Approximations of Bayesian networks through KL minimisation

Wim WIEGERINCK and Bert KAPPEN

RWCP Theoretical Foundation SNN University of Nijmegen
Geert Grooteplein Noord 21, 6525 EZ Nijmegen, The Netherlands

{wimw,bert}@mbfys.kun.nl

Received 30 August 1999

Abstract Exact inference in large, complex Bayesian networks is computationally intractable. Approximate schemes are therefore of great importance for real world computation. In this paper we consider an approximation scheme in which the original Bayesian network is approximated by another Bayesian network. The approximating network is optimised by an iterative procedure, which minimises the Kullback-Leibler divergence between the two networks. The procedure is guaranteed to converge to a local minimum of the Kullback-Leibler divergence. An important question in this scheme is how to choose the structure of the approximating network. In this paper we show how redundant structures of the approximating model can be pruned in advance. Simulation results of model selection and model optimisation are provided to illustrate the methods.

Keywords Bayesian Networks, Variational Approximations, Mean Field Theory, Kullback-Leibler Divergence

New Generation Computing, 18(2000)167-175
Ohmsha, Ltd. and Springer Verlag

§1 Introduction

Bayesian networks provide a rich framework for probabilistic modelling and reasoning [1, 2, 3]. Their graphical structure provides an intuitively appealing modularity and is well suited to the incorporation of prior knowledge. Bayesian networks are often used in a domain with causal structures, such as speech recognition and medical diagnosis. The invention of algorithms for exact inference during the last decades has lead to the rapid increase in popularity of Bayesian networks in modern AI. However, exact inference is NP-hard [4]. In practice, this is reflected in the fact that large densely connected networks, which can be expected to appear in real-world applications, are intractable for exact computations [5].

In this paper, we address the problem of approximate inference in intractable Bayesian networks. In this context, the variational methods gain increasingly interest [6, 7, 8, 9]. An advantage of these methods is that they provide bounds on the approximation error. This is in contrast to stochastic sampling methods [3, 8] which may yield unreliable results due to finite sampling times. Until now, variational approximations have been less widely applied than Monte Carlo methods, arguably since their use is not so straightforward.

The paper is organised as follows. In section 2 we present a variational framework for approximate inference in intractable Bayesian networks using (simpler) approximating Bayesian networks. An iterative algorithm is presented to optimise the parameters of the approximating networks such that the Kullback-Leibler (KL) divergence is minimised. In section 3 we address the problem how to choose the structure of the approximating model. We show that redundant structures can be pruned in advance. In section 4, we consider extremely dense connected networks. For these networks, the optimisation of the approximating network by KL minimisation is intractable. A way out is to minimise an approximation of KL instead. In section 5, we present simulation results on Lauritzen's chest clinic (ASIA) model and a random intractable network to illustrate the method. We conclude with a discussion and future plans in section 6.

§2 Approximating Bayesian networks

Basically, the problem of inference in a Bayesian network P is to find the conditional probability distribution $P(S_i|E)$ of each of the nodes i given the evidence E. If P is intractable, one has to approximate these conditional

probabilities. In the variational method [6, 8], the intractable probability distribution $P(S|E) = P_E(S)$ is approximated by a tractable distribution $Q(S)$ (on the non-evidential nodes). Then Q is used to compute the node probabilities $Q(S_i)$. In the standard (mean field) approach, Q is assumed to be completely factorised $Q(S) = \prod_i Q(S_i)$. We take the more general approach [10, 9], with Q being a tractable, but otherwise fully unconstrained Bayesian network. Therefore, to construct Q we first have to define a suitable tractable structure for Q: $Q(S) = \prod_i Q(S_i|\pi_i^q)$, where π_i^q denote the parents of node i in graphical structure Q. The next step is to optimise the parameters of Q such that the Kullback-Leibler (KL) divergence between Q and P_E,

$$D(Q, P_E) = \sum_{\{S\}} Q(S) \log \frac{Q(S)}{P_E(S)} \equiv \left\langle \log \frac{Q(S)}{P_E(S)} \right\rangle_Q \tag{1}$$

is minimised. The KL-divergence is related to the difference of the marginals of Q and P_E,

$$\max_i |P(S_i|E) - Q(S_i)| \leq \sqrt{\frac{1}{2} D(Q, P_E)} \tag{2}$$

(see [11]). The KL-divergence satisfies $D(Q, P_E) \geq 0$, and $D(Q, P_E) = 0 \Leftrightarrow Q = P_E$. Using $P(S|E) = P(S, E)/P(E)$ and substituting the graphical structures for P and Q, we can rewrite D as

$$D(Q, P_E) = \left\langle \sum_i \log Q(S_i|\pi_i^q) - \sum_i \log P(S_i|\pi_i^p) \right\rangle_Q + \text{constant}. \tag{3}$$

