for where for each $w \in \boldsymbol{w}$ there exists $v \in \boldsymbol{v}$ with $\sigma_h \models w = v$. In that case the left hand side of (15) equals 1 for all $\mathscr{M}_n$ and all $\boldsymbol{a}_n$ with $\mathscr{M}_n \models \tau(\boldsymbol{a}_n)$. With $q = 0$, $m = 1$ then (15) is satisfied.

Next, consider the case where $\sigma_h \models \wedge_{v\in\boldsymbol{v}} w \neq v$ for at least one $w \in \boldsymbol{w}$. Without loss of generality, assume that this is the case exactly for $h = 1, \ldots, K'$ ($K' < K$). We then define the "proper parts" $\tilde{\tau}(\tilde{\boldsymbol{v}}), \tilde{\sigma}_h(\tilde{\boldsymbol{v}}, \tilde{\boldsymbol{w}})$

of τ, σ_h as follows: first, $\tilde{\tau}(\tilde{\boldsymbol{v}})$ is chosen as the restriction of $\tau(\boldsymbol{v})$ to a maximal subset $\tilde{\boldsymbol{v}} \subseteq \boldsymbol{v}$ of variables such that $\tau \models v \neq v'$ for all $v, v' \in \tilde{\boldsymbol{v}}$. Subsequently, $\tilde{\sigma}_h(\tilde{\boldsymbol{v}}, \tilde{\boldsymbol{w}})$ is chosen as the restriction of $\sigma(\boldsymbol{v}, \boldsymbol{w})$ to a maximal subset $\tilde{\boldsymbol{v}} \cup \tilde{\boldsymbol{w}} \subseteq \boldsymbol{v} \cup \boldsymbol{w}$ such that $\sigma_h \models u \neq u'$ for all $u, u' \in \tilde{\boldsymbol{v}} \cup \tilde{\boldsymbol{w}}$. Now define

$$\mathsf{Ax}_{\text{new}}(\tau) := \{(\tilde{\tau}(\tilde{\boldsymbol{v}}), \tilde{\sigma}_h(\tilde{\boldsymbol{v}}, \tilde{\boldsymbol{w}})) \mid h = 1, \ldots, K'\}. \quad (16)$$

It now follows directly from the definitions that whenever $\mathscr{M}_n \models \mathit{pea}(\tilde{\tau}, \tilde{\sigma}_h, q, s_n)$ for some q, we have that (15) is satisfied with parameters $q, l = |\tilde{\boldsymbol{w}}|, m = 0$.

To conclude the proof, it only remains to note that when $p_{jh} = p_{j'h'}$ for two (or more) pairs of indices $1 \leq j, j' \leq L$, $1 \leq h, h' \leq K$, the total number of elements in the interval $p_{jh} + o(c^n)$ is bounded by $(1 \pm s_n)qn^l + m + (1 \pm s'_n)q'n^{l'} + m'$ (or longer sums of similar terms), which, in turn, can be bounded by expressions $(1 \pm s^*_n)q^*n^{l^*} + m^*$ for suitable parameters l^*, m^*, q^*, s^*_n. □

To prepare the proof of our main theorem, we first formulate a special version in a family of theorems that are collectively known as Chernoff bounds (see (Hagerup & Rüb 1990) for a useful overview).

Theorem 3.7 Let $p, q > 0$, $s_n > 0$ $(n \in \omega)$ with $s_n \to 0$. For each $n \in \omega$ let

$$k(n) \in [(1 - s_n)qn, (1 + s_n)qn],$$

and let

$$X^n_1, \ldots, X^n_{k(n)}$$

be 0,1-valued, mutually independent random variables with

$$P(X^n_i = 1) \in [p - a_n, p + a_n]$$

for some $a_n \in [0, 1]$ $(i = 1, \ldots, k(n); n \in \omega)$. Let $S_n := \sum_{i=1}^{k(n)} X^n_i$.