Parent sets π_i^q and π_i^p are understood with respect to the probability distribution Q and P, respectively and are in principle different. As a consequence of the factorisation of $P(S, E)$ into conditionals, the average $\langle \log P(S, E) \rangle_Q$ reduces to the sum of local averages $\sum_i \langle \log(P(S_i|\pi_i^p) \rangle_Q$, which facilitates the tractability of D.

$D(Q, P_E)$ depends on the numerical values of the conditional probability tables $Q(S_i|\pi_i^q)$. One way to optimise the tables is minimising the KL-divergence by a gradient based method. Another approach is to set the gradient of D with respect to these parameters equal to zero, yielding the equations

$$Q(S_i|\pi_i^q) = \frac{1}{Z_i} \exp \left\langle \sum_k \log P(S_k|\pi_k^p) - \sum_{k \neq i} \log Q(S_k|\pi_k^q) \right\rangle_{Q^c}. \tag{4}$$

in which the average $\langle \ldots \rangle_{Q^c}$ is taken with respect to the conditional probability distribution $Q^c = Q(S|S_i\pi_i^q)$, in which node i and its parents are clamped to the states S_i and π_i^q, respectively. Z_i is a normalisation factor. Eqs. 4 are a coupled set of non-linear equations that must be solved for $Q(S_i|\pi_i^q)$. For each i, the right hand side of Eqs. 4 does not depend on the parameters $Q(s_i|\pi_i^q)$. This means that asynchronous iteration of Eqs. 4 is guaranteed to converge to a local minimum of the KL-divergence.

The quality of the approximation depends strongly on the structure of Q. The simplest approach is the so called mean-field approach, in which the graph of Q is completely disconnected, i.e. $Q(S) = \prod_i Q(S_i)$. Then Eqs. 4 reduces to the standard mean field equations

$$Q(S_i) = \frac{1}{Z_i} \exp \left\langle \sum_k \log P(S_k|\pi_k^p) \right\rangle_{Q^c}$$

The other extreme is to factorise Q according to a triangulated graph [12, 3] of P. In this case, iteration of Eqs. 4 leads to the solution $Q = P_E$ and $D = 0$. This solution is only theoretically of interest, since the computational complexity of Q in this case is equal to the original inference problem. However, it indicates that the variational approach using structure interpolates between the standard mean field theory and the exact solution. In general one must choose a structure for Q that is a good compromise between approximation error and complexity.

§3 Pruning the approximating model

An important question is how to choose the structure of Q to get the best out of the approximation. In principle, the number of possible structures grows exponentially with the number of nodes. A sensible heuristic is to try making the graphical overlap between Q and P_E as large as computationally possible [10, 9]. The following lemma indeed shows that graphical structures in Q which do not appear in P are redundant if they satisfy the following lemma.

Lemma 3.1

Let P be an Bayesian network, let Q be an approximating Bayesian network and let $Q(S_i|S_p\pi_{ir}^q)$ be the conditional probability table of node i. (i.e. node i conditionally depends on node p and some other parents π_{ir}^q) If each family set (i.e., the set of a node and its parents) FAM from the graph of P and from the graph of Q except for the family set of i in Q, satisfies at least one of the

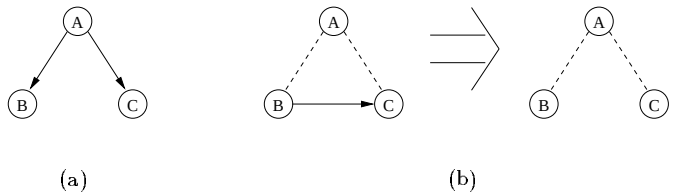


Fig. 1 Example of network pruning. (a): Graph of exact model $P(A)P(B|A)P(C|A)$. (b): Left: graphs of original approximating model $Q(A)Q(B)Q(C|B)$. Right: graph of pruned model $Q(A)Q(B)Q(C)$. In (b), arrows indicate the graphical structure of the approximating model. Dashed lines indicate the underlying family sets $\{A, B\}$ and $\{A, C\}$ in the exact model. Since there is no family set with both B and C as a member, the link is removed. Note, however, that in the exact model B and C are marginally dependent (via A).

following two independency relations:

$$Q(FAM|S_i S_p \pi^q_{ir}) = Q(FAM|S_i \pi^q_{ir})$$

$$Q(FAM|S_i S_p \pi^q_{ir}) = Q(FAM|S_p \pi^q_{ir})$$

then the table $Q(S_i|S_p\pi^q_{ir})$ reduces to $Q(S_i|\pi^q_{ir})$ by iteration of (4).

Proof Let $\{k'\}$ be the nodes for which the family sets in P satisfy $Q(S_{k'}, \pi^p_{k'}|S_i S_p \pi^q_{ir}) = Q(S_{k'}, \pi^p_{k'}|S_i \pi^q_{ir})$, and let $\{l'\}$ be nodes other than i for which the family sets in Q satisfy $Q(S_{l'}, \pi^q_{l'}|S_i S_p \pi^q_{ir}) = Q(S_{l'}, \pi^q_{l'}|S_i \pi^q_{ir})$, By assumption, only for these nodes the averages $\langle \log P(S_{k'}|\pi^p_{k'})\rangle_{Q(S|S_i S_p \pi^q_{ir})}$ and $\langle \log Q(S_{l'}|\pi^q_{l'})\rangle_{Q(S|S_i S_p \pi^q_{ir})}$ depend on the clamped state S_i. As a consequence, equation (4) can be reduces to

$$Q(S_i|S_p\pi^q_{ir}) = \frac{1}{Z_i} \exp\left\langle \sum_{k'} \log P(S_{k'}|\pi^p_{k'}) - \sum_{l'} \log Q(S_{l'}|\pi^q_{l'}) \right\rangle_{Q^c}, \qquad (5)$$

(with $S_i S_p \pi^q_{ir}$ clamped), since the contributions of nodes other than k' and l' can be absorbed in Z_i. By construction, the family sets of k' in P and l' in Q are conditionally independent of S_p given $S_i\pi^q_{ir}$. A direct consequence is that expression (5) does not depend on the state of of S_p, i.e. $Q(S_i|S_p\pi^q_{ir}) = Q(S_i|\pi^q_{ir})$ □

In fig. 1 a simple illustration of the lemma is given.

§4 Approximated minimisation

The complexity of the variational method is at least exponential in the parent size of the exact model P, since it requires the computation of averages of the form $\langle \log P(S_i|\pi_i^p) \rangle_Q$. This means that computational advantage can only be obtained if the parent size is much smaller than the clique size of P [12, 3]. Since the storage space of probability tables is exponential in the parent size, in practical applications probability tables with large number of parents will be parametrised. Popular parametrisations are noisy-OR gates [1] and weighted sigmoid functions [13]. For these parametrisations $\langle \log P(S_i|\pi_i^p) \rangle_Q$ can be approximated by a tractable quantity $\mathcal{E}_i(Q, \xi)$ (which may be defined using additional variational parameters ξ). As an example, consider tables parametrised as sigmoid functions,

$$P(S_i = 1|\{S_k\}) \;=\; (1 + \exp(z_i))^{-1} \tag{6}$$

where z_i is the weighted input of the node, $z_i = \sum_k w_{ik} S_k + h_i$. In this case, the averaged log probability is intractable for large parent sets. To proceed we can use the approximation proposed in [6]

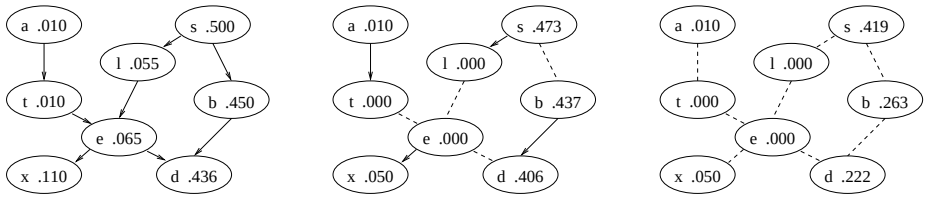
$$\langle \log(1 + e^{z_i}) \rangle_Q \leq \xi_i \langle z_i \rangle + \log \left\langle e^{-\xi_i z_i} + e^{(1-\xi_i) z_i} \right\rangle \equiv \mathcal{E}_i(Q, \xi) \tag{7}$$

which is tractable if Q is tractable. Numerical optimisation of $\mathcal{L}(Q, \xi) \equiv \mathcal{E}(Q, \xi) - \langle \log(Q) \rangle_Q$ with respect to Q and ξ leads to local minimum of an upper bound of the KL-divergence. Note however, that iteration of fixed point equations derived from $\mathcal{L}(Q, \xi)$ does not necessarily lead to convergence, due to the nonlinearity of $\mathcal{E}$ with respect to Q.