If $p > 0$ and $a_n = o(1)$, or if $p = 0$ and $a_n = o(c^n)$, then for all $e \in (0, 1)$

$$P(|S_n - qnp| > eqnp) = o(c^n). \quad (17)$$

Proof: It is clearly sufficient to prove the theorem for the case $k(n) = n$, i.e. $q = 1$, $s_n = 0$. Also, we can assume that the sequence a_n is decreasing in n. Let $e \in (0, 1)$ be fixed.

First consider the case $p > 0$. For $n \in \omega$ define

$$p_n := \frac{1}{n} \sum_{i=1}^{n} P(X^n_i = 1).$$

Then $p_n \in [p - a_n, p + a_n]$. From a version of Chernoff's theorem given in (Hagerup & Rüb 1990), we get that for every $e' > 0$

$$P(|S_n - np_n| > e'np_n) = o(c^n). \quad (18)$$

There exists $n_0 \in \omega$ such that for all $m \geq n_0$

$$|p_m - p| < \frac{1}{2}ep. \quad (19)$$

Since $p > 0$, then $e' := \frac{1}{2}e\frac{p}{p + a_{n_0}} \in (0, 1)$, and we obtain for all $m \geq n_0$:

$$\begin{aligned} |S_m - mp| > emp \Rightarrow & \\ |S_m - mp_m| &> emp - \tfrac{1}{2}emp \\ &= \tfrac{1}{2}emp \\ &= e'm(p + a_{n_0}) \\ &\geq e'mp_m. \end{aligned}$$

Hence (18) implies (17).

In the case $p = 0$ the left hand side of (17) becomes $P(S_n > 0)$. Under the assumption that $a_n = o(c^n)$, we obtain

$$P(S_n > 0) \leq 1 - (1 - a_n)^n = o(c^n)$$

as in the proof of lemma 3.5 □

We are now ready to formulate and prove the main technical result.

Theorem 3.8 Let N be a relational Bayesian network for S that only contains exponentially convergent combination functions. Let $\tau(\boldsymbol{v}) \subseteq \sigma(\boldsymbol{v}, \boldsymbol{w})$ be proper S-types. Then there exists $p \in [0, 1]$ such that for all $s \in (0, 1)$

$$P^N_n(\mathit{pea}(\tau, \sigma, p, s)) \to 1 \quad (n \to \infty). \quad (20)$$

Given N, τ, and σ we can compute p.

Proof: First we observe that it is sufficient to prove the theorem for the special case $|\boldsymbol{w}| = 1$. To see why this is the case, consider types $\tau(\boldsymbol{v}) \subset \sigma_1(\boldsymbol{v}, w_1) \subset \sigma_2(\boldsymbol{v}, w_1, w_2)$. Let s be given, and assume that p_1, p_2 are such that (20) holds for $\mathit{pea}(\tau, \sigma_1, p_1, s)$ and $\mathit{pea}(\sigma_1, \sigma_2, p_2, s)$. With

$$\begin{aligned} P^N_n(\mathit{pea}(\tau, \sigma_1, p_1, s) \wedge \mathit{pea}(\sigma_1, \sigma_2, p_2, s)) \geq & \\ 1 \quad & -(1 - P^N_n(\mathit{pea}(\tau, \sigma_1, p_1, s))) \\ & -(1 - P^N_n(\mathit{pea}(\sigma_1, \sigma_2, p_2, s))) \end{aligned}$$

and

$$\begin{aligned} \models \mathit{pea}(\tau, \sigma_1, p_1, s) \wedge \mathit{pea}(\sigma_1, \sigma_2, p_2, s) \to & \\ \mathit{pea}(\tau, \sigma_2, p_1p_2, 2s) & \end{aligned}$$

it follows that (20) holds for $pea(\tau, \sigma_2, p_1p_2, 2s)$. The argument obviously can be extended for $|\boldsymbol{w}| > 2$.

We now prove the theorem for $\tau(\boldsymbol{v}), \sigma(\boldsymbol{v}, w)$ by induction on the size of S. As the base case, take $S = \emptyset$. Then $\mathrm{Mod}_n(S)$ contains the single structure n; N is the empty network, and $P_n^N(n) = 1$. The proper S-types τ and σ are

$$\tau(\boldsymbol{v}) \equiv \bigwedge v_i \neq v_j, \quad \sigma(\boldsymbol{v}, w) \equiv \bigwedge v_i \neq v_j \wedge \bigwedge v_i \neq w,$$

where the conjunctions range over all pairs of distinct variables. Then,

$$n \models pea(\tau, \sigma, 1, |\boldsymbol{v}|/n) \tag{21}$$

for all n, and therefore $P_n^N(pea(\tau, \sigma, 1, s)) \to 1$ for all $s > 0$.

Now, let N_k be a relational Bayesian network for $S_k = \{r_1, \dots, r_k\}$, and assume that the theorem holds for all networks for $S_{k-1} = \{r_1, \dots, r_{k-1}\}$. In particular, it holds for the network N_{k-1} obtained from N_k by removing the probability formula F_{r_k} for r_k. For the record, we note that, naturally, the marginal distribution of $P_n^{N_k}$ on $\mathrm{Mod}_n(S_{k-1})$ is just $P_n^{N_{k-1}}$.

We partition $\tau(\boldsymbol{v})$ and $\sigma(\boldsymbol{v}, w)$ into two, respectively four, conjuncts:

$$\begin{aligned} \tau(\boldsymbol{v}) &\equiv \tau^{S_{k-1}}(\boldsymbol{v}) \wedge \tau^{r_k}(\boldsymbol{v}) \\ \sigma(\boldsymbol{v}, w) &\equiv \tau^{S_{k-1}}(\boldsymbol{v}) \wedge \tau^{r_k}(\boldsymbol{v}) \\ &\quad \wedge \sigma^{S_{k-1}, w}(\boldsymbol{v}, w) \wedge \sigma^{r_k, w}(\boldsymbol{v}, w) \end{aligned}$$

where $\tau^{S_{k-1}}$ contains all conjuncts of τ that are S_{k-1}-literals (including the (in-)equality formulas), τ^{r_k} contains all r_k-literals of τ, $\sigma^{S_{k-1}, w}$ contains all S_{k-1}-literals of σ that contain the variable w, and $\sigma^{r_k, w}$ contains all r_k-literals of σ containing w. We also use the following abbreviations:

$$\begin{aligned} \sigma^{S_{k-1}}(\boldsymbol{v}, w) &:\equiv \tau^{S_{k-1}}(\boldsymbol{v}) \wedge \sigma^{S_{k-1}, w}(\boldsymbol{v}, w) \\ \sigma^{w}(\boldsymbol{v}, w) &:\equiv \sigma^{S_{k-1}, w}(\boldsymbol{v}, w) \wedge \sigma^{r_k, w}(\boldsymbol{v}, w). \end{aligned}$$

We prove the theorem by showing that

$$P_n^{N_k}(pea(\tau^{S_{k-1}}, \sigma^w, p, s)) \to 1 \tag{22}$$

for some p. This then proves $P_n^{N_k}(pea(\tilde{\tau}, \tilde{\sigma}, p, s)) \to 1$ for every S_k-type $\tilde{\tau}$ for $\boldsymbol{v}$ that is consistent with $\tau^{S_{k-1}}(\boldsymbol{v})$, and the corresponding extension $\tilde{\sigma}$ by σ^w, including, of course, the original τ, σ we started out with.

By our induction hypothesis there exists $q \in [0, 1]$ and a sequence $(s_n)_n \subset [0, 1]$ with $s_n \to 0$, s.t.

$$P_n^{N_{k-1}}(pea(\tau^{S_{k-1}}, \sigma^{S_{k-1}}, q, s_n)) \to 1 \tag{23}$$

If $q = 0$ then (22) holds with $p = 0$, and we are done.