§5 Numerical results

We illustrate the theory by two toy problems. The first one is inference in Lauritzen's chest clinic model (ASIA), defined on 8 binary variables $\{a, t, s, l, b, e, x, d\}$ (see [12] for more details about the model). We compared exact marginals with approximate marginals using the approximating models in figure 2. From the results, we can conclude that adding structure to the approximating network decreases the error in the approximation. However, we also can see from the simulation results that even the fully disconnected mean field approximation is qualitatively correct (maximum error between marginals $P(S_i)$ and $Q(S_i)$ is about 0.2).

In the second toy problem we simulated approximate inference in a structure that has both tractable substructures and sigmoidal nodes with large parent sets. We generated models with graphical structure as in figure 3. The upper

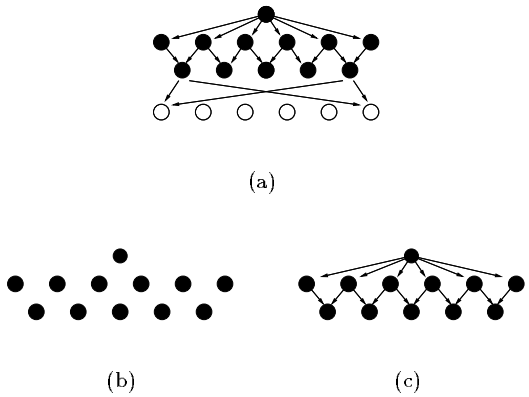


(a) KL = 0 (b) KL = 0.11 (c) KL = 0.35

Fig. 2 Chest clinic model (ASIA). (a): Exact model with marginal probabilities. (b-c): Approximating models with approximated marginal probabilities. Arrows indicate the graphical structure of the exact and approximating models. Dashed lines indicate the underlying links in the exact model. KL is the KL-divergence $D(Q, P)$ between the approximating model Q and the true model P.

node is a mixture node with m mixture components. The next layer consists of $n + 1$ binary nodes. The third layer consists of n binary nodes S_i. Each of these nodes has two parents in the preceding layer. Up to this layer the network is tractable. We refer to this part of the network as $\mathcal{N}_1$. This part of the network represents some underlying causal structure in the model, e.g. a causal structure of diseases and pathophysiological mechanisms in a model for medical diagnosis, and may have been designed using expert knowledge. Finally, there is a layer of n_v observables S_v. These are parametrised by sigmoid functions, receiving weighted inputs from all the nodes of the preceding layer. The goal is to find marginal probabilities of the nodes in the third layer given evidence on the observable nodes. Exact computation of these probabilities is intractable for large n.

We choose $m = 10$, $n_v = 50$ and varied $n = 8, \ldots, 15$. Networks of this size are still tractable for exact computation. The values in the probability tables of $\mathcal{N}_1$ are drawn uniformly. The weights in the sigmoidal functions are drawn from the Gaussian distribution with zero mean and standard deviation $1/\sqrt{n}$. We computed exact and approximated marginals for the third layer S_i. As approximating models we used a factorised model and a model with the tractable structure $\mathcal{N}_1$ (fig. 3). In figure 4 we plotted the maximal error $\max_d |Q(S_i) - P(S_i|S_v)|$ as a function of the network size. We also plotted the required computer time for exact and approximate inference as a function of the network size. In the optimisation of the approximating model with structure, we



(a) (b) (c)

Fig. 3 (a): Graphical structure of artificially generated probability distribution P. Non-evidential nodes are black. Evidential nodes are white. (b) and (c): Graphical structure on the non-evidential nodes of the approximating distributions Q.

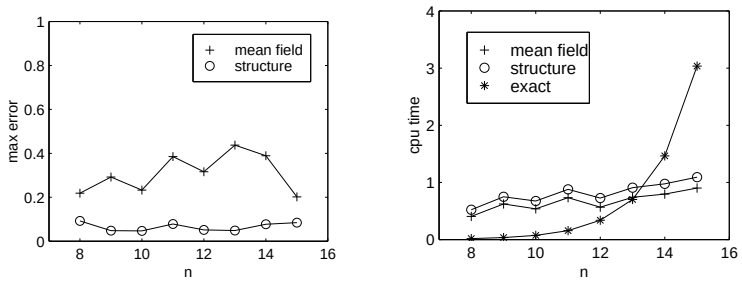


Fig. 4 Left: The maximal error as a function of the network size. Right: CPU-time (arbitrary units) for exact and approximate inference as a function of the network size

used the optimised factorised model as initialisation. Thus, the computation of the structured model can be seen as post-processing step after the optimisation of the factorised model. This is reflected in the plotted CPU-times.