Suppose, then, that $q > 0$. Consider some fixed $\mathscr{M}_n \in \mathrm{Mod}_n(S_{k-1})$ and $\boldsymbol{a} \in n^{|\boldsymbol{v}|}$ with

$$\mathscr{M}_n \models pea(\tau^{S_{k-1}}, \sigma^{S_{k-1}}, q, s_n) \wedge \tau^{S_{k-1}}(\boldsymbol{a}).$$

Then

$$\begin{aligned} &|\{b \mid \mathscr{M}_n \models \sigma^{S_{k-1}, w}(\boldsymbol{a}, b)\}| \\ &\qquad \in [(1 - s_n)qn, (1 + s_n)qn]. \end{aligned} \tag{24}$$

Consider a given $b \in n$ with $\mathscr{M}_n \models \sigma^{S_{k-1}, w}(\boldsymbol{a}, b)$. The literals in $\sigma^{r_k, w}(\boldsymbol{a}, b)$ are of the form $r_k(\boldsymbol{c})$ and $\neg r_k(\boldsymbol{c})$ where $\boldsymbol{c}$ varies over all tuples of length $|r_k|$ with elements from $\boldsymbol{a}$ and b that contain b at least once. Denote by C^+ the set of all such $\boldsymbol{c}$ that appear in a positive literal $r(\boldsymbol{c})$, and by C^- those $\boldsymbol{c}$ that appear in a literal $\neg r(\boldsymbol{c})$. Then, by definition,

$$\begin{aligned} &P_n^{N_k}(\sigma^{r_k, w}(\boldsymbol{a}, b) \mid \mathscr{M}_n) = \\ &\qquad \prod_{\boldsymbol{c} \in C^+} F_{r_k}(\boldsymbol{c})[\mathscr{M}_n] \prod_{\boldsymbol{c} \in C^-} (1 - F_{r_k}(\boldsymbol{c})[\mathscr{M}_n]). \end{aligned} \tag{25}$$

On the left hand side of (25) we view $\mathscr{M}_n$ as a subset of $\mathrm{Mod}_n(S_k)$: the set of all S_k-structurese whose S_{k-1}-reduct is $\mathscr{M}_n$.

Since the $\boldsymbol{c}$ all contain at least one b, for $b \neq b'$ the events $\sigma^{r_k, w}(\boldsymbol{a}, b)$ and $\sigma^{r_k, w}(\boldsymbol{a}, b')$ are conditionally independent, given $\mathscr{M}_n$.

Next, we will show that the probability (25) is close to a constant value q' that is independent of the particular choice of $\mathscr{M}_n, \boldsymbol{a}_n$, and b. Then, using, the independence of the $\sigma^{r_k, w}(\boldsymbol{a}, b)$, we obtain that with high probability the number of elements b, that satisfy $\sigma^w(\boldsymbol{a}, b)$ is very near $qq'n$.

With the assumptions made so far we cannot yet bound the probability (25). In order to do this with the aid of theorem 3.6, we need to exploit our induction hypothesis some more.

Let Ax be the set of M proportional extension axiom schemas as given by theorem 3.6 for F_{r_k}. By our induction hypothesis, there exist parameters $\boldsymbol{q} = (q_1, \dots, q_M)$ with

$$P_n^{N_{k-1}}(\mathsf{Ax}(\boldsymbol{q}, s_n)) \to 1 \tag{26}$$

(for convenience we may here assume the same sequence $(s_n)_n$ as appears in (23)).

Now consider a factor $F_{r_k}(\boldsymbol{c})[\mathscr{M}_n]$ in (25). By theorem 3.6 we have

$$F_{r_k}(\boldsymbol{c})[\mathscr{M}_n] = \tilde{q} + o(c^n) \tag{27}$$

for $\mathscr{M}_n \models \mathsf{Ax}(\boldsymbol{q}, s_n)$, with $\tilde{q}$ only depending on the S_{k-1}-type of $\boldsymbol{c}$. When in (25) we vary n, $\mathscr{M}_n$, $\boldsymbol{a}$, and b, only requiring that $\mathscr{M}_n \models \sigma^{S_{k-1}}(\boldsymbol{a}, b)$, then in each case the right hand side of (25) contains the same number of factors $F_{r_k}(\boldsymbol{c})[\mathscr{M}_n]$ and $(1 - F_{r_k}(\boldsymbol{c})[\mathscr{M}_n])$ for each possible S_{k-1}-type of $\boldsymbol{c}$. Hence, from (27) it follows that for such $\mathscr{M}_n, \boldsymbol{a}, b$ with $\mathscr{M}_n \models \mathsf{Ax}(\boldsymbol{q}, s_n)$

$$P_n^{N_k}(\sigma^{r_k,w}(\boldsymbol{a}, b) \mid \mathscr{M}_n) = q' + o(c^n) \qquad (28)$$

for some $q' \in [0,1]$.

Now let $s > 0$ be given. For this s and $p = qq'$ we want to prove (22). Let

$$\mathscr{M}_n \models \mathit{pea}(\tau^{S_{k-1}}, \sigma^{S_{k-1}}, q, s_n) \wedge \mathsf{Ax}(\boldsymbol{q}, s_n) \wedge \tau^{S_{k-1}}(\boldsymbol{a}).$$

From the independence of the events $\sigma^{r_k,w}(\boldsymbol{a}, b)$ for different b, applying theorem 3.7, we obtain

$$P_n^{N_k}(\mid \sigma^{r_k,w}(\boldsymbol{a}, w) \mid_w = (1 \pm s)q'q \mid \mathscr{M}_n) = 1 - o(c^n)$$

Then, for every

$$\mathscr{M}_n \models \mathit{pea}(\tau^{S_{k-1}}, \sigma^{S_{k-1}}, q, s_n) \wedge \mathsf{Ax}(\boldsymbol{q}, s_n)$$

we get

$$\begin{aligned}
&P_n^{N_k}(\forall \boldsymbol{v}(\tau^{S_{k-1}}(\boldsymbol{v}) \rightarrow \\
&\qquad \mid \sigma^{r_k,w}(\boldsymbol{v}, w) \mid_w = (1 \pm s)qq') \mid \mathscr{M}_n) \\
&= P_n^{N_k}(\bigcap_{\substack{\boldsymbol{a} \in n^{|\boldsymbol{v}|} \\ \mathscr{M}_n \models \tau^{S_{k-1}}(\boldsymbol{a})}} \mid \sigma^{r_k,w}(\boldsymbol{a}, w) \mid_w = (1 \pm s)qq') \mid \mathscr{M}_n) \\
&\geq 1 - n^{|\boldsymbol{v}|} \cdot o(c^n) \\
&= 1 - o(c^n),
\end{aligned}$$

Finally, with (23) and (26) we obtain (22).

At first glance it may appear as though we had proved not only the convergence stated in (8), but even exponential convergence. This is not the case, however, because the claim to exponential convergence gets lost when for (23) and (26) we translate convergence for every $s \in (0,1)$ into convergence for some sequence $s_n \rightarrow 0$.

The proof of (8) gives rise to a recursive procedure for the computation of p: we have $p = qq'$ with q determined by (23), and q' determined by several factors $\tilde{q}$ as provided by (27). The computation of q is accomplished by recursion that is grounded on the case of $\emptyset$-types solved by (21). The computation of $\tilde{q}$ is accomplished by computing the parameters $\boldsymbol{q}$ in (26) by further recursive calls, and by using theorem 3.6 to compute $\tilde{q}$ from Ax and $\boldsymbol{q}$. □

With theorem 3.8 it now is not difficult to prove the following result.