We conclude that variational methods using structure significantly improves the quality of approximation, within feasible computer time. In a network with tractable substructures, as can be expected in many practical applications such as medical diagnosis, these substructures provide a useful starting point for the approximating model.

§6 Discussion and future plans

Finding accurate approximations of Bayesian networks is crucial if their

application to large scale problems is to be realised. We have presented a scheme to use (simpler) approximating Bayesian networks, tuned via minimisation of the Kullback-Leibler divergence. We have addressed the question of selecting the structure of the approximating model, and showed that redundant structures can be pruned in advance. Parametrised models with large parent sets can be dealt with by minimising an approximation of the KL divergence between true and approximating model.

Numerical results reported here, as well as results on the Asia problem with evidence (not reported here) show that the factorised variational approximation is qualitatively correct in the sense that it correctly estimates whether probabilities are high or low. However, the numerical errors can be rather large. The results of approximations using structure gives a significant improvement. Our results seem to indicate that this improvement is independent of the problem size.

One of the current research items is to further investigate the optimal structure for Q. In addition, we intent to build a package of C++ routines for (automated) model selection, model optimisation and approximate inference. These routines are to be contributed to an RWCP library. Currently we are also involved in a joint project with Utrecht University Hospital to build a large and detailed system for diagnosis in internal medicine. This system will be based on a Bayesian network with many tractable substructures. Our aim is to use our approximate routines for this system as a RWCP demonstration project.

6.1 Acknowledgements

We thank David Barber for stimulating discussions. We are grateful to Ali Taylan Cemgil for sharing his matlab routines.

References

1) J. Pearl. *Probabilistic Reasoning in Intelligent systems: Networks of Plausible Inference.* Morgan Kaufmann Publishers, Inc., 1988.

2) F.V. Jensen. *An introduction to Bayesian networks.* UCL Press, 1996.

3) E. Castillo, J. M. Gutierrez, and A. S. Hadi. *Expert Systems and Probabilistic Network Models.* Springer, 1997.

4) G.F. Cooper. The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence*, 42(393-405), 1990.

5) M.A Shwe, B. Middleton, D.E. Heckerman, M. Henrion, Horvitz E.J., H.P. Lehman, and G.F. Cooper. Probabilistic Diagnosis Using a Reformulation of

the Internist-1/ QMR Knowledge Base. *Methods of Information in Medicine*, 30:241–55, 1991.

6) L.K. Saul, T. Jaakkola, and M.I. Jordan. Mean field theory for sigmoid belief networks. *Journal of Artificial Intelligence Research*, 4:61–76, 1996.

7) T.S. Jaakkola and M.I. Jordan. Variational probabilistic inference and the QMR-DT network. *Journal of Artificial Intelligence Research*, 10:291–322, 1999.

8) M. I. Jordan, editor. *Learning in Graphical Models*, volume 89 of *NATO ASI, Series D: Behavioural and Social Sciences*. Kluwer, 1998.

9) D. Barber and W. Wiegerinck. Tractable variational structures for approximating graphical models. In M.S. Kearns, S.A. Solla, and D.A. Cohn, editors, *Advances in Neural Information Processing Systems*, volume 11. MIT Press, 1999.

10) W. Wiegerinck and D. Barber. Mean field theory based on belief networks for approximate inference. In L. Niklasson, M. Bodén, and T. Ziemke, editors, *ICANN'98: Proceedings of the 8th International Conference on Artificial Neural Networks, Skövde, Sweden, 2-4 September 1998*, volume 2, pages 499–504, London, 1998. Springer.

11) J. Whittaker. *Graphical models in applied multivariate statistics*. Wiley, Chichester, 1990.

12) S.L. Lauritzen and D.J. Spiegelhalter. Local computations with probabilties on graphical structures and their application to expert systems. *J. Royal Statistical society B*, 50:154–227, 1988.

13) R. Neal. Connectionist learning of belief networks. *Artificial Intelligence*, 56:71–113, 1992.