Theorem 3.9 Let N be a relational Bayesian network that only contains exponentially convergent combination functions. Let $\phi(\boldsymbol{v})$ be a first-order formula, and $\boldsymbol{a} \subset \omega$. Then there exists $p \in [0,1]$, s.t.

$$P_n^N(\phi(\boldsymbol{a})) \rightarrow p \qquad (n \rightarrow \infty).$$

Proof: According to our observations at the beginning of this section, we may confine ourselves to the case $\phi(\boldsymbol{v}) = r(\boldsymbol{v})$ for some $r \in S$. Since $r(\boldsymbol{v})$ is equivalent to a finite disjunction of mutually exclusive proper types, it furthermore is sufficient to consider the case $\phi(\boldsymbol{v}) = \tau(\boldsymbol{v})$ for some proper type τ.

Let

$$K_n = n(n-1)\cdots(n - |\boldsymbol{v}| + 1).$$

We partition $\mathrm{Mod}_n(S)$ into subsets

$$M_1, \ldots, M_{K_n}$$

such that

$$\mathscr{M} \in M_j \quad \Leftrightarrow \quad |\{\boldsymbol{a}' \mid \mathscr{M} \models \tau(\boldsymbol{a}')\}| = j.$$

Then

$$P_n^N(\tau(\boldsymbol{a})) = \sum_{j=1}^{K_n} P_n^N(\tau(\boldsymbol{a}) \mid M_j) P_n^N(M_j). \qquad (29)$$

Using that $P_n^N(\tau(\boldsymbol{a}) \mid M_j) = P_n^N(\tau(\boldsymbol{a}') \mid M_j)$ for all tuples $\boldsymbol{a}, \boldsymbol{a}'$ of distinct elements, it follows that

$$P_n^N(\tau(\boldsymbol{a}) \mid M_j) = \frac{j}{K_n}. \qquad (30)$$

By theorem 3.8 there exists $p \in [0,1]$ such that for all $s > 0$

$$P_n^N(\{M_j \mid \frac{j}{n^{|\boldsymbol{v}|}} \in [(1-s)p, (1+s)p]\}) \rightarrow 1.$$

Since $n^{|\boldsymbol{v}|}/K_n \rightarrow 1$ for $n \rightarrow \infty$, this means that also

$$P_n^N(\{M_j \mid \frac{j}{K_n} \in [(1-s)p, (1+s)p]\}) \rightarrow 1.$$

This, together with (29) and (30) then shows that

$$P_n^N(\tau(\boldsymbol{a})) \rightarrow p.$$

□

Theorem 3.9 contains several previous convergence results as special cases. First, it is obviously possible to define the uniform distribution U_n on $\mathrm{Mod}_n(S)$ by a combination function free relational Bayesian network (with $F_{r_i} = 1/2$ for all $r_i \in S$). Hence, theorem 3.9 is applicable and yields the original 0-1-law for first-order logic in the form proved by Glebskiĭ et al. (1969) (i.e. for a language with constants).

The networks defining uniform distributions on parametric classes, which were mentioned in the remark following example 2.5, only use the combination function *max*. Hence, theorem 3.9 again is applicable, and we gain the conditional convergence laws for parametric conditions, as originally shown by Oberschelp (1982).

Convergence laws for sparse random graphs, on the other hand, do not follow from our result, because $count^{\alpha}$ is not exponentially convergent. For completeness' sake, we also mention that convergence holds, again, for the approximation to Gilbert random graphs described in example 2.7.

When we apply theorem 3.8 to the network defining the uniform distributions U_n, we will get (20) with $p > 0$ for all types τ, σ (more precisely, we will obtain that p is $1/m$ with m the number of possible extensions of τ for $\boldsymbol{w}$). Thus, theorem 3.8 implies the original result by Fagin (1976) that every extension axiom is satisfied by almost all structures. Kolaitis and Vardi (1990) strengthened this to the statement that the number of elements that realize a given type extension, almost surely, is at least $\sqrt{n}$. This statement, in turn, is strengthened by theorem 3.8 by showing linear growth of the number of realizing tuples.

From the AI point of view, theorem 3.9 is not completely satisfying: in practical applications of relational Bayesian networks we will almost always be interested in conditional probabilities $P_n^N(\phi(\boldsymbol{a}) \mid \psi(\boldsymbol{a}))$, rather than unconditional ones. It is well known that even in the special case where $P_n^N = U_n$, and ϕ, ψ are first-order sentences, the conditional probability $U_n(\phi \mid \psi)$ need not converge (Fagin 1976). Fortunately, this negative result does not affect us as much as it does classical finite model theory. There the focus is on languages without constants, whose probabilities then can only converge to either 0 or 1. For conditional probabilities $U_n(\phi \mid \psi)$ of interest (i.e. with ψ being some non-trivial condition) it then will be the case that $U_n(\psi) \to 0$, which is a necessary condition for the non-convergence of $U_n(\phi \mid \psi)$. The conditioning events $\psi(\boldsymbol{a})$ interesting for practical applications, in contrast, may very well have a positive limiting probability, in which case the convergence of $P_n^N(\phi(\boldsymbol{a}) \mid \psi(\boldsymbol{a}))$ is ensured. As a matter of fact, one may argue that in cases where $\psi(\boldsymbol{a})$ represents some observed evidence (which is the standard case), the positivity of $lim_n P_n^N(\psi(\boldsymbol{a}))$ is a necessary condition for the limiting behavior of $P_n^N(\phi(\boldsymbol{a}) \mid \psi(\boldsymbol{a}))$ to carry meaningful information, since otherwise we would derive our inferences in a model that tells us that the observed evidence was a virtual impossibility – which should cast some doubt on the adequacy of our probabilistic model for the situation at hand in the first place.

In a companion paper (Jaeger 1998) it is shown that relational Bayesian networks (again under suitable restrictions imposed on the combination functions) also define a distribution P_ω^N on $\mathrm{Mod}_\omega(S)$, the set of S-structures over ω. It is shown that in analogy to results in classical finite model theory there exists an isomorphism class $M \subset \mathrm{Mod}_\omega(S)$ with $P_\omega^N(M) = 1$. The class M is axiomatized by a set of non-standard extension axioms that for pairs $\tau(\boldsymbol{v}) \subset \sigma(\boldsymbol{v}, \boldsymbol{w})$ of types either say that there exist infinitely many tuples realizing the extension, or that there are no realizations of the extension. When a network N only contains combination functions that are exponentially convergent, and are suitable for defining P_ω^N (this includes *max* and *noisy-or*), then we furthermore get $lim_{n\to\infty} P_n^N(\phi(\boldsymbol{a})) = P_\omega^N(\phi(\boldsymbol{a}))$.

4 Conclusion

Relational Bayesian networks are a method for defining probability distributions on finite relational structures. We here have shown that for a subclass of relational Bayesian networks, characterized by a condition on the convergence properties of admissible combination functions, the distributions defined are asymptotically convergent.

The condition of exponential convergence we here imposed on combination functions certainly is very restrictive, even though it is satisfied by *noisy-or*, which is the combination function most frequently used in practical applications. In theorem 3.8 we have established a very strong sufficient condition for the convergence of $P_n^N(\phi(\boldsymbol{a}))$ for first-order ϕ. This condition is by no means a necessary one, and one might hope to obtain alternative convergence results via completely different arguments, and for classes of relational Bayesian networks characterized by conditions quite different from the ones introduced here. Counterexamples in (Shelah & Spencer 1988), in conjunction with example 2.6, however, caution us that even with quite simple relational Bayesian networks we can produce non-convergent probability measures.

From a finite model-theory perspective, another interesting issue is whether theorem 3.9 can be extended to non first-order properties $\phi(\boldsymbol{a})$. Seeing that previous convergence results for extensions of first-order logic and the uniform distributions (Blass et al. 1985, Kolaitis & M.Y.Vardi 1990, Kolaitis & M.Y.Vardi 1992), like the original 0-1 laws for first-order logic, are essentially based on the existence of a suitable, almost sure "extension theory", in light of theorem 3.8, one may conjecture that similar proofs go through for the measures P_n^N. From the artificial intelligence perspective, however, results for extended logics are of limited relevance, because there we are mostly interested in the convergence behavior of those properties that are expressible in the query language provided for relational Bayesian networks, which is essentially quantifier free first-order logic.

Acknowledgments

Most of this work was conducted during a visit to Stanford University made possible by Daphne Koller. This work was funded in part by DARPA contract DACA76-93-C-0025, under subcontract to Information Extraction and Transport, Inc.

References

Bacchus, F. (1990), *Representing and Reasoning With Probabilistic Knowledge*, MIT Press.

Bacchus, F., Grove, A. J., Halpern, J. Y. & Koller, D. (1997), 'From statistical knowledge bases to degrees of belief', *Artificial Intelligence* **87**, 75–143.

Blass, A., Gurevich, Y. & Kozen, D. (1985), 'A zero-one law for logic with fixed point operator', *Information and Control* **67**, 70–90.

Cooper, G. F. (1990), 'The computational complexity of probabilistic inference using bayesian belief networks', *Artificial Intelligence* **42**, 393–405.

Ebbinghaus, H.-D. & Flum, J. (1995), *Finite Model Theory*, Perspectives in Mathematical Logic, Springer Verlag.

Fagin, R. (1976), 'Probabilities on finite models', *Journal of Symbolic Logic* **41**(1), 50–58.

Glebskiĭ, Y., Kogan, D., Liogon'kiĭ, M. & Talanov, V. (1969), 'Range and degree of realizability of formulas in the restricted predicate calculus', *Kibernetika* **2**. English translation in *Cybernetics* **5** (1972).

Grove, A., Halpern, J. & Koller, D. (1992), Random worlds and maximum entropy, *in* 'Proc. 7th IEEE Symp. on Logic in Computer Science'.

Hagerup, T. & Rüb, C. (1990), 'A guided tour of Chernoff bounds', *Information Processing Letters* **33**, 305–308.

Jaeger, M. (1997), Relational bayesian networks, *in* 'Proceedings of UAI-97'.

Jaeger, M. (1998), Reasoning about infinite random structures with relational bayesian networks, *in* 'Principles of Knowledge Representation and Reasoning: Proceedings of the Sixth International Converence (KR'98)'. to appear.

Kolaitis, P. G. & M.Y.Vardi (1990), '0-1 laws and decision problems for fragments of second-order logic', *Information and Computation* **87**, 302–338.

Kolaitis, P. G. & M.Y.Vardi (1992), 'Infinitary logics and 0-1 laws', *Information and Computation* **98**, 258–294.

Lynch, J. F. (1997), 'Infinitary logics and very sparse random graphs', *Journal of Symbolic Logic* **62**(2), 609–623.

McColm, G. L. (1996), Zero-one laws for gilbert random graphs, *in* 'Proceedings Eleventh Annual IEEE Symposium on Logic in Computer Science', pp. 360–369.

Oberschelp, W. (1982), Asymptotic 0-1 laws in combinatorics, *in* D. Jungnickel, ed., 'Combinatorial Theory', Vol. 969 of *Lecture Notes in Mathematics*, Springer Verlag.

Pearl, J. (1988), *Probabilistic reasoning in intelligent systems : networks of plausible inference*, The Morgan Kaufmann series in representation and reasoning, rev. 2nd pr. edn, Morgan Kaufmann, San Mateo, CA.

Shelah, S. & Spencer, J. (1988), 'Zero-one laws for sparse random graphs', *Journal of the American Mathematical Society* **1**(1), 97–115.

Wellman, M. P., Breese, J. S. & Goldman, R. P. (1992), 'From knowledge bases to decision models', *The Knowledge Engineering Review* **7**(1), 35–53.